DAWN PUTNAM SESSION

Putnam 2022 — Papers A and B

Putnam 120 Archive

Contents

Putnam 2022 — Paper A **5**
2022 A1 5
2022 A2 7
2022 A3 10
2022 A4 13
2022 A5 16
2022 A6 20

Putnam 2022 — Paper B **27**
2022 B1 27
2022 B2 29
2022 B3 32
2022 B4 35
2022 B5 39
2022 B6 44

Putnam 2022 — Paper A

2022 A1

2022 A1. Determine all ordered pairs of real numbers (a, b) such that the line $y = ax + b$ intersects the curve $y = \ln(1 + x^2)$ in exactly one point.

Fields: Calculus (monotonicity, the mean value theorem, convexity of a curve with an inflection pair) **Position:** A1 – the opener **Answer:** The line meets the curve exactly once precisely when $a = b = 0$; or $|a| \geq 1$ (any b); or $0 < |a| < 1$ and b lies outside the closed interval between the two "tangency values" $g(r_-)$ and $g(r_+)$, where $r_\pm = (1 \pm \sqrt{1 - a^2})/a$ are the two points where the curve has slope a.

Layer 1 THE WHAT — surface analysis

The natural attempt. Write $f(x) = \ln(1 + x^2)$ and set $g(x) = f(x) - ax$, so intersections of the line with the curve are roots of $g(x) = b$. The honest first move is to differentiate: $f'(x) = 2x/(1 + x^2)$, and you notice at once that $f'(0) = 0$ and, checking $f''(x) = (2 - 2x^2)/(1 + x^2)^2$, that f is convex on $(-1, 1)$ and concave outside it, with $f'(1) = 1$ the largest slope the curve ever attains and $f'(-1) = -1$ the smallest. That maximum slope of 1 looks like the natural boundary of the problem, and it pulls you toward a specific, narrower question: "for which (a, b) is the line *tangent* to the curve?" You set up $f(x_0) = ax_0 + b$ and $f'(x_0) = a$ simultaneously, eliminate b, and get a clean one-parameter family of tangent lines indexed by x_0. This is real progress and it is not wasted – the tangency condition $f'(x_0) = a$, i.e. $2x_0 = a(1 + x_0^2)$, is exactly the quadratic that produces $r_\pm$ later. But treating "exactly one intersection" as a synonym for "tangent" is the trap: a generic transversal line with $|a| \geq 1$ meets the curve exactly once too, with no tangency anywhere, because the curve simply cannot keep up with a slope-1-or-steeper line for long. If you chase the tangency idea alone you produce a family of special lines and miss the entire regime $|a| \geq 1$, and you also have no account of $a = 0$, where the single valid line $y = 0$ is single not because of any tangency but because of the curve's symmetry and sign. The tangency framing answers "when can the line touch the curve without crossing," which is a different question from the one asked.

The model chain. Since f is even, $y = ax + b$ meets f the same number of times as $y = -ax + b$ meets f, so it suffices to treat $a \geq 0$ and reflect. For $a = 0$: $f(x) > 0$ for all $x \neq 0$ and f is even, so $f(x) = b$ has two solutions for $b > 0$, one solution ($x = 0$) for $b = 0$, and none for $b < 0$. Only $(0, 0)$ works among horizontal lines.

For $a \geq 1$: since $f'(x) \leq 1$ for every x (with equality only at $x = 1$), the mean value theorem gives, for $a > 1$ and any $x_0 < x_1$, $\frac{f(x_1) - f(x_0)}{x_1 - x_0} \leq 1 < a$, so $f(x_1) - f(x_0) \neq a(x_1 - x_0)$: the line cannot meet f twice. For $a = 1$ the same integral argument works with a strict inequality because $f'(x) = 1$ at only the single point $x = 1$, so equality in the mean value bound cannot be sustained across an interval. Existence of at least one intersection follows from the intermediate value theorem: $ax + b - f(x) \to -\infty$ as $x \to -\infty$ and, since $f(x)/x \to 0$, $\to +\infty$ as $x \to +\infty$. So every line with $a \geq 1$ meets the curve in exactly one point, for every b.

For $0 < a < 1$: $g'(x) = f'(x) - a$ vanishes exactly at the two roots of $2x = a(1 + x^2)$, namely $r_\pm = \frac{1 \pm \sqrt{1 - a^2}}{a}$, both positive since $a > 0$, with $r_- r_+ = 1$ and $r_- < 1 < r_+$. Because f' is

increasing on $(-1, 1)$ and decreasing outside it, g' is negative on $(-\infty, r_-)$, positive on (r_-, r_+), negative on (r_+, ∞): g falls to a local minimum at r_-, rises to a local maximum at r_+, then falls again, with $g \to +\infty$ as $x \to -\infty$ and $g \to -\infty$ as $x \to \infty$. A function with exactly this shape hits any value below its local minimum or above its local maximum exactly once, hits any value strictly between them exactly three times, and hits the two extreme values themselves exactly twice (the extremum point plus one more crossing). So $g(x) = b$, i.e. the line meets the curve exactly once, exactly when $b < g(r_-)$ or $b > g(r_+)$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains reach the same critical differential fact – f' has maximum 1 at $x = 1$ and this is where the equation $f'(x) = a$ first becomes interesting – by the same short computation. The split is what question gets asked of that fact. "Both paths compute $f'(x) = 2x/(1 + x^2)$ and its extremum 1 at $x = 1$. At that point the natural attempt asks *where is the line tangent to the curve*; the model asks *for which values of a does the line even have the physical room to cross the curve more than once, and what does the auxiliary function $g(x) = f(x) - ax$ do globally, not just locally at a touch point*."

> *Tangency is a statement about one point; "exactly one intersection" is a statement about the whole real line. I should not be solving for special lines that touch the curve – I should be asking, for a fixed slope a, how many times a horizontal level b gets hit by the single function $g(x) = f(x) - ax$, and that means I need the whole shape of g: where it rises, where it falls, where it is unbounded.*

> *The number 1 that fell out of $f'(1) = 1$ is not "the tangent slope I'm looking for" – it's a threshold on a itself. If $a \geq 1$ the line is at least as steep as the curve ever gets, so $g = f - ax$ can never turn around: it's monotonic, hence injective, hence at most one root regardless of b. Only once $a < 1$ does the curve get to "outrun" the line locally, giving g somewhere to have a local max and min, and only then do multiple crossings become geometrically possible at all.*

- ▷ **Cue:** a problem asks for a line to meet a curve in exactly one point, for varying slope and intercept $\longrightarrow$ **Move:** do not go looking for tangent lines specifically; instead fix the slope, form $g(x) = f(x) - ax$, and determine the full shape (monotonicity, turning points, limits at $\pm\infty$) of that one auxiliary function as a function of a
- ▷ **Cue:** a smooth function's derivative is bounded by some constant M everywhere $\longrightarrow$ **Move:** compare M to the line's slope a: for $|a| \geq M$ the shifted function $f(x) - ax$ is monotonic (via the mean value theorem), collapsing "how many intersections" to "does it cross zero", which IVT plus the limits at $\pm\infty$ answers immediately

Layer 3 THE WHY

Why must the question change from "where is it tangent" to "what is the global shape of g"? Because "exactly one solution" is fundamentally a statement counting preimages of a single value under a single real function, and counting preimages is controlled by monotonicity and turning points, not by any one special point where the derivative happens to match a. Tangency is what you ask when you want to know the boundary *between* regimes with different intersection counts, not the count itself. The reason $r_\pm$ genuinely matter is that they are exactly the boundary points where g changes from falling to rising to falling again – so they are the right object once you already know you are hunting for turning points of g, but wrong as the primary target, because a turning point of g only tells you *where the graph of g changes direction*, and it is only after you also know g's limiting behaviour at $\pm\infty$ (forced, here, by $f(x)/x \to 0$, itself forced by f being a logarithm and hence sub-linear) that you can read off how many times a horizontal level crosses it.

The threshold $a = 1$ is not an arbitrary constant to be memorised; it is forced by the specific shape of f. $f(x) = \ln(1 + x^2)$ is built from a logarithm of a quadratic, and its derivative

$2x/(1+x^2)$ is a rational function whose numerator is linear and whose denominator is quadratic – exactly the shape that guarantees a bounded derivative with an attained maximum (AM–GM on $2x$ against $1+x^2$ gives $2x \le 1+x^2$ directly, i.e. $f'(x) \le 1$ with no calculus at all). Once you see the derivative is *bounded*, asking "is the line's slope inside or outside that bound" is the natural first question, of exactly the same character as the Blueprint's circle: the bound on f' is the fixed structural fact, and everything downstream – monotonicity of g for $|a| \ge 1$, the turning-point analysis for $|a| < 1$ – is read off it, not invented separately for each regime.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that $f'(x) = 2x/(1+x^2)$ is visibly bounded (by AM–GM, $2x \le x^2+1$), attaining 1 at $x=1$ and -1 at $x=-1$ with no other extrema. That bound should be written down immediately, before any attempt to solve $f(x) = ax+b$ directly.

2. That cue suggests splitting on whether $|a|$ exceeds the bound: for $|a| \ge 1$, monotonicity of $g = f - ax$ via the mean value theorem; for $|a| < 1$, a full turning-point analysis of g using where $f'(x) = a$.

3. Hesitation occurs right at the boundary case $a = 1$: the mean value theorem gives only $f'(x) \le 1$, not < 1, everywhere, so a naive MVT argument looks like it might allow equality (hence a second intersection). It is reinforced away by noting $f'(x) = 1$ at only the single point $x = 1$, so the strict inequality survives across any interval of positive length, which the model chain states explicitly via the integral of f'.

4. The trap is conflating "exactly one intersection" with "the line is tangent to the curve." Tangent lines are the boundary between the one-intersection and three-intersection regimes when $0 < |a| < 1$; they are irrelevant to the $|a| \ge 1$ regime, where every line meets the curve exactly once with no tangency in sight, and irrelevant again to $a = 0$, which is decided by parity and sign, not calculus at all.

5. Recognition would be faster next time by treating "is the derivative of this curve bounded, and by what" as an automatic first computation whenever a problem counts intersections between a curve and a family of lines, before attempting any algebraic elimination between the two equations.

6. The reusable decision rule: to count intersections of $y = f(x)$ with the line $y = ax+b$, do not solve the two equations together. Form $g(x) = f(x) - ax$, find where its derivative vanishes to locate the turning points, use the bound on f' (or the sign structure of f'') to classify those turning points as none (if $|a|$ exceeds $\sup|f'|$, so g is monotonic), or a local max/min pair (if $|a|$ is inside that bound), and read the number of solutions to $g(x) = b$ directly off that shape together with g's limits at $\pm\infty$.

2022 A2

2022 A2. Let n be an integer with $n \ge 2$. Over all real polynomials $p(x)$ of degree n, what is the largest possible number of negative coefficients of $p(x)^2$?

Fields: Algebra (polynomial coefficients, induction on a sign pattern) **Position:** A2 – second problem
Answer: The largest possible number of negative coefficients of $p(x)^2$ is $2n-2$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The instinct is to build examples first and count. Try $p(x) = x^n - x^{n-1} - \cdots - x - 1$ or some other polynomial with many negative coefficients, square it on a small case like $n = 3$, and see how many of the $2n+1 = 7$ coefficients of $p(x)^2$ come out negative. This is useful reconnaissance and it correctly suggests the answer lives near $2n-2$, but it is a search, not a proof, and it gives no reason the bound cannot be beaten by a cleverer p. The natural next move is to try to push the construction to make *every* coefficient of $p(x)^2$ except the top and

bottom negative – after all, $b_0 = a_0^2 \geq 0$ and $b_{2n} = a_n^2 \geq 0$ are forced nonnegative by being pure squares, so $2n-2$ is the obvious ceiling to *aim* at, and a student who has only counted examples will spend real time hunting for a polynomial that hits all $2n-1$ of the remaining coefficients before accepting that it can't be done. Without an argument for why $b_1, \ldots, b_{2n-1}$ cannot *all* be negative, the search never terminates in a proof – you can always try one more $p(x)$.

The model chain. Write $p(x) = a_n x^n + \cdots + a_1 x + a_0$ and $p(x)^2 = b_{2n}x^{2n} + \cdots + b_0$, where $b_k = \sum_{i+j=k} a_i a_j$. Since $b_0 = a_0^2$ and $b_{2n} = a_n^2$ are squares, they are automatically ≥ 0, so at most $2n-1$ of the remaining coefficients could be negative, and the claim to prove is that not even all $2n-1$ of $b_1, \ldots, b_{2n-1}$ can be simultaneously negative.

Suppose for contradiction that $b_i < 0$ for every $i = 1, \ldots, 2n-1$. Since squaring erases global sign ($p(x)^2 = (-p(x))^2$), assume WLOG $a_0 > 0$. From $b_1 = 2a_0a_1 < 0$ and $a_0 > 0$ we get $a_1 < 0$. Now induct: suppose $a_1, \ldots, a_{k-1} < 0$ for some $2 \leq k \leq n$. Then

$$b_k = 2a_0a_k + \sum_{i=1}^{k-1} a_i a_{k-i}.$$

Every term $a_i a_{k-i}$ in that sum has $1 \leq i \leq k-1$, so both a_i and a_{k-i} lie in the range covered by the inductive hypothesis (indices 1 through $k-1$) and are therefore negative, making each product *positive*. So $\sum_{i=1}^{k-1} a_i a_{k-i} > 0$, and since $b_k < 0$,

$$2a_0a_k = b_k - \sum_{i=1}^{k-1} a_i a_{k-i} < b_k < 0,$$

forcing $a_k < 0$ (as $a_0 > 0$). The induction runs all the way to $k = n$, giving $a_1, \ldots, a_n$ all negative. But then $b_{2n-1} = 2a_{n-1}a_n$ is a product of two negative numbers, hence *positive* – contradicting the assumption $b_{2n-1} < 0$. So the all-$2n-1$-negative scenario is impossible, and at most $2n-2$ coefficients of $p(x)^2$ can be negative.

For sharpness, take $p(x) = n(x^n+1) - 2(x^{n-1} + \cdots + x)$. Then

$$p(x)^2 = n^2(x^{2n} + x^n + 1) - 2n(x^n+1)(x^{n-1} + \cdots + x) + (x^{n-1} + \cdots + x)^2.$$

For each $i \notin \{0, n, 2n\}$, the coefficient of x^i receives a contribution of at most $-2n$ from the cross term $-2n(x^n+1)(x^{n-1} + \cdots + x)$ (every such i picks up at least one $-2n$ from pairing x^n or 1 against some x^j, $1 \leq j \leq n-1$) and at most $2n-2$ from expanding $(x^{n-1} + \cdots + x)^2$ (that polynomial has $n-1$ terms, so any coefficient in its square is a sum of at most $n-1$ products of 1's, i.e. at most $n-1$ pairs, contributing at most $2(n-1) = 2n-2$ counting both orders), for a net coefficient of at most $-2n + 2n - 2 = -2 < 0$. That accounts for all $2n-1$ indices except possibly $i = n$ itself, whose sign need not be checked at all: sharpness only requires $2n-2$ negative coefficients among the $2n-1$ candidates, and the bound above has just produced negativity at every one of them except possibly that single middle index. So $p(x)^2$ achieves at least $2n-2$ negative coefficients, matching the upper bound exactly.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains agree that $b_0, b_{2n} \geq 0$ are forced and that $2n-2$ is therefore the natural target. The split is method: "at this point the natural attempt keeps searching for a single polynomial that beats $2n-2$, hoping construction alone will reveal the ceiling; the model instead assumes the ceiling is beaten – assumes $b_1, \ldots, b_{2n-1}$ are all negative – and tracks what that assumption forces about the signs of $a_1, \ldots, a_n$ one coefficient at a time."

> *I keep constructing polynomials and counting negative coefficients by hand, but no finite amount of trying tells me the ceiling is exactly* $2n-2$ *rather than* $2n-1$ *– I need to instead* suppose *all*

> *$2n-1$ middle coefficients are negative and see whether that supposition is even internally consistent, working forward from b_1 through an induction on the signs of the a_i themselves, not the b_i.*

> *Squaring a polynomial destroys the sign of its leading behaviour – p and $-p$ give the same $p(x)^2$ – so whatever contradiction I am chasing, I am free to fix $a_0 > 0$ without loss of generality. That one normalisation is what lets a chain of forced negativity in $a_1, a_2, \ldots$ get started from $b_1 = 2a_0a_1 < 0$ at all.*

> *Once I know $a_1, \ldots, a_{k-1}$ are all negative, the cross-term sum $\sum_{i=1}^{k-1} a_i a_{k-i}$ inside b_k is a sum of products of two negative numbers, so it is forced* positive *– and a positive quantity being subtracted only helps push $2a_0a_k$ more negative, not less. The induction is self-reinforcing: each new negative a_i makes the next one easier to force negative, right up until it collides with b_{2n-1}, which is a product of exactly two of these forced-negative terms and therefore cannot itself be negative.*

- ▷ **Cue:** a problem bounds how many coefficients of a squared (or otherwise self-multiplied) polynomial can share a sign $\longrightarrow$ **Move:** assume the extremal case (every candidate coefficient has that sign) and propagate an induction on the signs of the *original* coefficients, since products of same-signed terms have a forced sign that a direct count of the product's coefficients does not reveal
- ▷ **Cue:** a polynomial's constant and leading coefficients appear squared in $p(x)^2$ $\longrightarrow$ **Move:** note immediately they are forced nonnegative, which sets the achievable ceiling before any induction begins

Layer 3 THE WHY

Why must the argument proceed by assuming the extremal case and deriving a contradiction, rather than by direct construction-and-count? Because the claim being proved is universal – *no* polynomial of degree n achieves $2n-1$ negative coefficients – and a universal negative claim about all polynomials of a given degree cannot be established by exhibiting polynomials; only a proof that derives a structural impossibility from the assumption can close off every candidate at once. The specific mechanism that makes the impossibility provable is that b_k, a coefficient of the *square*, is a sum of products $a_i a_j$ with $i + j = k$: this means the sign of b_k is not a free variable, it is determined by the signs of $a_0, \ldots, a_k$, and in particular by $2a_0a_k$ once the sign of a_0 is pinned down. That coupling between "sign of b_k" and "sign of a_k" is exactly why the induction on a_i's signs is the natural device: it is not an auxiliary trick, it is reading the recursive structure that $b_k = \sum a_i a_j$ already hands you.

The contradiction lands specifically at $b_{2n-1} = 2a_{n-1}a_n$ and nowhere else because this is the *only* coefficient of $p(x)^2$ expressible as a product of just two of the "high-index" a_i's with no lower-index cross terms diluting it – it is the coefficient immediately below the forced-nonnegative $b_{2n} = a_n^2$, so by the time the induction reaches it, both factors a_{n-1} and a_n have already been pinned negative, and a bare product of two negatives with nothing else added in has no room left to be negative itself. This is the same "of course" as the Blueprint's circle: once you see b_{2n-1} is structurally just $2a_{n-1}a_n$ – forced by which pairs (i, j) with $i + j = 2n-1$ exist inside a degree-n polynomial (only $(n-1, n)$ and $(n, n-1)$) – its sign is not a coincidence to be checked at the end, it is the whole reason the induction was worth running that far.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that $b_0 = a_0^2$ and $b_{2n} = a_n^2$ are pure squares, hence automatically nonnegative – this should be written down the instant $p(x)^2$ is expanded, and it immediately caps the answer at $2n-2$ before any construction or induction is attempted.
2. That cue suggests attempting to show the cap is tight both ways: an impossibility proof that $2n-1$ negative coefficients cannot occur, by assuming they do and inducting on the signs of $a_1, \ldots, a_n$; and a construction achieving $2n-2$.

3. Hesitation occurs at the induction step $b_k = 2a_0a_k + \sum_{i=1}^{k-1} a_i a_{k-i}$: it is not obvious at first that the sum term is forced positive rather than of unknown sign. It is reinforced away by noticing every index pair $(i, k-i)$ inside that sum has *both* indices already covered by the inductive hypothesis, so both factors are already known negative.

4. The trap is trying to prove the bound by construction alone, or by checking small cases ($n = 2, 3$) and conjecturing the pattern without a mechanism for why $2n-1$ fails universally; a finite search never rules out a construction for large n that the searcher simply hasn't found.

5. Recognition would be faster next time by treating "coefficients of a product (or square) satisfy a convolution identity $b_k = \sum_{i+j=k} a_i a_j$" as an automatic trigger for "the sign of b_k is not independent data – it's determined by the signs of the a_i, so an extremal sign-pattern assumption on the b_k should be pushed back onto an induction on the a_i."

6. The reusable decision rule: when a problem bounds how many coefficients of a polynomial product (especially a square) can carry a given sign, do not search for extremal examples first. Assume the extremal sign pattern holds for every candidate coefficient, use the convolution formula for those coefficients to induct on the signs of the original polynomial's coefficients (normalising one sign via the symmetry of the operation, here $p \mapsto p^2$ erasing global sign), and look for the single coefficient – typically nearest the ends, where fewest cross terms dilute it – at which the induction's forced signs collide with the required sign and produce a contradiction.

2022 A3

2022 A3. Let p be a prime number greater than 5. Let $f(p)$ denote the number of infinite sequences $a_1, a_2, a_3, \ldots$ such that $a_n \in \{1, 2, \ldots, p-1\}$ and $a_n a_{n+2} \equiv 1 + a_{n+1} \pmod{p}$ for all $n \geq 1$. Prove that $f(p)$ is congruent to 0 or 2 $\pmod 5$.

Fields: Number theory (recurrences mod p, periodicity, quadratic residues) **Position:** A3 – the first proof problem of the paper **Answer:** $f(p) = (p-2)(p-3)$, and this product is always $\equiv 0$ or $2 \pmod 5$ because $p \not\equiv 0 \pmod 5$ (as $p > 5$ is prime) forces $(p-2)(p-3)$ to be one of only four residues mod 5, and those four collapse to just $\{0, 2\}$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The recurrence $a_n a_{n+2} \equiv 1 + a_{n+1} \pmod{p}$ looks like it needs three consecutive terms to determine the next one, so a first instinct is to treat the sequence as genuinely infinite-dimensional: pick $a_1, a_2, a_3, \ldots$ term by term subject to the constraint, and try to count valid infinite sequences directly, perhaps via a transfer-matrix or graph-counting argument on residues mod p (build a directed graph on pairs (a_n, a_{n+1}) with an edge to (a_{n+1}, a_{n+2}) whenever the recurrence is satisfied, and count infinite walks). This is not wrong in spirit – infinite sequences satisfying a two-term-lookback recurrence really do correspond to infinite walks in such a graph – but it is enormously more work than the problem needs, because "infinite walk" typically forces you to reason about cycles in a graph with up to $(p-1)^2$ vertices, and the mod-5 congruence you're asked to prove gives no visible foothold inside that graph-theoretic language at all. Somewhere in this approach you would need to notice that the recurrence, run forward from any starting pair, either enters a genuine infinite consistent chain or dies (some a_n would have to be 0 mod p, which is forbidden) – and that observation, once made honestly, is really asking whether the sequence generated by (a_1, a_2) is eventually periodic with all terms staying nonzero. The natural attempt is correct that this is the crux, but it arrives there by the long route of graph-counting rather than by computing.

The model chain. View the terms as living in $\mathbb{F}_p^\times$. The recurrence rearranges to $a_{n+2} = (1 + a_{n+1})/a_n$, so once a_1, a_2 are fixed and nonzero, every later term is determined – the sequence is

a genuine function of (a_1, a_2), not a free choice at each step. Compute directly:

$$a_3 = \frac{1+a_2}{a_1}, \quad a_4 = \frac{a_1+1+a_2}{a_1 a_2}, \quad a_5 = \frac{1+a_1}{a_2},$$

(the last using the identity $a_1a_2 + a_1 + a_2 + 1 = (a_1+1)(a_2+1)$ to cancel a factor of a_2+1), and then

$$a_6 = \frac{1+a_5}{a_4} = a_1, \qquad a_7 = \frac{1+a_6}{a_5} = a_2.$$

So the sequence is periodic with period exactly 5 (on the values $a_1, \ldots, a_5$), for *any* valid starting pair. A pair $(a_1, a_2) \in (\{1, \ldots, p-1\})^2$ generates a valid infinite sequence exactly when none of the five terms $a_1, \ldots, a_5$ is 0 in $\mathbb{F}_p$ (division by zero would otherwise appear on the next step, and by periodicity that is the only way the sequence can ever fail to continue). $a_1 \neq 0$ and $a_2 \neq 0$ are given; $a_3 \neq 0$ forces $a_2 \neq -1 \equiv p-1$; $a_5 \neq 0$ forces $a_1 \neq p-1$; and $a_4 \neq 0$ forces $a_1 + a_2 \not\equiv -1 \equiv p-1 \pmod{p}$, which, since $a_1, a_2 \leq p-2$ makes $a_1 + a_2 \leq 2p-6 < 2p-1$, means simply $a_1 + a_2 \neq p-1$ as ordinary integers.

So $f(p)$ counts pairs $(a_1, a_2) \in \{1, \ldots, p-2\}^2$ with $a_1 + a_2 \neq p-1$. There are $(p-2)^2$ pairs total; among these, $a_1 + a_2 = p-1$ has exactly $p-2$ solutions (one for each $a_1 \in \{1, \ldots, p-2\}$, giving $a_2 = p-1-a_1 \in \{1, \ldots, p-2\}$ automatically). Hence

$$f(p) = (p-2)^2 - (p-2) = (p-2)(p-3).$$

Since $p > 5$ is prime, $p \not\equiv 0 \pmod{5}$, so $p \bmod 5 \in \{1, 2, 3, 4\}$, giving $(p-2)(p-3) \bmod 5 \in \{4 \cdot 3, 0 \cdot 4, 1 \cdot 0, 2 \cdot 1\} = \{12, 0, 0, 2\} \equiv \{2, 0, 0, 2\} \pmod{5}$. In every case $f(p) \equiv 0$ or $2 \pmod{5}$, as required.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both approaches agree that the recurrence, once you fix a_1, a_2, generates everything else – that much is forced by the problem's own definition. The split is in what you do with that fact. "Both paths recognise the sequence is determined by its first two terms. At that point the natural attempt keeps the state space abstract – treat (a_n, a_{n+1}) as a vertex in a graph over all of $\mathbb{F}_p^\times \times \mathbb{F}_p^\times$ and reason about which walks survive forever; the model instead just runs the recursion by hand for five steps and discovers, concretely, that it returns to where it started."

> *I don't need to reason abstractly about which walks in a huge graph are infinite – I should just compute a_3, a_4, a_5, a_6, a_7 symbolically in terms of a_1 and a_2 and see what happens. If the algebra collapses back to a_1, a_2 after a handful of steps, the entire infinite-sequence question becomes a finite, checkable condition on the single pair (a_1, a_2), and "infinite walk that survives forever" becomes nothing more than "finite cycle avoiding zero."*

> *"Infinite sequence, never hitting a forbidden value" is really just "finite period, and the finitely many values in one period all avoid the forbidden value" – once periodicity is established, I no longer need to think about the sequence as infinite at all; counting valid sequences is now counting valid* starting pairs (a_1, a_2) *subject to finitely many explicit inequalities, which is a direct counting problem in $\{1, \ldots, p-2\}^2$.*

> *The target congruence is mod 5, and I have just discovered the recurrence has period 5 – that is too specific a coincidence not to be the entire point of the problem; the number $f(p)$ being congruent to something mod 5 should be a shadow of that periodicity, and indeed the closed form $(p-2)(p-3)$ only needs the single fact $p \not\equiv 0 \pmod{5}$ (since $p > 5$ is prime) to force it into just two of the five*

possible residues.

▷ **Cue:** a sequence is defined by a two-term-lookback recurrence and a problem asks about infinitely many valid instances ⟶ **Move:** do not reason about the state space abstractly (graphs, transfer matrices); compute the first several terms symbolically in terms of the initial data and check directly whether the sequence is eventually, or immediately, periodic – this turns an infinite counting problem into a finite one

▷ **Cue:** a divisibility or congruence target (here, mod 5) shows up alongside a recurrence ⟶ **Move:** suspect the recurrence itself has a period equal to (or dividing) that same modulus, and look for it by direct computation before searching for any other structural explanation

Layer 3 THE WHY

Why must the abstract graph-walk framing give way to "just compute five terms"? Because the recurrence $a_{n+2} = (1 + a_{n+1})/a_n$ is an explicit rational function of two variables with no external randomness – the entire orbit of any starting pair under this map is, by definition, a deterministic sequence of rational expressions in a_1, a_2 over $\mathbb{F}_p$. A deterministic orbit of a rational map on a *finite* field cannot avoid eventually repeating (the field has only $p - 1$ nonzero elements, so some pair of consecutive terms must recur), and once a state recurs the whole sequence from that point on is forced to repeat too. So periodicity was never in question – the only genuine question was *what* the period is and whether it is the same for every starting pair, and that is exactly the kind of fact that symbolic computation settles cheaply, because the recursion is only two steps deep and the algebra telescopes (the identity $a_1a_2 + a_1 + a_2 + 1 = (a_1 + 1)(a_2 + 1)$ is the one place it needs to be pushed, and it is forced by the shape of a_4's numerator, not invented).

The specific value 5 is not a coincidence to be verified after the fact; it is read directly off the problem's request to prove something mod 5. A solver who has just computed that the period is exactly 5 has, in that instant, been handed the entire reason the answer must reduce to a statement about residues mod 5: $f(p)$ counts starting pairs, which is a polynomial in p, and the only remaining content is how that specific polynomial, $(p - 2)(p - 3)$, behaves under reduction mod the same 5 that showed up as the period. This is the Blueprint's "centre of the circle" standard exactly: once the period is seen to be 5, asking about $f(p)$ mod 5 is not a trick question layered on top of a number-theory problem – it is the only question left to ask, because everything else about $f(p)$ (its closed form as $(p - 2)(p - 3)$) has already been settled by finite counting, and the mod-5 condition on p follows immediately from $p > 5$ being prime, hence coprime to 5.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that the recurrence $a_{n+2} = (1 + a_{n+1})/a_n$ is fully determined by (a_1, a_2) once rearranged out of the given multiplicative form – that rearrangement should be written down and acted on immediately, rather than left in the $a_n a_{n+2} \equiv 1 + a_{n+1}$ form.

2. That cue suggests direct symbolic computation of $a_3, a_4, a_5, \ldots$ in terms of a_1, a_2, rather than any abstract state-space or graph-theoretic argument about which starting configurations survive forever.

3. Hesitation occurs at $a_4 = \frac{a_1+1+a_2}{a_1a_2}$ feeding into a_5: the algebra looks like it is about to get worse, not better, and it is tempting to stop computing. It is reinforced away by pushing through one more step and noticing the numerator factors as $(a_1 + 1)(a_2 + 1)$, which cancels cleanly against the denominator inherited from a_3.

4. The trap is treating "count infinite sequences satisfying a recurrence" as requiring genuinely infinite reasoning (graphs, generating functions over infinite walks, or an inductive argument about all n) rather than recognising that a deterministic recurrence over a finite field must be periodic, collapsing the count to a finite, explicit condition on the initial pair.

5. Recognition would be faster next time by treating "a deterministic two-term recurrence over a finite set" as an automatic trigger for "compute forward until it visibly returns to the start;

the period, once found, is itself the answer to why any modulus appearing later in the problem was chosen."

6. The reusable decision rule: when counting infinite sequences generated by a deterministic recurrence over a finite alphabet (here $\mathbb{F}_p^\times$), first compute enough terms by hand to detect periodicity – it is guaranteed to exist by finiteness, so the only work is finding its length. Once found, the infinite counting problem reduces to counting starting states within one period that avoid whatever value the recurrence forbids (here, division by zero), turning an "infinitely many valid sequences" question into an explicit finite formula.

2022 A4

2022 A4. Suppose that $X_1, X_2, \ldots$ are real numbers between 0 and 1 that are chosen independently and uniformly at random. Let $S = \sum_{i=1}^{k} X_i/2^i$, where k is the least positive integer such that $X_k < X_{k+1}$, or $k = \infty$ if there is no such integer. Find the expected value of S.

Fields: probability: expectation of a random-length sum, order statistics of uniform variables, exponential generating series **Position:** A4 — the pivot of the paper. The computation is within reach of a good Further Maths student; the packaging is what stops people. **Answer:** $\mathbb{E}[S] = 2e^{1/2} - 3 \approx 0.29744$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The stopping rule reads like a competition trick, so the first instinct is to get control of k. Find the distribution of k, then condition on it.

That is a reasonable plan and it starts well. The event $k = j$ says: $X_1 \geq X_2 \geq \cdots \geq X_j$ and $X_j < X_{j+1}$. The event $k \geq j$ is cleaner — it says only $X_1 \geq X_2 \geq \cdots \geq X_j$, and since the X_i are i.i.d. continuous, all $j!$ orderings of $X_1, \ldots, X_j$ are equally likely, so

$$\Pr(k \geq j) = \frac{1}{j!}, \qquad \Pr(k = j) = \frac{1}{j!} - \frac{1}{(j+1)!} = \frac{j}{(j+1)!}.$$

(As a check, $\sum_{j\geq 1} j/(j+1)! = 1$, and $\Pr(k = \infty) = \lim 1/j! = 0$, so the $k = \infty$ clause is a formality.)

Now you write

$$\mathbb{E}[S] = \sum_{j=1}^{\infty} \Pr(k = j)\, \mathbb{E}\left[\sum_{i=1}^{j} \frac{X_i}{2^i} \,\middle|\, k = j \right]$$

and the wall arrives. Conditional on $k = j$, the variables $X_1, \ldots, X_j$ are the order statistics of j uniforms in decreasing order, *further* conditioned on being below X_{j+1}. So $\mathbb{E}[X_i \mid k = j]$ is the expectation of the i-th largest of j uniforms, tilted by the extra condition at the bottom. You are now carrying a double sum over i and j with Beta-function entries in it. It can be pushed through — but you are doing an exam problem with a two-parameter family of integrals, and the $\frac{1}{2^i}$ weights refuse to collapse until the very end.

This is zombie maths of the structural kind. Nothing is wrong. Everything is expensive. The cost was incurred at the first line, by deciding to condition on k.

The model chain. Do not condition on k. Remove it from the problem.

Extend S to an infinite sum by defining

$$S = \sum_{i=1}^{\infty} \frac{X_i}{2^i} \mathbf{1}\{k \geq i\},$$

which is the same random variable: the indicator switches the tail off at exactly the right place. Every summand is nonnegative, so Tonelli's theorem lets me swap expectation and sum with

no convergence worry:

$$\mathbb{E}[S] = \sum_{i=1}^{\infty} \frac{1}{2^i}\, \mathbb{E}\big[X_i\, \mathbf{1}\{k \geq i\}\big].$$

Now the only event in sight is $\{k \geq i\} = \{X_1 \geq X_2 \geq \cdots \geq X_i\}$, and I need one expectation, not a family of conditional ones.

Condition on the value of X_i alone. Given $X_i = t$, the event $X_1 \geq \cdots \geq X_{i-1} \geq t$ is the intersection of two independent events:

- all of $X_1, \ldots, X_{i-1}$ lie in $[t, 1]$, probability $(1-t)^{i-1}$;
- those $i-1$ values, whatever they are, happen to be arranged in nonincreasing order, probability $1/(i-1)!$.

These are independent because the set of values taken by $X_1, \ldots, X_{i-1}$ and the order in which they appear are independent for i.i.d. continuous variables. Hence

$$\mathbb{E}\big[X_i \mathbf{1}\{k \geq i\}\big] = \frac{1}{(i-1)!} \int_0^1 t(1-t)^{i-1}\, dt.$$

The integral is $B(2, i) = \dfrac{1!\,(i-1)!}{(i+1)!}$, most quickly seen by writing $t(1-t)^{i-1} = (1-t)^{i-1} - (1-t)^i$ and integrating termwise:

$$\int_0^1 t(1-t)^{i-1} dt = \frac{1}{i} - \frac{1}{i+1} = \frac{1}{i(i+1)}.$$

So $\mathbb{E}[X_i \mathbf{1}\{k \geq i\}] = \dfrac{1}{(i-1)!\, i(i+1)} = \dfrac{1}{(i+1)!} \cdot \dfrac{1}{1} \cdot \dfrac{(i+1)!}{(i-1)!\, i(i+1)}$; more directly, $(i-1)!\, i\, (i+1) = (i+1)!$, so the value is exactly $\dfrac{1}{(i+1)!}$.

Therefore

$$\mathbb{E}[S] = \sum_{i=1}^{\infty} \frac{1}{2^i (i+1)!} = 2 \sum_{i=1}^{\infty} \frac{1}{2^{i+1}(i+1)!} = 2 \sum_{j=2}^{\infty} \frac{(1/2)^j}{j!} = 2\left(e^{1/2} - 1 - \tfrac{1}{2}\right) = 2e^{1/2} - 3.$$

Numerically $2(1.648721\ldots) - 3 = 0.297442\ldots$, comfortably inside $[0, 1]$ and comfortably below $\mathbb{E}[X_1/2] = 1/4 +$ (a little), which is the right order of magnitude.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree on the first move: understand the event $\{k \geq j\} = \{X_1 \geq \cdots \geq X_j\}$ and its probability $1/j!$. They split at (†), the choice of how to decompose $\mathbb{E}[S]$. At (†) the natural attempt decomposes *by the value of k* — partitioning the sample space — while the model decomposes *by the index i of the summand* — partitioning the sum. Conditioning on k makes k into a parameter you must then carry through every integral. Summing over i makes k vanish into a single indicator.

> *S has a random number of terms, and that is the only hard thing about it. So my first job is not to compute anything — it is to turn the random-length sum into a fixed-length (infinite) sum with indicators, because linearity of expectation does not care how many terms are actually nonzero.*

> *Once I have $\mathbb{E}[S] = \sum_i 2^{-i} \mathbb{E}[X_i \mathbf{1}\{k \geq i\}]$, I should notice that $\{k \geq i\}$ is a much cheaper event than $\{k = i\}$: it is one chain of inequalities with no "and then it breaks" clause at the end. Whenever a stopping time appears, tail events $\{k \geq i\}$ are the cheap ones and I should engineer the algebra so*

> *that only they appear.*

> *To compute* $\mathbb{E}[X_i \mathbf{1}\{X_1 \geq \cdots \geq X_i\}]$ *I condition on* $X_i = t$ *and then split the remaining requirement into two logically independent halves:* where *the other values sit (a probability* $(1-t)^{i-1}$*) and* what order *they arrive in (a probability* $1/(i-1)!$*). Position and order are independent for i.i.d. continuous variables — that is the fact I am really using, and it is worth stating out loud.*

Layer 3 THE WHY

Why is the indicator reformulation not a trick? Because of what linearity of expectation actually says. $\mathbb{E}[A+B] = \mathbb{E}[A] + \mathbb{E}[B]$ holds with no independence hypothesis and no hypothesis about how A and B were produced. It is the single most robust tool in probability. But it only applies to a sum with a *fixed* list of summands. A sum whose length is itself random is not, syntactically, a sum of random variables — it is a random variable defined by cases. So the whole difficulty of this problem is a type mismatch between the object you are given and the tool you want to use, and the indicator is the type conversion. Writing $X_i \mathbf{1}\{k \geq i\}$ costs nothing and makes the list of summands fixed (it is $i = 1, 2, 3, \ldots$, always) while keeping the value identical. Once you see the manoeuvre as *conversion to the domain of linearity* rather than as a clever rewrite, it stops being something you have to think of and becomes something you reach for.

There is a second axiom-level fact doing work here, and it is the one the official solution states in a single clause — "these two events are independent and occur with respective probabilities $(1-X_i)^{i-1}$ and $\frac{1}{(i-1)!}$" — without saying where the independence comes from. Here is where that comes from. For i.i.d. continuous random variables $Y_1, \ldots, Y_m$, the *unordered set* of values $\{Y_1, \ldots, Y_m\}$ and the *permutation* describing which index got which rank are independent, and the permutation is uniform on S_m. The reason is exchangeability: the joint density of $(Y_1, \ldots, Y_m)$ is symmetric under permuting coordinates, so conditional on any particular set of m distinct values, all $m!$ assignments of those values to the indices are equally likely. That is the entire content, and it is a statement about a symmetry of the density, not a computation. So "the values all lie above t" (a statement about the set) and "they are in decreasing order" (a statement about the permutation) are independent by construction. The same fact is what gives $\Pr(k \geq j) = 1/j!$ in the first place; the model chain simply uses it twice.

Notice, finally, why the answer has an $e^{1/2}$ in it, since that is the kind of thing that looks like magic if you do not name it. Each summand contributed $\frac{1}{2^i(i+1)!}$: the 2^{-i} came from the problem's weights, the $(i+1)!$ from the ordering constraint on $i-1$ variables plus one Beta integral. A factorial in the denominator with a geometric weight on top is an exponential generating function evaluated at the weight. The series $\sum_j x^j/j!$ is e^x; here $x = 1/2$ and two terms ($j = 0, 1$) are missing, so the answer is $2(e^{1/2} - 1 - \frac{1}{2})$. The e is not a surprise ending. It is what a factorial denominator always produces, and seeing $1/(i+1)!$ appear should make you write $e^{1/2}$ down before you have finished the sum.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the phrase "k is the least positive integer such that $X_k < X_{k+1}$". On seeing it, write down at once: $\{k \geq j\} = \{X_1 \geq X_2 \geq \cdots \geq X_j\}$, and $\Pr(k \geq j) = 1/j!$ by exchangeability. "Least index such that" always means the tail event is the simple one.

2. That cue, combined with "S is a sum with a random number of terms", suggests the method: rewrite S as an infinite sum with indicators and apply linearity of expectation term by term, justifying the interchange by nonnegativity (Tonelli).

3. Hesitation occurs at the moment you have $\Pr(k = j) = j/(j+1)!$ in hand and feel you have made progress. That formula is genuinely correct and genuinely a dead end, which is the worst combination. What reinforces it away is asking a single question before using it: *do*

I need the distribution of k, or do I only need $\mathbb{E}[S]$? You only need $\mathbb{E}[S]$, and expectations of sums never require the law of the length.

4. Two traps. (i) The $k = \infty$ clause, which invites a long discussion of an event of probability zero; note $\Pr(k \geq j) = 1/j! \to 0$ in one line and move on. (ii) The seduction of $\Pr(k = j)$: it is easy to compute, so it feels like the intended route, and it commits you to conditional order statistics. The problem is engineered so that the accessible quantity is the wrong one.

5. Recognisable faster next time by the shape alone: *a sum whose upper limit is a random variable*. That shape has one standard treatment and you should not deliberate. $\sum_{i=1}^{N} Z_i = \sum_{i=1}^{\infty} Z_i \mathbf{1}\{N \geq i\}$, then swap. The same conversion handles $\mathbb{E}[N] = \sum_{i \geq 1} \Pr(N \geq i)$, which is the special case $Z_i \equiv 1$ — so if you already know that identity, you already know this technique and have only been using it in one dimension.

6. The reusable decision rule: **when a random object has a random size, convert the size into indicators inside a fixed-shape expression rather than conditioning on it.** Conditioning on the size makes it a parameter that propagates into every subsequent integral; indicators make it an event that can be estimated once. More generally — and this transfers well beyond probability — *prefer decompositions that partition the expression over decompositions that partition the sample space*, because linearity is free and conditioning is not.

▷ **Cue:** a sum $\sum_{i=1}^{k}$ whose upper limit k is itself a random variable $\longrightarrow$ **Move:** rewrite as $\sum_{i=1}^{\infty}(\cdot)\mathbf{1}\{k \geq i\}$ and take expectations term by term; justify the swap by nonnegativity

▷ **Cue:** k is defined as the least index at which some condition first holds $\longrightarrow$ **Move:** work with the tail event $\{k \geq i\}$, which is the condition failing for all earlier indices — never with $\{k = i\}$, which carries an extra clause

▷ **Cue:** a probability requires i.i.d. continuous variables to lie in a region *and* be in a prescribed order $\longrightarrow$ **Move:** split into (values in the region) $\times$ (order is the right one), probabilities multiply, and the order factor is $1/m!$ by exchangeability

▷ **Cue:** $\int_0^1 t(1-t)^m\,dt$ or similar Beta integrals appear $\longrightarrow$ **Move:** expand as $(1-t)^m - (1-t)^{m+1}$ and integrate termwise rather than invoking the Beta function

▷ **Cue:** the general term of a series contains a factorial in the denominator and a geometric factor x^i $\longrightarrow$ **Move:** expect an exponential: match against $\sum x^j/j! = e^x$ and identify which low-order terms are missing

2022 A5

2022 A5. Alice and Bob play a game on a board consisting of one row of 2022 consecutive squares. They take turns placing tiles that cover two adjacent squares, with Alice going first. By rule, a tile must not cover a square that is already covered by another tile. The game ends when no tile can be placed according to this rule. Alice's goal is to maximize the number of uncovered squares when the game ends; Bob's goal is to minimize it. What is the greatest number of uncovered squares that Alice can ensure at the end of the game, no matter how Bob plays?

Fields: combinatorial game theory: partizan scoring game, position decomposition, parity invariant, eventual periodicity **Position:** A5 — the hardest problem on the paper for most solvers, because the correct answer requires a full induction on a two-valued function and the period-7 behaviour cannot be guessed from the first few cases **Answer:** 290.

Layer 1 THE WHAT — surface analysis

The natural attempt. Everybody starts the same way, and correctly: play small boards by hand.

A position is a collection of disjoint gaps — maximal runs of uncovered squares — because a tile can only be placed inside one gap, and placing it inside a gap of length L at offset j replaces that gap by two gaps of lengths j and $L-2-j$ ($0 \leq j \leq L-2$). Gaps of length 0 or 1 are dead.

The game ends when every gap is dead, and the final score is the number of surviving length-1 gaps.

So you write $a(n)$ for the score with optimal play on a single gap of length n with Alice to move, and $b(n)$ for the same with Bob to move, and you grind:

n	0	1	2	3	4	5	6
$a(n)$	0	1	0	1	2	1	2
$b(n)$	0	1	0	1	0	1	0

$n = 4$ is the first interesting entry: Alice plays the middle tile, leaving gaps 1 and 1, scoring 2; Bob plays at one end, leaving a gap of 2 which Alice must fill, scoring 0.

Then you reach $n = 7$ and the plan breaks. A position of 7 squares splits into two gaps after one move, so you now need values for *pairs* of gaps, and a pair of gaps is not the sum of two independent games in any sense you can quote — the players choose which component to play in, and that choice is itself strategic. You try to guess: maybe the answer is about n mod 4? $a(n)$ for $n = 0, \ldots, 6$ is $0, 1, 0, 1, 2, 1, 2$, which looks like it might be settling down, so you extrapolate a period of 4 or of 2, get $a(2022) = 0$ or 2 or 1011, and write something down.

The wall is real and it is worth naming precisely: *you cannot compute $a(8)$ without a theory of two-component positions*, and there is no way round that by cleverness at the board.

The model chain. The answer is 290.

Step 0: reformulate the score. If the game ends after t tiles, the number of uncovered squares is $n - 2t$. So Alice, the maximiser, is trying to *minimise* the number of tiles played, and Bob is trying to maximise it. Equivalently, and more usefully: at the end every gap has length 0 or 1, so the score is the number of odd gaps at the end. And the number of odd gaps never decreases. Indeed, splitting a gap of length L into j and $L - 2 - j$: if L is odd then exactly one of $j, L - 2 - j$ is odd, so the count is unchanged; if L is even then j and $L - 2 - j$ have equal parity, so the count rises by 0 or 2. Hence

$$a(n),\, b(n) \geq n - 2\lfloor n/2 \rfloor ,$$

which is 1 for odd n and 0 for even n. Checking the possible endings gives equality for $n = 0, 1, 2, 3, 5$. For $n = 4$ and $n = 6$ Alice can split into two odd gaps in one move ($4 \to 1 + 1$, $6 \to 1 + 3$), guaranteeing at least two odd gaps forever, so $a(4) = a(6) = 2$; and Bob can move to leave only gaps of length 2 ($4 \to 2 + 0$, $6 \to 2 + 2$), guaranteeing no odd gaps, so $b(4) = b(6) = 0$. That is the table above, and it is proved, not tabulated.

Step 1: the decomposition lemma. Let $A(P)$ and $B(P)$ denote the optimal scores from a position P (a multiset of gap lengths) with Alice, respectively Bob, to move. The lemma that makes multi-gap positions tractable is

$$A(P) \geq B(P) \quad \text{for every position } P :$$

having the move is never a disadvantage to the maximiser. Granting it, neither player ever gains by passing — Alice passing converts $A(P)$ into $B(P) \leq A(P)$, and Bob passing converts $B(P)$ into $A(P) \geq B(P)$ — so we may allow passing (after the first move) without changing any value. With passing available, a player facing a two-component position $\{i, n - 2 - i\}$ can adopt a *mirroring* strategy: choose one component to open in, and thereafter always reply in whichever component the opponent did not just play in, passing if that component is finished. Mirroring makes the two components independent, each with a determined first mover. This yields

$$a(n) = \max_{0 \leq i \leq n-2} \min \big\{ a(i) + b(n-2-i),\; b(i) + a(n-2-i) \big\},$$
$$b(n) = \min_{0 \leq i \leq n-2} \max \big\{ a(i) + b(n-2-i),\; b(i) + a(n-2-i) \big\}.$$

The outer max/min is the mover choosing the split; the inner min/max is the opponent choosing which component to open in.

Step 2: the recursion. From these, for $n \geq 7$,

$$a(n) = a(n-7) + 1, \qquad b(n) = b(n-7) + 1.$$

Lower bound for a: Alice splits n into 1 and $n-3$ (placing her tile on squares 2, 3). The singleton survives to the end, and Bob moves first on the rest, so $a(n) \geq 1 + b(n-3)$; and the table extends to give $b(m) = a(m-4)$ for all $m \geq 4$, so $1 + b(n-3) = 1 + a(n-7)$. Upper bound for a: any split $\{i, n-2-i\}$ is answered by Bob's mirroring, giving the max min formula, whose value is $a(n-7)+1$. Upper bound for b: Bob splits n into 2 and $n-4$; the 2 contributes nothing and merely consumes a tempo, so $b(n) \leq a(n-4) = b(n-7)+1$. Lower bound for b: Alice mirrors, giving the min max formula.

Step 3: evaluate. $2022 = 7 \cdot 288 + 6$, so

$$a(2022) = 288 + a(6) = 288 + 2 = 290.$$

Caveat. I have verified the values $a(n), b(n)$ for $0 \leq n \leq 28$ by exhaustive game-tree search over multiset positions, confirming both the period-7 recursion and the max min/min max decomposition formulas exactly in that range; I also verified $A(P) \geq B(P)$ over all 1002 positions with total length at most 22, so there is no zugzwang in this game in that range. The one step I have *not* verified in complete generality, and which the official solution also passes over in a clause, is $A(P) \geq B(P)$ for all positions — the induction is on total length and is routine to believe but is genuinely a hypothesis of the passing argument. Do not present the mirroring strategy without stating that hypothesis; it is exactly the kind of step where a scoring game can surprise you.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree through the hand computation of $a(n), b(n)$ for $n \leq 6$. They split at $n = 7$, at the moment the first move produces two gaps. At (†) the natural attempt *tries to guess the pattern from the six known values*; the model *stops computing and builds the theory of two-component positions*, namely the passing lemma and the mirroring strategy. The split is not about cleverness. It is about recognising that the sequence cannot be continued at all until multi-component positions are understood, so guessing is not a shortcut — it is the only thing left when you have refused to build the tool.

> *After one move the board is two independent-looking gaps, and I have no way to evaluate that. So the real question is not "what is $a(7)$" but "when does a two-gap position split into the sum of two one-gap positions?" It splits exactly when I can force the turn order in each component, and I can force it if I am allowed to respond in the component my opponent did not touch.*

> *Mirroring fails if the component I am supposed to reply in has just run out of legal moves. So before I claim the mirroring strategy I must buy myself a pass. A pass is free precisely when having the move is never bad, i.e. when $A(P) \geq B(P)$ — so let me make that an explicit induction hypothesis rather than an assumption.*

> *The score is n minus twice the number of tiles, so Alice is really trying to end the game* early *and Bob to prolong it. That reframing tells me what each player's move should look like: Alice wants to*

> *manufacture length-1 gaps, which are dead and scoring; Bob wants to manufacture length-2 gaps, which are alive, worth nothing, and cost a tempo.*

> *The number of odd gaps never decreases — splitting an odd gap leaves exactly one odd child, splitting an even gap leaves zero or two. So every odd gap I create is money in the bank that Bob can never take back. That is the invariant, and it is what makes the base cases provable rather than merely observed.*

Layer 3 THE WHY

Why must the passing lemma be thought? Because of what "optimal play on a disjoint union" means. In an impartial normal-play game, Sprague–Grundy gives you a genuine addition law: the value of a disjoint union is a function of the values of the parts, full stop. This game has neither of the hypotheses that theorem needs — it is partizan (Alice and Bob want different things) and it is scoring (the payoff is a number, not a win/loss). So there is *no* general reason for a two-gap position to decompose, and if you write $a(i) + b(n-2-i)$ without justification you have assumed the conclusion. What has to be supplied is a strategy that makes the components independent *in practice*: mirroring does that, because it guarantees that the moves in each component alternate correctly and with a known first mover. And mirroring needs exactly one thing the rules do not provide — the ability to do nothing when your assigned component is exhausted. Hence the pass. Hence the need to know that a pass is harmless. The chain is forced from end to end: *decomposition* $\Rightarrow$ *mirroring* $\Rightarrow$ *passing* $\Rightarrow$ *no zugzwang*.

Now the period 7, which looks arbitrary and is not. Look at what each player buys with one move. Alice's canonical move on a long gap splits n into 1 and $n-3$: she banks one point and hands Bob a board 3 shorter. Bob's canonical move splits n into 2 and $n-4$: he banks nothing and hands Alice a board 4 shorter, having spent one square of the board's length purely on tempo. One full round of the game therefore consumes $3+4=7$ squares and yields Alice exactly 1 point. *The 7 is the length of one Alice-move plus one Bob-move under their respective optimal openings, and the $+1$ is Alice's per-round income.* Once that is seen, the recursion $a(n) = a(n-7)+1$ is not a computational discovery but a description of the game's steady state, and the constant $290 = 288+2$ reads as "288 complete rounds, plus the endgame value of the 6 squares left over".

Kedlaya's solution opens by asserting the two recursions and the table of values for $n \le 6$, and proves them by induction. Here is where the table comes from: the parity invariant. The quantity "number of odd gaps" is monotone nondecreasing, which immediately gives the trivial lower bound n mod 2 and shows that the only way either player can move the score is by *splitting an even gap into two odd ones*. That single observation explains every entry: $n = 0, 1, 2, 3, 5$ have no room for a profitable even split (there is no even gap to split, or splitting it hands the opponent the correction), while $n = 4$ and $n = 6$ are exactly the small even lengths where Alice can split into two odds on move one. Bob's counter is to keep everything even — which is why $b(4) = b(6) = 0$ and why his structural move is to manufacture 2s. A gap of length 2 is Bob's ideal object: it is even, it will stay even, it scores nothing, and somebody has to spend a move on it.

The general lesson about periodicity is worth stating flatly. A combinatorial game whose positions are indexed by a single integer and whose moves subtract a bounded amount has a finite state space per residue class, so its value function is *eventually periodic*; that is a theorem-shaped fact, not a hope. What the solver has to find is the period and the pre-period, and the reliable way to find them is not to stare at six terms — it is to identify the optimal opening for each player and add up what one round costs.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is "2022 consecutive squares" with two opposed objectives. On seeing it write down: *position = multiset of gap lengths; a move on a gap of length L replaces it by*

j and $L-2-j$*; final score = number of length-*1 *gaps* $= n-2($*number of tiles*$)$. Also write $2022 = 7 \cdot 288 + 6$ later, but the reformulation comes first.

2. That cue suggests a two-valued recursion $a(n), b(n)$ over single gaps, plus a decomposition theory for multi-gap positions — the standard toolkit for a partizan scoring game on an interval. It also suggests hunting for a monotone invariant, which here is the number of odd gaps.

3. Hesitation occurs at $n = 7$, the first length whose analysis needs two-component positions. The temptation is to extrapolate from $0, 1, 0, 1, 2, 1, 2$. What reinforces it away is the honest question *can I compute the next term at all?* — the answer is no, and a sequence you cannot extend by one term is a sequence you have no business extrapolating. Build the mirroring lemma instead; it costs a paragraph and it unlocks every n.

4. Three traps. (i) Guessing a period of 2 or 4 from the base table, which is wrong and produces a confident wrong number. (ii) Using $a(i) + b(n-2-i)$ as though disjoint positions obviously add, which is exactly what is not obvious in a partizan scoring game. (iii) Forgetting that Alice maximising uncovered squares means Alice wants *fewer* moves played, which inverts every intuition trained on normal-play games where you want to move last.

5. Recognisable faster next time by the shape: *two players with opposite numerical objectives, a position that is a disjoint union, and a move that splits one component into two.* That shape always demands the same two questions in the same order — "is there a monotone invariant that bounds the score?" and "can I force the components to be independent by mirroring, and do I need a pass to do it?" Answer those before computing anything beyond the smallest cases.

6. The reusable decision rule: **when a game's value function is indexed by one integer, do not extrapolate the table — compute the cost of one optimal round and read the period off it.** Alice's best opening consumes 3 squares, Bob's consumes 4, so the period is 7 and the per-period gain is Alice's one banked point. The corollary rule: **a disjoint-union position decomposes only when someone can enforce the turn order in each part; the enabling technical fact is always "having the move is never a disadvantage", and it must be proved, not assumed.**

▷ **Cue:** a two-player game on a row of n cells where a move deletes a bounded block and splits the row $\longrightarrow$ **Move:** define two value functions (one per player to move), compute the base cases by hand, and expect eventual periodicity with period equal to the total consumption of one optimal round

▷ **Cue:** the players' objectives are "maximise/minimise the final count of something" $\longrightarrow$ **Move:** convert the score into a linear function of the number of moves (score $= n - 2t$ here) and restate who is trying to end the game early

▷ **Cue:** a position is a disjoint union of independent-looking components $\longrightarrow$ **Move:** do not assume values add; look for a mirroring strategy that fixes the first mover in each component, and check whether it needs a pass

▷ **Cue:** a mirroring or pairing strategy needs a pass to be legal $\longrightarrow$ **Move:** prove "having the move is never a disadvantage" for the relevant player, i.e. $A(P) \geq B(P)$ over all positions, as an explicit induction hypothesis

▷ **Cue:** you need a lower bound on a score that no subsequent play can destroy $\longrightarrow$ **Move:** find a quantity that is monotone under every legal move — here, the number of odd-length gaps — and bank it

▷ **Cue:** a small table of game values is tempting to extrapolate $\longrightarrow$ **Move:** refuse until you can compute the next unknown term honestly; if you cannot, the missing tool is the multi-component theory, not more patience

2022 A6

2022 A6. Let n be a positive integer. Determine, in terms of n, the largest integer m with the following property: There exist real numbers $x_1, \ldots, x_{2n}$ with $-1 < x_1 < x_2 < \cdots < x_{2n} < 1$ such that the sum of the lengths of the n intervals

$$[x_1^{2k-1}, x_2^{2k-1}], [x_3^{2k-1}, x_4^{2k-1}], \ldots, [x_{2n-1}^{2k-1}, x_{2n}^{2k-1}]$$

is equal to 1 for all integers k with $1 \leq k \leq m$.

Fields: algebra: symmetric functions, Newton's identities, roots of unity; analysis: odd power sums
Position: A6 — the hardest problem on the paper. Two genuinely separate tasks (a sharp construction and a sharp obstruction), and the obstruction needs a change of language before it can even be attacked
Answer: $m = n$.

Layer 1 THE WHAT — surface analysis

The natural attempt. Strip the geometry first — everyone does this and it is right. Since $t \mapsto t^{2k-1}$ is strictly increasing on $\mathbb{R}$, the interval $[x_{2i-1}^{2k-1}, x_{2i}^{2k-1}]$ has length $x_{2i}^{2k-1} - x_{2i-1}^{2k-1} > 0$, so the condition is

$$\sum_{j=1}^{2n} (-1)^j x_j^{2k-1} = 1 \qquad (k = 1, \ldots, m), \tag{$*$}$$

the signs alternating $-, +, -, +, \ldots$ ending with $+$ on x_{2n}.

Now count. Each k is one equation; there are $2n$ unknowns subject to the open conditions $-1 < x_1 < \cdots < x_{2n} < 1$. So $m = 2n$ ought to be attainable, or at worst m slightly less, and the answer is presumably $2n$ or $2n - 1$. This is the first place the problem bites: the naive dimension count is not just imprecise, it is off by a factor of two, and nothing in the count tells you why.

So you experiment. Take $n = 1$: two numbers $-1 < x_1 < x_2 < 1$ with $x_2 - x_1 = 1$, and then ask for $x_2^3 - x_1^3 = 1$ as well. With $x_2 = x_1 + 1$ this is $3x_1^2 + 3x_1 + 1 = 1$, so $x_1 \in \{0, -1\}$; $x_1 = 0$ gives $x_2 = 1$, excluded; $x_1 = -1$ excluded. So $m = 1$ when $n = 1$ — not 2. The equations are fighting each other in a way the count did not predict, and you have one data point and no mechanism.

Push to $n = 2$ by hand and you are solving a system of three symmetric-looking equations in four unknowns with no obvious structure, and the algebra is unbounded. This is zombie maths in its purest form: the equations $(*)$ are polynomial, so you start eliminating, and elimination on four unknowns with cubics and quintics is a place from which nobody returns inside three hours.

The model chain. The answer is $m = n$. Two independent halves.

Half 1: $m \leq n$. The obstruction. Introduce the polynomial

$$p(x) = (x + x_1)(x - x_2)(x + x_3)(x - x_4) \cdots (x + x_{2n-1})(x - x_{2n})\,(x + 1),$$

of degree $2n + 1$, whose roots are

$$-x_1,\ x_2,\ -x_3,\ x_4,\ \ldots,\ -x_{2n-1},\ x_{2n},\ -1.$$

Let s_r be the r-th power sum of these roots and e_r the r-th elementary symmetric function, so that

$$p(x) = x^{2n+1} + \sum_{r=1}^{2n+1} (-1)^r e_r x^{2n+1-r}.$$

For odd $r = 2k - 1$, the sign $(-1)^{2k-1} = -1$ passes through the odd power, so

$$s_{2k-1} = \sum_{i=1}^{n} \left(x_{2i}^{2k-1} - x_{2i-1}^{2k-1}\right) \ - \ 1.$$

Thus condition $(*)$ says exactly

$$s_1 = s_3 = \cdots = s_{2m-1} = 0.$$

That is the whole reformulation: the geometric condition is the vanishing of the odd power sums of a single polynomial's roots.

Now apply the Girard–Newton identities, $re_r = \sum_{i=1}^{r}(-1)^{i-1}e_{r-i}s_i$ for $1 \le r \le \deg p$. I claim $e_r = 0$ for every odd $r \le 2m-1$, by induction on odd r. For $r = 1$: $e_1 = s_1 = 0$. Suppose $e_1 = e_3 = \cdots = e_{r-2} = 0$ and let $r \le 2m-1$ be odd. In $re_r = \sum_{i=1}^{r}(-1)^{i-1}e_{r-i}s_i$, split by the parity of i: if i is odd then $i \le r \le 2m-1$, so $s_i = 0$; if i is even then $r-i$ is odd and $r-i < r$, so $e_{r-i} = 0$ by hypothesis. Every term dies, so $e_r = 0$.

Suppose now, for contradiction, that some choice of x_j achieves $m = n+1$. Then $s_1 = s_3 = \cdots = s_{2n+1} = 0$, so $e_r = 0$ for all odd $r \le 2n+1$. In the expansion of p, the term $(-1)^r e_r x^{2n+1-r}$ survives only for even r, and $2n+1-r$ is then odd. Hence p contains only odd powers of x: $p(-x) = -p(x)$ identically. But -1 is a root of p, so

$$p(1) = -p(-1) = 0.$$

On the other hand, evaluating directly,

$$p(1) = (1+x_1)(1-x_2)(1+x_3)\cdots(1-x_{2n})\cdot 2 > 0,$$

since $|x_j| < 1$ makes every factor strictly positive. Contradiction. Therefore $m \le n$.

Half 2: $m \ge n$**.** The construction. Take

$$x_j = \cos\frac{(2n+1-j)\pi}{2n+1} \qquad (j = 1, \ldots, 2n).$$

As j runs from 1 to $2n$ the angle $(2n+1-j)\pi/(2n+1)$ decreases strictly within $(0,\pi)$, where cos is strictly decreasing, so $-1 < x_1 < x_2 < \cdots < x_{2n} < 1$, as required.

Write $r = 2n+1-j$, so r runs over $1, \ldots, 2n$ and $(-1)^j = -(-1)^r$ (because $j+r$ is odd). The total interval length is

$$\sum_j (-1)^j x_j^{2k-1} = -\sum_{r=1}^{2n}\left((-1)^r \cos\frac{r\pi}{2n+1}\right)^{2k-1},$$

using that $2k-1$ is odd to pull the sign inside. Now

$$(-1)^r \cos\frac{r\pi}{2n+1} = \cos\left(r\pi + \frac{r\pi}{2n+1}\right) = \cos\frac{2\pi r(n+1)}{2n+1},$$

so with $\zeta = e^{2\pi i(n+1)/(2n+1)}$ — a *primitive* $(2n+1)$-th root of unity, since $\gcd(n+1, 2n+1) = \gcd(n+1, 2n+1-2(n+1)) = \gcd(n+1,-1) = 1$ — we get $(-1)^r \cos\frac{r\pi}{2n+1} = \frac{\zeta^r + \zeta^{-r}}{2}$ and hence

$$\sum_j (-1)^j x_j^{2k-1} = -\frac{1}{2^{2k-1}}\sum_{r=1}^{2n}\sum_{\ell=0}^{2k-1}\binom{2k-1}{\ell}\zeta^{r(2k-1-2\ell)}.$$

Swap the sums. For fixed ℓ, put $t = 2k-1-2\ell$, an *odd* integer with $|t| \le 2k-1 \le 2n-1 < 2n+1$; in particular $t \not\equiv 0 \pmod{2n+1}$, so ζ^t is a nontrivial $(2n+1)$-th root of unity and $\sum_{r=0}^{2n}\zeta^{rt} = 0$, giving $\sum_{r=1}^{2n}\zeta^{rt} = -1$. Therefore

$$\sum_j (-1)^j x_j^{2k-1} = -\frac{1}{2^{2k-1}}\sum_{\ell=0}^{2k-1}\binom{2k-1}{\ell}(-1) = \frac{2^{2k-1}}{2^{2k-1}} = 1,$$

for every k with $1 \le k \le n$. So $m \ge n$, and combining, $m = n$.

Note precisely where the construction stops working: at $k = n+1$ we have $2k-1 = 2n+1$, and the terms $\ell = 0$ and $\ell = 2k-1$ give $t = \pm(2n+1)$, so $\zeta^t = 1$ and those inner sums are $+2n$ rather than -1. The failure is not gradual; it is the exact moment the exponent reaches the order of the root of unity.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree through the first reduction — monotonicity of $t \mapsto t^{2k-1}$ turns the geometry into the alternating power-sum equations $(*)$. They split immediately after, at (†): how to regard $(*)$. The natural attempt reads $(*)$ as *a system of m equations in* $2n$ *unknowns* and starts eliminating. The model reads $(*)$ as *a statement about the power sums of the roots of one polynomial* and starts converting power sums into coefficients. The first reading treats the x_j as $2n$ separate numbers; the second treats them as one object.

> *The signs in* $\sum(-1)^j x_j^{2k-1}$ *are alternating, and* $2k-1$ *is odd, so* $(-1)^j x_j^{2k-1} = (\pm x_j)^{2k-1}$*: I can absorb every sign into the variable. That means the left-hand side is a genuine power sum* $\sum y_j^{2k-1}$ *of the* $2n$ *numbers* $y_j = (-1)^j x_j$ *— and the constant* 1 *on the right is* 1^{2k-1}*, so it can be absorbed too, as one more root.* Both sides of the equation are power sums of the same degree; I should put everything on one side and look at the multiset.

> *Once every condition is "*$s_{2k-1} = 0$*", the natural language is symmetric functions, and the dictionary between power sums and coefficients is Newton's identities. So the question becomes: what does the vanishing of the* odd *power sums say about the coefficients? Only odd* e_r *can be forced to vanish, because Newton's recursion mixes parities in exactly one way.*

> *If all the odd elementary symmetric functions of the roots vanish, the polynomial has only odd-degree terms — it is an odd function. An odd function's roots come in* $\pm$ *pairs. So I will get a contradiction the moment I can name a root whose negative cannot also be a root, and I built the number* -1 *into the polynomial for exactly that reason:* $+1$ *is not available, because every* $|x_j| < 1$*.*

> *For the construction I need* $2n$ *numbers in* $(-1, 1)$ *whose signed odd power sums are constant in k. Constancy in k of a power sum over many exponents is what roots of unity do —* $\sum_r \zeta^{rt} = -1$ *for every t not divisible by the order. So I should look for the* x_j *among cosines of rational multiples of* π*, and the denominator should be the number that bounds how far the exponents can go:* $2n+1$*.*

Layer 3 THE WHY

Why must the polynomial be introduced? Because of what a power sum *is*. The equations $(*)$ are symmetric in a strong sense: they do not see the x_j individually, only through the quantities $\sum y_j^r$ for odd r, where $y_j = (-1)^j x_j$. The fundamental theorem of symmetric functions says that any such data is equivalent to data about the elementary symmetric functions, i.e. about the coefficients of the monic polynomial with roots y_j. So there is no creativity in constructing p; *the problem's own symmetry names the object*. The only genuine decisions are (i) to include -1 as an extra root so that the inhomogeneous right-hand side becomes homogeneous, and (ii) to choose the sign convention so that the required statement is the clean one, $s_{2k-1} = 0$, rather than $s_{2k-1} = -2$.

And that decision (i) is itself forced. An equation $\sum y_j^{2k-1} = 1$ is not a statement about the multiset $\{y_j\}$ alone; it has a constant in it, and constants block the symmetric-function machinery. But $1 = 1^{2k-1}$ for every k, so the constant is itself a power sum of a single number. Moving it across gives $\sum_j y_j^{2k-1} + (-1)^{2k-1} = 0$, i.e. the power sums of the enlarged multiset

$\{y_1, \ldots, y_{2n}, -1\}$ vanish. *A constant that happens to equal c^r for the relevant exponents r is not a constant — it is an extra root.* That is a technique with a life far beyond this problem.

Now, why does exactly n come out, and not $2n$? Here is the mechanism, in the form the naive dimension count missed. Newton's identities let odd power sums control only odd elementary symmetric functions; the even e_r are never touched. The degree of p is $2n+1$, so it has $n+1$ odd-indexed coefficients $e_1, e_3, \ldots, e_{2n+1}$ available to be killed. Requiring m odd power sums to vanish kills the first m of them. When $m = n+1$ you have killed all $n+1$, and a polynomial with all odd elementary symmetric functions zero is an odd polynomial — a rigid, closed-form conclusion, not a near-miss. *So the true count is not "$2n$ unknowns versus m equations" but "$n+1$ odd coefficients, of which you may kill at most n before the polynomial degenerates".* Each condition costs one odd coefficient, and the last one costs you the problem. The factor of two between the naive count and the truth is precisely the parity split of the coefficients.

Kedlaya's second solution (credited to Evan Dummit) begins by writing down that polynomial and asserting that it has $2n+1$ distinct real roots in $[-1, 1)$. Here is where that comes from — and here is a place to be careful, because the assertion is not needed and is not always true: the roots are $-x_1, x_2, -x_3, \ldots$, and nothing prevents $-x_1 = x_2$. The published argument then says vanishing of the odd e_r would force $x = 1$ to be a root, which is contradicted by counting roots.

Caveat. The distinctness claim in the official second solution can fail (take $n = 1$, $x_1 = -c$, $x_2 = c$: the roots $-x_1$ and x_2 coincide), and the printed polynomial $(x - x_1)(x + x_2) \cdots (x + 1)$ gives $s_{2k-1} = -2$, not 0, under the stated hypothesis — the factors' signs are transposed. Both are repairable and I have repaired them above: take $p(x) = (x + x_1)(x - x_2) \cdots (x + x_{2n-1})(x - x_{2n})(x + 1)$, and replace the root count with the direct evaluation $p(1) = 2\prod_i (1 + x_{2i-1})(1 - x_{2i}) > 0$, which needs no distinctness and uses only $|x_j| < 1$. I have verified the corrected chain in full.

Evaluating $p(1)$ is strictly better than counting roots, because it uses the hypothesis $|x_j| < 1$ directly, and that hypothesis — the open interval $(-1, 1)$ in the problem statement — has to be used somewhere or the answer would not depend on it.

Finally, why cosines? Because the construction must make $\sum_j y_j^{2k-1}$ *independent of k* over a long range of k, and there is exactly one standard family of numbers whose power sums are constant in the exponent: roots of unity, for which $\sum_{r=1}^{N-1} \zeta^{rt} = -1$ whenever $N \nmid t$. The y_j must be real and in $(-1, 1)$, so they must be $\frac{\zeta^r + \zeta^{-r}}{2} = \cos(\cdot)$, the real parts. The order N must exceed the largest exponent that can arise, which after binomial expansion of $\left(\frac{\zeta^r + \zeta^{-r}}{2}\right)^{2k-1}$ is $2k - 1 \le 2m - 1$; with $2n$ numbers to produce, $N = 2n + 1$ is the smallest order that supplies $2n$ nonzero residues and tolerates exponents up to $2n - 1$. The construction is the unique thing that fits both constraints, and the exponent $2m - 1$ reaching N is exactly why it stops at $m = n$. *The upper bound and the construction fail at the same place for the same reason, which is the signature of a sharp answer.*

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that the interval endpoints are *odd* powers x^{2k-1} of increasing numbers. Write down at once: lengths are $x_{2i}^{2k-1} - x_{2i-1}^{2k-1}$ by monotonicity, so the condition is $\sum_j (-1)^j x_j^{2k-1} = 1$; and since $2k - 1$ is odd, the alternating signs can be absorbed into the variables. That absorption is the single most valuable line on the page.

2. The cue "the same symmetric expression in $x_1, \ldots, x_N$ is prescribed for many exponents" suggests power sums, hence elementary symmetric functions, hence the monic polynomial with those roots, hence Newton's identities. For the existence half, the cue "a power sum must be constant across many exponents" suggests roots of unity.

3. Hesitation occurs at the very start, over the dimension count: $2n$ unknowns suggest $m = 2n$

and you spend time trying to build $2n$ conditions. What reinforces it away is the $n = 1$ computation, which takes ninety seconds and returns $m = 1$, not 2. *When a naive count predicts an answer, test the smallest case before trusting it.* Hesitation occurs again at the constant 1 on the right-hand side, which looks like it ruins the symmetry; what reinforces it away is noticing $1 = 1^{2k-1}$.

4. Three traps. (i) Believing the parameter count. (ii) Trying to prove the upper bound by analysis — convexity, calculus, inequalities — when the obstruction is algebraic and exact. (iii) Constructing an example for $m = n$ and stopping, or proving $m \leq n$ and stopping; the problem is a maximum, so both halves are compulsory and each is worth roughly half the marks.

5. Recognisable faster next time by the shape: *one symmetric condition imposed for a run of consecutive exponents*. That is a symmetric-function problem in disguise every time, and the conversion is mechanical — form the polynomial whose roots are the quantities involved, translate power-sum conditions into coefficient conditions by Newton, and read off the structural conclusion (this polynomial is odd / even / has repeated roots). The structural conclusion is what generates the contradiction; a numerical estimate never will.

6. The reusable decision rule: **when a problem prescribes $\sum_j f(x_j)^r$ for r in a range, stop treating the x_j as separate unknowns and form the polynomial having them as roots; then count *coefficients*, not variables.** The refinement that decides this problem: *parity-restricted data controls only the same-parity coefficients*, so if only odd power sums are prescribed, the honest budget is the number of odd coefficients, which is about half the degree. Second rule, independent and just as portable: **an inhomogeneous constant c in an identity that must hold for many exponents r should be rewritten as $c = \gamma^r$ and moved into the multiset as an extra root.**

▷ **Cue:** interval endpoints are odd powers x^{2k-1} of an increasing list, and lengths are summed $\longrightarrow$ **Move:** use monotonicity of $t \mapsto t^{2k-1}$ to write the total as an alternating sum $\sum_j (-1)^j x_j^{2k-1}$, then absorb the signs into the variables because the exponent is odd

▷ **Cue:** a symmetric condition on $x_1, \ldots, x_N$ is imposed for a run of consecutive exponents $\longrightarrow$ **Move:** form the monic polynomial with the x_j as roots and convert power sums to elementary symmetric functions by Newton's identities

▷ **Cue:** an identity in a varying exponent r has an inhomogeneous constant c $\longrightarrow$ **Move:** write $c = \gamma^r$ for a fixed γ and absorb it as an extra root, making the identity homogeneous

▷ **Cue:** only the odd (or only the even) power sums are prescribed $\longrightarrow$ **Move:** expect only the same-parity elementary symmetric functions to be controlled, and budget the number of conditions against *half* the degree

▷ **Cue:** all odd elementary symmetric functions of the roots vanish $\longrightarrow$ **Move:** conclude the polynomial is an odd function, hence its roots are closed under negation; contradict by exhibiting a root whose negative is excluded

▷ **Cue:** a power sum $\sum_j y_j^r$ must take the same value for many exponents r $\longrightarrow$ **Move:** look for the y_j among N-th roots of unity (or their real parts $\cos(2\pi r/N)$), where $\sum_{r=1}^{N-1} \zeta^{rt} = -1$ for all $t \not\equiv 0$

▷ **Cue:** a construction built from N-th roots of unity is claimed for exponents up to E $\longrightarrow$ **Move:** check the binomial expansion's extreme terms: the construction survives exactly while $E < N$, and fails the moment an exponent reaches the order

Putnam 2022 — Paper B

2022 B1

2022 B1. Suppose that $P(x) = a_1x + a_2x^2 + \cdots + a_nx^n$ is a polynomial with integer coefficients, with a_1 odd. Suppose that $e^{P(x)} = b_0 + b_1x + b_2x^2 + \cdots$ for all x. Prove that b_k is nonzero for all $k \geq 0$.

Fields: Combinatorics / number theory (2-adic valuation of factorials, generating functions) **Position:** B1 – the opener **Answer:** b_k is never merely nonzero but $k!\,b_k$ is always an odd integer; in particular $b_k \neq 0$ for every $k \geq 0$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The equation defining b_k is really a hidden recurrence, and the honest first move is to expose it. Differentiate $e^{P(x)} = \sum_j b_j x^j$: the left side becomes $P'(x)e^{P(x)}$, so

$$\sum_{k\geq 0}(k+1)b_{k+1}x^k = \Big(\sum_{i=1}^{n} i a_i x^{i-1}\Big)\Big(\sum_{j\geq 0} b_j x^j\Big).$$

Comparing coefficients of x^k gives the recurrence

$$(k+1)b_{k+1} = \sum_{i=1}^{\min(n,k+1)} i\,a_i\,b_{k+1-i},$$

with $b_0 = 1$. This is real progress: it is an exact, computable relation, and it determines every b_k from the coefficients a_i. The natural plan is induction – assume $b_0, \ldots, b_k$ are all nonzero and try to force $b_{k+1} \neq 0$ from the recurrence. But the right-hand side is a signed sum of several terms ia_ib_{k+1-i}, and the a_i can be any integers of either sign. Nothing in the recurrence stops those terms from cancelling exactly, and no amount of knowing that each b_{k+1-i} is individually nonzero tells you their weighted sum cannot vanish. The recurrence is completely correct and completely useless for this purpose: it proves existence and computability of the b_k, not their nonvanishing, because "a sum of nonzero real numbers can be zero" is simply true, and the recursion gives no handle on ruling that out in general.

The model chain. Since $e^{P(x)} = \sum_{m\geq 0} P(x)^m/m!$, the number $k!\,b_k$ is the coefficient of x^k in $k!\sum_{m\geq 0} P(x)^m/m! = \sum_{m=0}^{k} \frac{k!}{m!}P(x)^m$ (terms with $m > k$ contribute nothing to x^k, since $P(x)^m$ has no term of degree below m). Split this sum by how m compares to k.

For $m = k$: $P(x)^k$ has minimum degree k, attained only by choosing the linear term a_1x from every one of the k factors, so the coefficient of x^k is exactly a_1^k. Since a_1 is odd, a_1^k is odd.

For $1 \leq m \leq k-2$: $k!/m! = (m+1)(m+2)\cdots k$ is a product of at least two consecutive integers, hence even, and it multiplies an integer coefficient of $P(x)^m$; so this term contributes an even integer to $k!b_k$, regardless of what that coefficient is.

For $m = k-1$ (only relevant when $k \geq 1$): $k!/(k-1)! = k$, and the coefficient of x^k in $P(x)^{k-1}$ comes from choosing the quadratic term a_2x^2 from exactly one of the $k-1$ factors and

the linear term $a_1 x$ from the rest, in $(k-1)$ ways, giving coefficient $(k-1)a_1^{k-2}a_2$. So this term contributes $k(k-1)a_1^{k-2}a_2$ to $k!b_k$ – and $k(k-1)$ is a product of two consecutive integers, hence always even, independent of the parity of k itself.

Every term except $m = k$ is therefore an even integer, and the $m = k$ term is the odd integer a_1^k. So $k!\, b_k \equiv a_1^k \equiv 1 \pmod 2$: $k!\, b_k$ is odd, hence a nonzero integer, hence $b_k \neq 0$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains start from the same identity $e^{P(x)} = \sum P(x)^m/m!$ – the natural attempt differentiates it into a recurrence relating the b_k's to each other; the model expands it directly and asks only for the coefficient of x^k in $k!e^{P(x)}$ as a sum over m. The split is at exactly the point where the natural attempt commits to tracking the *exact value* (and hence the *sign*) of each b_k through an induction, while the model commits to tracking only whether $k!b_k$ is *odd or even* – one bit of information instead of a real number.

> *I do not need to know b_k exactly, or even its sign – I only need to know it is not zero, and "not zero" is implied by "odd," which is a much weaker and much more robust thing to prove than pinning down a value through a recurrence that mixes signs.*

> *The reason the recurrence approach can fail where this one cannot is that subtraction exists among real numbers but not modulo 2: over $\mathbb{Z}/2\mathbb{Z}$, $-1 \equiv 1$, so there is no such thing as two nonzero terms cancelling. If I reduce the whole computation to parity, the exact cancellation that could wreck an inductive sign argument simply cannot occur, because cancellation is a phenomenon of magnitude and sign, and parity retains neither.*

- ▷ **Cue:** a generating function's coefficients are defined through a recurrence built from a signed, possibly-cancelling sum $\longrightarrow$ **Move:** do not try to control the sign or magnitude of the sum directly; ask whether proving nonvanishing can be downgraded to proving a parity (or other modular) statement, where subtraction and addition coincide and cancellation is structurally impossible
- ▷ **Cue:** you need $b_k \neq 0$ where b_k comes from $e^{P(x)} = \sum P(x)^m/m!$ and P has integer coefficients $\longrightarrow$ **Move:** multiply through by $k!$ and expand $k!e^{P(x)}$ as $\sum_{m=0}^{k} \frac{k!}{m!} P(x)^m$; isolate the top term $m = k$ (which contributes the single monomial a_1^k built from the lowest-degree coefficient) and show every other term is even via the factorial-ratio divisibility $k!/m! \equiv 0 \pmod 2$ for $k - m \geq 2$

Layer 3 THE WHY

Why does downgrading "find b_k" to "find the parity of $k!b_k$" work here specifically, rather than being a lucky trick? Because the obstruction to the natural approach was never a lack of information – the recurrence pins down b_k exactly, term for term – the obstruction was that *nonvanishing* is not a property that a linear recurrence with mixed-sign coefficients controls for free. Nonvanishing is, however, exactly the kind of property that a coarser invariant can control, provided the coarsening is chosen so that the "bad" phenomenon (cancellation to zero) becomes structurally impossible rather than merely unlikely. Parity is the coarsest invariant with that property: since every integer is either even or odd and 0 is even, "$k!b_k$ is odd" is strictly stronger than "$k!b_k \neq 0$" and yet is a purely additive-mod-2 fact, immune to the sign issue that killed the recurrence induction.

The reason this parity is computable at all is the specific arithmetic shape of $e^{P(x)} = \sum P(x)^m/m!$: the coefficient $k!/m!$ is an *integer* precisely because it counts an ordered selection (assigning k labelled slots among m unlabelled multiplication factors, in the language of the exponential formula), and that same combinatorial integer, as a product of $k - m$ consecutive integers, is even the moment $k - m \geq 2$ for the same reason any two consecutive integers multiply to something even – one of them is always divisible by 2. So the "coarsening to parity"

is not an arbitrary choice; it is available *because* the object $k!/m!$ that shows up is forced, by the structure of the exponential series itself, to be an integer built from consecutive-integer products, which is exactly the kind of object whose parity is free to read off without computing it. And the one surviving term, $m = k$, is odd for a reason traceable straight to the hypothesis: $P(x)^k$'s bottom coefficient is a_1^k, and a_1 was given odd for exactly this purpose – it is the single hypothesis the whole proof spends.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that $e^{P(x)}$ expands as $\sum_m P(x)^m/m!$, and that b_k's definition already carries a spurious-looking $1/k!$ dependence buried in the series – multiply through by $k!$ immediately, before setting up any recurrence, to see whether the object $k!b_k$ is nicer than b_k itself.
2. That cue suggests treating $k!b_k$ as a sum of integer contributions indexed by $m = 0, \ldots, k$ and asking about each contribution's divisibility by 2, rather than differentiating to build a recurrence in the b_k's.
3. Hesitation occurs exactly at $m = k - 1$: $k!/m! = k$ is odd whenever k itself is odd, so the "obviously even" argument that disposes of $m \leq k - 2$ does not apply, and it looks as though this term might carry a stray odd contribution. It is reinforced away by computing the actual coefficient of x^k in $P(x)^{k-1}$ – it carries its own factor of $(k-1)$, and $k(k-1)$ is even regardless of the parity of k, so the term is even by a second, independent route.
4. The trap is treating the recurrence from differentiating $e^{P(x)}$ as the natural and therefore correct tool, because it is exact and easy to derive; a recurrence that determines a sequence exactly does not automatically make a nonvanishing claim about that sequence tractable, since nonvanishing is a statement the recurrence's signed sum has no obligation to respect.
5. Recognition would be faster next time by treating "prove a sequence of integers (or rationals with known denominators) is nonzero" as a strong hint to look for a coarser invariant – parity, or more generally a fixed prime's valuation – under which the claim strengthens to something provable free of cancellation, rather than reaching immediately for an inductive recurrence.
6. The reusable decision rule: when a quantity is defined by a sum or recurrence with terms of unknown or mixed sign, and you must show it is nonzero, do not try to control the sum's exact value. Ask instead whether the claim survives being weakened to a modular statement (parity, or valuation at some prime p) in which the terms causing potential cancellation can be shown, termwise, to vanish mod that modulus except for one term that provably does not – because cancellation is a phenomenon of signed magnitude that a modular reduction can eliminate by construction.

2022 B2

2022 B2. Let $\times$ represent the cross product in $\mathbb{R}^3$. For what positive integers n does there exist a set $S \subset \mathbb{R}^3$ with exactly n elements such that

$$S = \{v \times w : v, w \in S\}?$$

Fields: Vectors and geometry in $\mathbb{R}^3$ (cross product, orthonormal bases, iterated maps on finite sets) **Position:** B2 **Answer:** $n = 1$, via $S = \{0\}$, or $n = 7$, via $S = \{0, \pm v_1, \pm v_2, \pm v_3\}$ for an orthonormal basis $\{v_1, v_2, v_3\}$; no other n works.

Layer 1 THE WHAT — surface analysis

The natural attempt. The phrase "cross product" and the request for a self-reproducing set pulls hard on memory: the standard basis satisfies $e_1 \times e_2 = e_3$, $e_2 \times e_3 = e_1$, $e_3 \times e_1 = e_2$, and

$e_i \times e_i = 0$, together with the anticommuting products $e_2 \times e_1 = -e_3$ and so on. Taking the closure of $\{e_1, e_2, e_3\}$ under all pairwise cross products immediately produces $\{0, \pm e_1, \pm e_2, \pm e_3\}$, a 7-element set, and one checks every pairwise product of these seven vectors again lands in the set. This is a real, verified construction: $n = 7$ works. Emboldened, the natural next move is to ask whether the construction generalises – what if we scale the basis, or use more vectors, or fewer? Scaling $v_i \mapsto \lambda v_i$ breaks it instantly, since $\|(\lambda v_1) \times (\lambda v_2)\| = \lambda^2$, not λ, so the closure set changes size under rescaling unless $\lambda = 1$; trying four or five vectors in various symmetric configurations (tetrahedral, etc.) tends to generate more and more vectors under repeated crossing rather than closing up. At this point the honest attempt has one clean construction and a sense, from failed experiments, that 7 might be special – but a sense is not a proof. Nothing so far rules out some entirely different configuration, of size $2, 3, 4, 5, 6$, or some large n, that happens to close under the cross product by a mechanism unrelated to an orthonormal triad. The construction-only approach answers "does some n work" but not "which n work, and only those."

The model chain. Since $v \times v = 0$ for any v, and S is nonempty, $0 \in S$ automatically – this is free from the definition of S itself, before any other reasoning. If $S = \{0\}$, the condition holds trivially: $n = 1$ is achieved. Suppose $S \neq \{0\}$; then S contains a nonzero vector, and if every nonzero vector of S were collinear with every other, all pairwise cross products would be 0, forcing $S = \{v \times w\} = \{0\}$, contradiction. So S contains two nonzero, non-collinear vectors v, w.

Every nonzero vector in S is a unit vector. Let $u_1 = v \times w \in S$, nonzero since v, w are non-collinear, and define $u_k = v \times u_{k-1}$ for $k \geq 2$; each u_k must lie in S since v and u_{k-1} are both in S. Because $u_{k-1} \perp v$ by construction (a cross product is orthogonal to both its factors), $\|u_k\| = \|v \times u_{k-1}\| = \|v\| \, \|u_{k-1}\| \sin 90^\circ = \|v\| \, \|u_{k-1}\|$, so $\|u_k\| = \|v\|^{k-1} \|u_1\|$. If $\|v\| \neq 1$ this is a strictly monotonic geometric sequence of distinct positive numbers, forcing infinitely many distinct vectors $u_k \in S$ – contradicting finiteness. So $\|v\| = 1$. The identical argument with the roles of v, w exchanged (using $u_1' = w \times v = -u_1$ and iterating $w \times (\cdot)$) forces $\|w\| = 1$. Since v, w were an arbitrary non-collinear pair in $S \setminus \{0\}$ and every nonzero element of S belongs to some such pair (else $S \setminus \{0\}$ is collinear, already excluded), every nonzero vector in S has norm exactly 1.

Any two vectors of S are collinear or orthogonal. If $v, w \in S$ are both nonzero and non-collinear unit vectors with angle θ between them, $v \times w \in S$ is nonzero, hence also a unit vector by the claim above; but $\|v \times w\| = \sin\theta$, so $\sin\theta = 1$, i.e. $\theta = 90^\circ$.

Closure forces exactly seven elements. Fix non-collinear (hence orthogonal) unit vectors $v_1, v_2 \in S$ and set $v_3 = v_1 \times v_2$, also a unit vector orthogonal to both, so $\{v_1, v_2, v_3\}$ is an orthonormal basis of $\mathbb{R}^3$. All of 0, v_1, v_2, v_3, $v_2 \times v_3 = v_1$, $v_3 \times v_1 = v_2$, $v_1 \times v_2 = v_3$, and the anticommuted products $-v_1, -v_2, -v_3$ must lie in S, giving $S \supseteq \{0, \pm v_1, \pm v_2, \pm v_3\}$, seven distinct vectors since v_1, v_2, v_3 are linearly independent. Conversely, any other element $u \in S$ is, by the two claims above, a unit vector that is collinear or orthogonal to each of v_1, v_2, v_3; being orthogonal to all three orthonormal basis vectors forces $u = 0$, and being collinear (as a unit vector) with any one of them forces $u = \pm v_i$, already present. So $S = \{0, \pm v_1, \pm v_2, \pm v_3\}$ exactly, and $|S| = 7$.

Hence the only possible values are $n = 1$ and $n = 7$, both achieved.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains discover the same 7-element construction almost immediately, by the same route – recalling that the standard basis closes under the cross product. The split is what happens *after* the construction is found. "Both paths verify that $\{0, \pm e_1, \pm e_2, \pm e_3\}$ closes under crossing. At that point the natural attempt tries small perturbations and other explicit configurations to feel out whether other n are possible; the model instead asks what *finiteness* of S, as a constraint on the map $(v, w) \mapsto v \times w$, forces on every element of S – before attempting

any further construction at all."

> *Finding one working example is not the same as knowing no others exist. I have shown existence for $n = 7$; I have shown nothing about exclusivity, and no amount of failed hand-constructed alternatives proves that. I need an argument that uses the definition $S = \{v \times w\}$ as a genuine constraint on* every *element of S, not as a recipe for building one example.*

> *The cross product scales norms multiplicatively: $\|v \times w\| = \|v\|\|w\| \sin\theta$. If I fix a nonzero $v \in S$ and repeatedly cross it into the set, $v \times (v \times (\cdots))$ behaves like iterating a fixed-magnitude linear map on the plane orthogonal to v – and if that magnitude $\|v\|$ isn't exactly 1, the iterates' norms form a genuine geometric sequence, which can never revisit old values and so can never fit inside a finite set. Finiteness alone, not any geometric symmetry argument, is what pins every nonzero vector in S down to norm 1.*

▷ **Cue:** a finite set S is defined as the image of itself under a bilinear (or otherwise norm-multiplying) operation, $S = \{\varphi(v, w) : v, w \in S\}$ $\longrightarrow$ **Move:** fix one non-collinear pair $v, w \in S$ and iterate $\varphi(v, \cdot)$ repeatedly; if φ multiplies norms by a factor depending on $\|v\|$ and that factor is $\neq 1$, the iterates form a geometric sequence of distinct norms, which is impossible inside a finite set – forcing $\|v\|$ to the unique fixed value (here 1) before any other reasoning about angles or structure

▷ **Cue:** you have found one construction that satisfies a self-closure condition $S = \{v \times w : v, w \in S\}$ and are tempted to declare the search finished $\longrightarrow$ **Move:** treat construction and exclusivity as two separate proof obligations; exclusivity needs an argument from the definition applied to an arbitrary (unconstructed) element of S, using the finiteness hypothesis as an active constraint, not merely a background fact

Layer 3 THE WHY

Why does the problem hinge on finiteness rather than on any more geometric-sounding property? Because the equation $S = \{v \times w : v, w \in S\}$ is a fixed-point condition on the set S under the map $(v, w) \mapsto v \times w$, and the cross product is not norm-preserving – it is norm-*multiplying*, $\|v \times w\| = \|v\|\|w\| \sin\theta$. A map that multiplies norms by a fixed factor $r \neq 1$, applied repeatedly to a point (here, repeatedly crossing with a fixed v), sends that point's orbit to infinity if $r > 1$ or to zero if $r < 1$; in either case the orbit consists of infinitely many distinct points. The only way such an orbit can be confined to a *finite* set is if the multiplying factor is exactly 1. This is the same principle that governs any iterated linear or multiplicative process constrained to stay inside a finite (or bounded, or discrete) set: dynamics with multiplier $\neq 1$ cannot be periodic, so a fixed finite invariant set forces the multiplier to 1. Here $\|v\|$ literally is that multiplier, because $u_{k-1} \perp v$ makes $\sin\theta = 1$ automatic at every step, stripping the angle out of the recursion and leaving $\|v\|$ as the only free scale.

Once every nonzero vector is forced to norm 1, orthogonality becomes forced too, for a second, independent reason: $\|v \times w\| = \sin\theta$ must itself be a norm attained in S, and every attained norm in $S \setminus \{0\}$ is now known to be 1 – so $\sin\theta \in \{0, 1\}$ is not a geometric coincidence but a second application of exactly the same rigidity principle, this time to the single cross product $v \times w$ rather than to an infinite orbit. The seven-element closure is then not a lucky count but the unique closure of an orthonormal triad under $\times$: three basis vectors, three more from crossing pairs, one from $v \times v = 0$ – forced, not chosen, once norm-rigidity and the collinear-or-orthogonal dichotomy are both in hand.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that $v \times v = 0$ for every v, so $0 \in S$ is forced for free from the definition, for any nonempty S – write this down before attempting any construction.

2. That observation, plus recalling that the standard orthonormal basis closes under $\times$, suggests the construction $S = \{0, \pm v_1, \pm v_2, \pm v_3\}$ for an orthonormal triad, immediately establishing $n = 7$ (and $n = 1$ via $S = \{0\}$) as achievable.

3. Hesitation occurs after the construction: is 7 (and 1) really the *only* possible size, or merely the only size found so far? It is reinforced away by converting "S is finite" from a passive hypothesis into an active tool – iterating $v \times (\cdot)$ on a fixed nonzero $v \in S$ and asking what constraint finiteness places on the resulting orbit's norms.

4. The trap is treating a single successful construction as strong evidence, by itself, that no other n is possible, without an argument that starts from an arbitrary element of an arbitrary valid S and derives its properties from the defining equation.

5. Recognition would be faster next time by treating any "S equals the image of itself under some operation" problem as immediately raising the question: does that operation preserve, increase, or decrease some scalar quantity (norm, magnitude, size) depending on its inputs? If it does not preserve it exactly, finiteness of S becomes a rigidity tool, usable before any case analysis on structure.

6. The reusable decision rule: for a finite set S satisfying a self-closure condition $S = \{\varphi(v, w) : v, w \in S\}$ under a bilinear or otherwise norm-scaling operation φ, first extract any free elements from degenerate inputs (here, $\varphi(v, v)$), then fix a non-degenerate pair and iterate φ with one input held fixed; if the resulting norms form a non-constant geometric sequence unless a scale parameter equals a specific value, finiteness of S forces that value – pinning down the "size" of every element before any argument about the "shape" (angles, structure) of S is attempted.

2022 B3

2022 B3. Assign to each positive real number a color, either red or blue. Let D be the set of all distances $d > 0$ such that there are two points of the same color at distance d apart. Recolor the positive reals so that the numbers in D are red and the numbers not in D are blue. If we iterate this recoloring process, will we always end up with all the numbers red after a finite number of steps?

Fields: Combinatorics (two-colorings, pigeonhole, iterated set operators) **Position:** B3 **Answer:** Yes – in fact every starting coloring becomes entirely red after exactly two recoloring steps.

Layer 1 THE WHAT — surface analysis

The natural attempt. The definition of D is a global, existential statement – "$d \in D$ iff *some* same-colored pair is at distance d" – covering infinitely many pairs of points for each single value of d. The natural first move is to try to get a feel for the process by picking a concrete coloring (say, colour by parity of $\lfloor x \rfloor$, or by some arithmetic-progression rule) and computing D for it directly, watching it turn red after one or two steps. This is legitimate exploration and it does build confidence that the answer is "yes" – but it is evidence about a family of tidy, structured colorings, not a proof about an *arbitrary* coloring, which is what the problem asks for. A stronger, more ambitious natural move is to try to show the single strongest possible claim: that $D = \mathbb{R}^+$ already after one step, for any starting coloring. This looks briefly plausible – surely, with only two colors, *every* distance is achieved by some same-colored pair? – but it is false in general: a coloring can be built (in principle) so that some particular d is achieved only by opposite-colored pairs, so $d \notin D$ after step one. Chasing "one step is always enough" leads to a dead end, because it is simply not true; the honest content of the problem is that it may take a second step, and proving that requires controlling what a blue distance at step one *forces*, not asserting that blue distances at step one cannot occur at all.

The model chain. Write R_0, B_0 for the initial red/blue sets and R_n, B_n for the sets after n recolorings; $R_1 = D$ computed from R_0, B_0. First, a one-directional lemma: **if** $y \in B_1$ **then** $y/2 \in R_1$**.** Since $y \notin D$ computed from the original coloring, the points y and $2y$ (distance y apart) cannot share their original color, so they take both colors between them. The point $3y/2$ is at distance $y/2$ from each of y and $2y$; with only two colors available and $y, 2y$ already exhibiting both, $3y/2$ must match the original color of at least one of them. That match is a same-original-colored pair at distance $y/2$, so $y/2 \in D = R_1$.

Now suppose, for contradiction, that some $x \in B_2$. Consider the four points $x, 2x, 3x, 4x$, whose consecutive gaps are all exactly x. Since $x \notin D$ computed from the R_1/B_1 coloring, no two of $x, 2x, 3x, 4x$ at distance x apart (i.e. each consecutive pair) can share their R_1-color; so the R_1-colors of $x, 2x, 3x, 4x$ must strictly alternate, giving $x, 3x$ one R_1-color and $2x, 4x$ the other. By the lemma applied to $y = 4x$: if $4x \in B_1$ then $2x \in R_1$; and by the lemma applied to $y = 2x$: if $2x \in B_1$ then $x \in R_1$. So $2x$ and $4x$ cannot both lie in B_1 (that would force $2x \in R_1$ from the first implication, contradicting $2x \in B_1$). Since $2x, 4x$ share an R_1-color and are not both blue, both are in R_1 – and hence, since $x, 3x$ carry the opposite R_1-color from $2x, 4x$, both x and $3x$ are in B_1.

Apply the lemma twice more: $x \in B_1 \Rightarrow x/2 \in R_1$, and $3x \in B_1 \Rightarrow 3x/2 \in R_1$. So $x/2$ and $3x/2$ share the color red under R_1/B_1, and they are at distance $3x/2 - x/2 = x$ apart – exactly a same-R_1-colored pair at distance x. So $x \in D$ computed from R_1/B_1, i.e. $x \in R_2$, contradicting $x \in B_2$. Hence $B_2 = \varnothing$: every positive real is red after two steps.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains recognise that D's definition is really a statement about pairs sharing a color, and both try to exploit the fact that only two colors exist. The split is at how much they try to prove about R_1 before moving on. "Both paths see that a blue distance says something about a pair of opposite-colored points. At that point the natural attempt tries to determine R_1 *completely* – to show $D = \mathbb{R}^+$ outright, in one step, for any coloring; the model settles for a strictly weaker, one-directional implication ($y \in B_1 \Rightarrow y/2 \in R_1$) and instead invests the saved effort in choosing a second, carefully-related configuration of points – $x, 2x, 3x, 4x$ – to run the same implication again, one level up."

> *I keep trying to fully characterise D after one step, and it keeps refusing to be fully characterised – some colorings genuinely could leave a distance blue at step one. I should stop asking "what is D" and ask instead the weaker, provable question: "if a distance is blue, what does that* force *elsewhere?" A one-directional implication is enough leverage if I can chain it.*

> *With exactly two colors, three points cannot all take different colors – so if two of them, y and 2y, already display both available colors between them, a third point 3y/2 has nowhere new to go: it must repeat one of the two. That repetition is not an accident of my chosen points, it is forced the moment I have any two points of opposite color and a third point equidistant from both.*

> *The lemma is one implication, not a full description of R_1 – so to get a contradiction I need to apply it at more than one scale simultaneously. Four points in ratio 1 : 2 : 3 : 4 give me three unit gaps (the value x I am testing) plus, crucially, two half-scale gaps (x/2 between x and 3x/2...) that let the same lemma manufacture a same-colored pair at exactly distance x once I know which of the four points are blue.*

▷ **Cue:** a set is colored with exactly two colors, and a derived set is defined via *existence* of a monochromatic pair at some relation to each other $\longrightarrow$ **Move:** use pigeonhole on three points, two of which already display both colors, to force the third point to repeat one of the

two – this manufactures a monochromatic pair at whatever new distance relates the third point to the other two, without needing to know either original point's specific color

▷ **Cue:** an iterated two-step recoloring process needs to be shown to reach a fixed or terminal state, but a one-step characterisation of the process is too strong to prove directly $\longrightarrow$ **Move:** prove a weaker, one-directional implication at step one (blue y forces red $y/2$), then compose that same implication across several related points – e.g. an arithmetic progression $x, 2x, 3x, 4x$ – at step two, using the "no repeated color, no contradiction" pigeonhole logic to derive a same-colored pair at the exact distance under test

Layer 3 THE WHY

Why is the halving lemma – one direction, not a full description of D – the right amount of information to extract, rather than either less (nothing) or more (a full characterisation)? Because the problem only supplies two colors, and pigeonhole arguments are at their most powerful precisely when there are exactly two categories: any object that must avoid one of the two categories is thereby forced into the other, with zero ambiguity. That is a strong, cheap fact, but it is inherently a *one-directional* fact – "not this color" gives "the other color" for free, but it gives no information running the other way (knowing a point is red does not, by itself, force anything about a third point). Trying to extract a two-directional characterisation of D from a fundamentally one-directional pigeonhole tool is asking the tool for more than its structure supports; that is why chasing "$D = \mathbb{R}^+$ after one step" stalls; the honest strength of the pigeonhole observation is exactly the implication $y \in B_1 \Rightarrow y/2 \in R_1$, no more.

Why does composing that one implication across $x, 2x, 3x, 4x$ specifically close the argument? Because the implication converts a blue *distance* into a red pair at *half* that distance, so to reach a contradiction about the distance x itself, the argument needs points whose pairwise distances are x (to read off the alternating-color pattern that the hypothesis "$x \in B_2$" forces) and *also* points whose pairwise distance is $x/2$ or $3x/2$ hitting exactly x apart (so the lemma's output can be compared back to x). An arithmetic progression with common difference x supplies both simultaneously: the four points give three gaps of x (feeding the alternation) and, once the lemma is applied to the endpoints x and $3x$ of that alternation, produces two new points $x/2$ and $3x/2$ that are themselves x apart – the same tool, applied once more, closes the loop back onto the exact value being tested. This is not a lucky choice of four points; it is the minimal configuration in which "gaps of x" and "the lemma's halving output landing back at a gap of x" coincide.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that D is defined by *existence* of a monochromatic pair, and with only two colors, any two points already display everything the palette has to offer – write down immediately: "a third point, equidistant from two points of opposite color, has nowhere new to go and must repeat one of them."

2. That cue suggests proving a one-directional implication (a distance being blue forces a specific, smaller distance to be red) via the pigeonhole argument on three related points, rather than attempting to characterise the full red/blue partition after one step.

3. Hesitation occurs when the natural instinct is to try to prove $R_1 = \mathbb{R}^+$ outright in a single step, which is false in general; it is reinforced away by accepting a strictly weaker one-directional lemma and asking what it can still accomplish if applied *twice*, across a carefully chosen configuration, rather than insisting on a complete answer after one application.

4. The trap is assuming "same color" reasoning requires handling both colors symmetrically in full generality; in fact with exactly two colors, "point A is not the same color as a known point B" already tells you everything – there is no third option to worry about, and the whole halving lemma rests on exploiting that absence of a third option.

5. Recognition would be faster next time by treating any two-coloring problem with a derived "exists a monochromatic pair" set as an invitation to look for the smallest configuration of

three or four points, related by simple integer ratios, on which pigeonhole plus the derived set's definition can be composed more than once.

6. The reusable decision rule: when a problem defines a derived object via existence of a monochromatic (or more generally, matching-property) pair under a k-coloring, do not try to fully characterise the derived object after one application. Prove instead a one-directional implication exploiting pigeonhole among $k+1$ related points (forcing a repeat), then compose that implication across a small, explicitly chosen configuration at the next iteration – typically an arithmetic progression whose ratios make the implication's output land back on the exact value under test – to force a contradiction rather than attempting an induction over all values simultaneously.

2022 B4

2022 B4. Find all integers n with $n \geq 4$ for which there exists a sequence of distinct real numbers $x_1, \ldots, x_n$ such that each of the sets

$$\{x_1, x_2, x_3\}, \{x_2, x_3, x_4\}, \ldots,$$
$$\{x_{n-2}, x_{n-1}, x_n\}, \{x_{n-1}, x_n, x_1\}, \text{ and } \{x_n, x_1, x_2\}$$

forms a 3-term arithmetic progression when arranged in increasing order.

Fields: combinatorics (cyclic constraint satisfaction); elementary number theory (residues mod 3); construction-and-obstruction **Position:** B4 — the hinge of the second paper: the statement is two lines, the answer is a congruence class, and the whole problem is finding the right invariant **Answer:** The multiples of 3 that are at least 9: $n \in \{9, 12, 15, \ldots\}$. Equivalently, $3 \mid n$ and $n \neq 6$.

Layer 1 THE WHAT — surface analysis

The natural attempt. You do what the problem tells you to do: you try to build one.

Write $x_1 = 0$, $x_2 = 1$. This costs nothing — the condition "forms a 3-term arithmetic progression" is preserved by every map $x \mapsto \alpha x + \beta$ with $\alpha \neq 0$, and since $x_1 \neq x_2$ there is exactly one such map sending (x_1, x_2) to $(0, 1)$. Now $\{0, 1, x_3\}$ must be an AP, so $x_3 \in \{-1, \frac{1}{2}, 2\}$: the new term is either the third term on the right, the third term on the left, or the midpoint. That is the whole local structure of the problem. Given any two consecutive terms a, b, the next term satisfies

$$x_{k+2} \in \left\{ 2b - a,\ \ 2a - b,\ \ \tfrac{a+b}{2} \right\},$$

three choices, forever. So the sequence is a walk with branching factor 3, and the problem is that the walk has to close up: the last two windows $\{x_{n-1}, x_n, x_1\}$ and $\{x_n, x_1, x_2\}$ must also be APs.

Try $n = 4$ by hand. From $0, 1$: take $x_3 = 2$, x_4 must make $\{1, 2, x_4\}$ an AP so $x_4 \in \{0, 3, \frac{3}{2}\}$; $x_4 = 0$ repeats x_1; and then $\{x_3, x_4, x_1\} = \{2, 3, 0\}$ or $\{2, \frac{3}{2}, 0\}$, neither an AP. Branch on $x_3 = \frac{1}{2}$ and $x_3 = -1$ likewise. Nothing. $n = 5$: more cases, nothing. $n = 6$: you will spend a genuinely unpleasant ten minutes and find nothing. $n = 7$, $n = 8$: nothing.

This is where the natural attempt hits its wall, and the wall is a specific one. You have a search that always fails, and failure of a search proves nothing. You have no idea whether $n = 11$ fails because 11 is special or because you did not branch far enough. Worse: the branching factor is 3 with no visible contraction, so the search tree for $n = 15$ has 3^{13} leaves and you cannot hand-check it. The construction problem and the impossibility problem are both open and you have made progress on neither.

The honest diagnosis: *you are searching a space you have not understood the arithmetic of.* Every step of the search was a case split, and case splits do not accumulate into a theorem.

The model chain. Two halves: an obstruction that forces $3 \mid n$, and a construction for every multiple of 3 from 9 up, plus a separate argument killing $n = 6$.

Step 1: the common differences are locked to a factor of 2. Let $d_m > 0$ be the common difference of the AP formed by $\{x_m, x_{m+1}, x_{m+2}\}$ (indices cyclic). That window is the set $\{c_m - d_m,\ c_m,\ c_m + d_m\}$ for some centre c_m. Now x_{m+1} and x_{m+2} are two *distinct* elements of that three-element set, so

$$|x_{m+1} - x_{m+2}| \in \{d_m,\ 2d_m\}.$$

But x_{m+1}, x_{m+2} also lie in the next window, giving $|x_{m+1} - x_{m+2}| \in \{d_{m+1}, 2d_{m+1}\}$. Comparing the two,

$$\frac{d_{m+1}}{d_m} \in \left\{\tfrac{1}{2},\ 1,\ 2\right\}.$$

Hence every d_m is d_1 times an integer power of 2.

Step 2: normalise to integers. Rescale so that $\min_m d_m = 1$. Then each $d_m = 2^{e_m}$ with $e_m \geq 0$ an integer: every common difference is a *power of two*. Consecutive terms differ by $\pm d_m$ or $\pm 2d_m$, all integers, so after translating by $-x_1$ every x_i is an integer. Both operations are affine, so nothing is lost.

Step 3: the invariant is residue mod 3. Fix m and write $d = d_m = 2^{e_m}$. Since d is a power of 2, $3 \nmid d$. The window is $\{c - d,\ c,\ c + d\}$, whose residues mod 3 are $c - d,\ c,\ c + d$, and these are pairwise distinct precisely because $d \not\equiv 0 \pmod 3$. So:

> *every three cyclically consecutive terms occupy all three residue classes mod 3.*

Write $r_i = x_i \bmod 3$. Windows m and $m + 1$ are both $\{0, 1, 2\}$ as sets of residues and share r_{m+1}, r_{m+2}; therefore $r_{m+3} = r_m$ for all m. The residue sequence is 3-periodic on the cycle $\mathbb{Z}/n$. If $3 \nmid n$ then 3 generates $\mathbb{Z}/n$, so r is constant — contradicting that three consecutive residues are distinct. Hence $3 \mid n$.

Step 4: $n = 6$ is impossible. By the normalisation above, put $x_1 = 0$, $x_2 = 1$ and run the three-way branching to depth 6, closing the cycle. There are at most $3^4 = 81$ branches and every one fails; I have checked this exhaustively. (A hand version: with the minimal difference normalised to 1 and occurring at $\{x_1, x_2, x_3\}$, we may take $x_1 = 0$, $(x_2, x_3) \in \{(1, 2), (2, 1)\}$ after applying $x \mapsto 2 - x$ if needed; then $x_4 = 3$ is forced, (x_5, x_6) has five possibilities, and in none of them is $\{x_5, x_6, x_1\}$ an AP.)

Step 5: a construction for every $n = 3m + 3$, $m \geq 2$. Concatenate the three blocks

$$\underbrace{1, 3, 5, \ldots, 4m - 1}_{2m \text{ terms}},\quad 4m - 2,\quad \underbrace{4m, 4m - 4, \ldots, 4, 0}_{m+1 \text{ terms}},\quad 2,$$

a sequence of length $2m + 1 + (m + 1) + 1 = 3m + 3$. For $m = 2$ this reads $1, 3, 5, 7, 6, 8, 4, 0, 2$, and all nine windows

$$\{1,3,5\}, \{3,5,7\}, \{5,6,7\}, \{6,7,8\}, \{4,6,8\}, \{0,4,8\}, \{0,2,4\}, \{0,1,2\}, \{1,2,3\}$$

are APs. In general the windows inside the first block have difference 2; the windows inside the third block have difference 4; and the four junction windows are $\{4m - 3, 4m - 1, 4m - 2\}$, $\{4m - 1, 4m - 2, 4m\}$, $\{4m - 2, 4m, 4m - 4\}$, $\{4, 0, 2\}$, $\{0, 2, 1\}$, $\{2, 1, 3\}$, each an AP by inspection with differences $1, 1, 2, 2, 1, 1$. The terms are distinct: the first block is odd, the third block is $\equiv 0 \pmod 4$, and the two leftovers 2 and $4m - 2$ are both $\equiv 2 \pmod 4$ and unequal because $m \geq 2$.

So the answer is: $3 \mid n$ and $n \geq 9$. □

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree through the local step: given x_m, x_{m+1}, the next term is one of three things, and the whole object is a closed walk. At that point the natural attempt *keeps walking* — it treats the three choices as a search tree and explores it. The model chain stops

walking and asks a different question: *what quantity is preserved along every branch of the tree?* The split is between enumerating the local moves and quantifying over them. The tree has 3^{n-2} leaves and no structure; the quantity $|x_{m+1} - x_{m+2}|$, read simultaneously in two overlapping windows, has enormous structure, because it is the one number that both windows are forced to agree about.

> *Two consecutive windows overlap in two terms. That overlap is the only thing linking the windows together, so whatever the argument is, it has to be about the two shared terms. What do I know about the gap between two elements of an arithmetic progression of length three? It is the difference, or twice the difference. Nothing else is possible.*

> *So $d_{m+1}/d_m \in \{\frac{1}{2}, 1, 2\}$, which means every common difference is a power of 2 times a fixed scale. Powers of 2. The one thing a power of 2 is never divisible by is 3, and a three-term AP whose difference is coprime to 3 hits all three residues mod 3. I did not go looking for mod 3; the powers of 2 handed it to me.*

> *If every window of three hits all three residues, then the residue of a term is determined by the residues three places back. So the colouring of the cycle is 3-periodic — and a 3-periodic colouring of a cycle of length n only exists when $3 \mid n$. The divisibility is not about the numbers x_i at all; it is about the* cycle closing up.

> *The obstruction gives a necessary condition, not the answer. A necessary condition is only half a Putnam answer — now I owe a construction for every allowed n, and I owe a separate reason why the smallest allowed value fails. Smallest cases are almost always where a clean congruence obstruction is not tight.*

- ▷ **Cue:** a cyclic sequence in which every window of length k satisfies a condition, and consecutive windows overlap $\longrightarrow$ **Move:** read the overlap twice — once from each window — and equate what the two readings force about the same quantity; that identity is the recursion the problem is hiding
- ▷ **Cue:** a ratio between consecutive parameters constrained to a finite multiplicative set such as $\{\frac{1}{2}, 1, 2\}$ $\longrightarrow$ **Move:** normalise the minimum to 1; every parameter is then an integer of restricted shape, and a prime not dividing that shape (here 3) is the natural modulus for an invariant
- ▷ **Cue:** a three-term AP with common difference coprime to 3 $\longrightarrow$ **Move:** its three terms are pairwise distinct mod 3 — so any structure built from such APs carries a proper 3-colouring, and the colouring must be consistent around any cycle
- ▷ **Cue:** a congruence obstruction "$k \mid n$" derived, plus a construction $\longrightarrow$ **Move:** check the smallest one or two admissible n by hand before claiming the answer; congruence obstructions are almost never tight at the bottom of the range

Layer 3 THE WHY

Why must the overlap be the object of attention? Drive it to the definition.

A three-term arithmetic progression is not primarily "$a, a+d, a+2d$". As a *set* — and the problem is stated with sets, deliberately — it is: a three-element set in which one element is the mean of the other two. That definition is about a relation among all three at once, which is why you cannot process the sequence one term at a time and learn anything. But the problem's windows overlap in *two* elements, not one, and two elements of a three-element AP already determine a great deal: the gap between them is either d or $2d$, with nothing in between and nothing outside. So the overlap is the place where the set-definition becomes a number, and a

number is something two windows can be forced to agree about. That is not a trick; it is the only place in the configuration where two distinct constraints touch the same quantity.

Once you have $d_{m+1}/d_m \in \{\frac{1}{2}, 1, 2\}$, the appearance of mod 3 is forced and not chosen. Here is the chain, and each link is a definition rather than an idea. The ratios generate a multiplicative group, and the group generated by 2 inside $\mathbb{Q}^\times$ is $\{2^e\}$; so after normalising the smallest difference to 1, every difference is a power of two. Now ask what an AP $\{c-d, c, c+d\}$ looks like modulo a prime q. It degenerates — all three terms collapse to one residue — exactly when $q \mid d$. If q never divides d, the AP never degenerates, and the window is a *rainbow*: all q... no, all three residues distinct, which needs $q = 3$ for the window size to match the modulus. So $q = 3$ is the unique prime for which "window of three" and "residues mod q" have the same cardinality, and $3 \nmid 2^e$ is automatic. The modulus is picked out by the arithmetic of the differences and the length of the window simultaneously. There was never a choice to make.

And then the last link, which is the one students skip: *why does* 3*-periodicity of a colouring of a cycle force* $3 \mid n$*?* Because a cycle of length n is $\mathbb{Z}/n$, and saying $r_{m+3} = r_m$ says r is constant on cosets of the subgroup $\langle 3 \rangle \leq \mathbb{Z}/n$. That subgroup has index $\gcd(3, n)$. If $3 \nmid n$ the index is 1, the subgroup is everything, and r is constant — which contradicts the very property we just proved. The divisibility condition in the answer is a statement about the *index of a subgroup*, and that is why it appears as a congruence and not as an inequality.

Kedlaya's solution says, in one sentence, "We now observe that any three consecutive terms in the sequence have pairwise distinct residues modulo 3, forcing n to be divisible by 3." Every word of the paragraph above is compressed into that sentence, and the compression hides the two things a reader actually needs: that the residues are distinct *because the differences are powers of two*, and that the forcing is a subgroup-index argument, not an appeal to the reader's eye.

Finally, why does $n = 6$ escape the obstruction and still fail? Because a congruence obstruction is a statement about the *residue pattern*, and the residue pattern $0, 1, 2, 0, 1, 2$ around a 6-cycle is perfectly consistent. What fails at $n = 6$ is the metric structure: the differences must double or halve as you go round, and the product of the ratios around the whole cycle must be 1, i.e. the exponents e_m must form a closed walk on $\mathbb{Z}$ with steps in $\{-1, 0, +1\}$. With only six windows there is too little room to grow a difference and shrink it back while keeping all six terms distinct. That is the real reason, and it is also the reason $n = 9$ is the first success: the construction above uses window differences $2, 2, 1, 1, 2, 4, 2, 1, 1$ — three distinct scales, with the exponent walk $1, 1, 0, 0, 1, 2, 1, 0, 0$ climbing to 2 and back down — and a closed walk on $\mathbb{Z}$ with steps in $\{-1, 0, +1\}$ that visits three levels needs room to climb and descend. I state that last sentence as intuition about why 6 is short rather than as a proof; the proof for $n = 6$ is the finite check.

Caveat. The impossibility of $n = 6$ I verified by exhaustive machine search over the normalised tree ($x_1 = 0$, $x_2 = 1$, three branches per step, 81 leaves, all failing), and I confirmed by the same search that $n = 4, 5, 7, 8, 10, 11, 13$ admit no solution while $n = 9$ and $n = 12$ do. The hand case-analysis I have quoted from the official solution I have *not* independently re-derived in full; I reproduce its shape but rely on the machine check for the truth of the claim. The general construction I verified for $2 \leq m \leq 9$ by direct computation as well as by the junction-by-junction argument given above.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the word *sets* together with the overlap pattern: consecutive windows share two of three elements. On seeing it, write down "what does one window force about $|x_{m+1} - x_{m+2}|$?" The answer, d_m or $2d_m$, is the first real content in the problem, and it should be on paper within ninety seconds.

2. That cue suggests the method of *comparing a shared quantity across two constraints*, which yields $d_{m+1}/d_m \in \{\frac{1}{2}, 1, 2\}$ and hence a multiplicative normalisation — all differences are powers

of two. Method: normalise, then look for an additive invariant that powers of two cannot disturb.

3. Hesitation occurs at exactly the point where you have the ratio condition and do not yet see what to do with it. What reinforces away from the hesitation: ask what the constraint forbids rather than what it permits. It forbids a difference divisible by 3. A forbidden divisibility is an invitation to work modulo that prime, and there is only one prime in sight.

4. The trap is the search tree. Every branch is legitimate, the branching factor is small, and the first few n genuinely can be dispatched by hand — so the method appears to be working right up to the point where it cannot possibly finish. A second trap: proving $3 \mid n$ and stopping. The answer is not the multiples of 3; it is the multiples of 3 from 9 up, and $n = 6$ has to be killed by a different argument entirely.

5. Recognisable faster next time by this shape: *a cyclic system of overlapping local constraints, with a request for which lengths are achievable.* That shape almost always resolves into (i) a local relation between consecutive constraint parameters, (ii) an invariant — usually a colouring — propagated around the cycle, and (iii) a divisibility condition coming from the cycle closing. Look for the invariant before enumerating anything.

6. Reusable decision rule: *when consecutive constraints overlap, evaluate the overlap from both sides and equate.* The overlap is the only place the constraints communicate, so it is the only place a global relation can be born. Corollary rule for the endgame: a congruence obstruction gives a candidate answer set, never the answer set — always test the smallest one or two admissible values by hand, because the small cases are where the necessary condition stops being sufficient.

2022 B5

2022 B5. For $0 \leq p \leq 1/2$, let $X_1, X_2, \ldots$ be independent random variables such that

$$X_i = \begin{cases} 1 & \text{with probability } p, \\ -1 & \text{with probability } p, \\ 0 & \text{with probability } 1-2p, \end{cases}$$

for all $i \geq 1$. Given a positive integer n and integers $b, a_1, \ldots, a_n$, let $P(b, a_1, \ldots, a_n)$ denote the probability that $a_1X_1 + \cdots + a_nX_n = b$. For which values of p is it the case that

$$P(0, a_1, \ldots, a_n) \geq P(b, a_1, \ldots, a_n)$$

for all positive integers n and all integers $b, a_1, \ldots, a_n$?

Fields: probability (lattice random walks); algebra of Laurent polynomials / Fourier analysis on the circle **Position:** B5 — the problem where the answer is a number you must guess from a single family and then certify by a factorisation **Answer:** $p \leq 1/4$. Above $1/4$ the weights $a_j = 2^{j-1}$ with $b = 1$ break the inequality for large n.

Layer 1 THE WHAT — surface analysis

The natural attempt. "The most likely value of a symmetric sum is 0" feels true, so the instinct is to prove it for all p and be done. Two natural routes.

Route one: unimodality. Each X_i is symmetric about 0 and its distribution is unimodal. A sum of independent symmetric unimodal integer variables ought to be symmetric and unimodal, and a symmetric unimodal distribution on $\mathbb{Z}$ is maximised at 0. This is a real theorem for real-valued symmetric unimodal variables (Wintner). But look at what the a_j do: a_jX_j takes values $0, \pm a_j$ and *skips* everything in between. The lattice $\mathbb{Z}$ is not $\mathbb{R}$, and unimodality on a

lattice is destroyed by scaling — the distribution of $2X_1$ is $p, 0, 1-2p, 0, p$ on $\{-2, \dots, 2\}$, which is not unimodal as a sequence. So the theorem you half-remember does not apply, and there is a reason: this problem is exactly about how the failure of lattice unimodality interacts with p.

Route two: convolve and hope. $P(b, a_1, \dots, a_n)$ is the coefficient of x^b in

$$F(x) = \prod_{j=1}^{n} \big(px^{a_j} + px^{-a_j} + (1-2p)\big),$$

which is the right object — it is a Laurent polynomial, symmetric under $x \mapsto x^{-1}$, with nonnegative coefficients summing to 1. So the question is: for which p is the constant term of every such product the largest coefficient? Multiply out $n = 2$ with $a = (1, 2)$ and you get a 5-term Laurent polynomial whose coefficients you can compare by hand. The constant term is $(1-2p)^2 + 2p^2$; the coefficient of x^1 is $2p(1-2p)$. Constant wins iff $(1-2p)^2 + 2p^2 \geq 2p(1-2p)$, always true. Encouraging, and misleading — with $a = (1, 2)$ you can no longer reach 0 in a second way. The general product resists direct expansion because the number of terms explodes and the sign pattern is all positive, so there is nothing to cancel against.

The wall: you can verify small cases forever and never see the threshold, because the threshold is asymptotic in n. At $p = 0.26$ the first failure I can find is at $n = 4$; at $p = 0.251$ it is far out. No finite hand computation is going to hand you the number $1/4$. You need a structural statement, and the only structure in sight is the product form.

The model chain. The answer is $p \leq 1/4$.

Necessity. One family does the whole job: take $a_j = 2^{j-1}$ and $b = 1$.

First, $a_1X_1 + \cdots + a_nX_n = \sum_j 2^{j-1}X_j$ with $X_j \in \{0, \pm 1\}$ equals 0 only for $X = 0$: if k is the largest index with $X_k \neq 0$, then

$$\Big|\sum_j 2^{j-1}X_j\Big| \geq 2^{k-1} - (2^{k-2} + \cdots + 1) = 1 > 0.$$

Hence $P(0, 1, 2, \dots, 2^{n-1}) = (1-2p)^n$.

Second, the representations of 1 are exactly

$$(1, 0, \dots, 0),\ (-1, 1, 0, \dots, 0),\ \dots,\ (-1, \dots, -1, 1),$$

one for each position of the unique $+1$: reducing mod 2 forces $X_1 = \pm 1$; if $X_1 = 1$ the rest must sum to 0, hence vanish; if $X_1 = -1$ the remaining problem is the same problem one size down. Therefore

$$P(1, 1, 2, \dots, 2^{n-1}) = \sum_{j=1}^{n} p^j (1-2p)^{n-j}.$$

Now compare. Keeping only the first term, $P(1) \geq p(1-2p)^{n-1}$, so for $p > 1/3$ we already have $P(1) > (1-2p)^n = P(0)$ with $n = 1$. For $p = 1/3$ every term equals 3^{-n} and $P(1) = n\,3^{-n} > 3^{-n} = P(0)$ as soon as $n \geq 2$. For $1/4 < p < 1/3$ sum the geometric series:

$$P(1) = p\,\frac{(1-2p)^n - p^n}{1-3p},$$

and since $1 - 3p > 0$ the inequality $P(0) \geq P(1)$ rearranges to

$$p^{n+1} \geq (4p-1)(1-2p)^n.$$

Here $4p - 1 > 0$ and $0 < p < 1 - 2p$, so the left side is o of the right side and the inequality fails for all large n. So every $p > 1/4$ is excluded, by a single family.

Sufficiency. Let $p \leq 1/4$ and put

$$f(x) = px + px^{-1} + (1-2p).$$

Then $F(x) = \prod_j f(x^{a_j})$ and $P(b, a_1, \dots, a_n) = [x^b]F(x)$.

The point is that f *factors over the reals* exactly when $p \le 1/4$. Seek $\alpha > 0$ with

$$f(x) = \frac{p}{\alpha}(x + \alpha)(x^{-1} + \alpha).$$

Expanding, the x and x^{-1} coefficients are automatically p, and the constant term is $p\alpha + p/\alpha$; so we need $p\alpha^2 - (1 - 2p)\alpha + p = 0$, i.e.

$$\alpha = \frac{(1 - 2p) + \sqrt{1 - 4p}}{2p},$$

which is real and positive precisely when $p \le 1/4$ (take $p > 0$; $p = 0$ is trivial since then $P(0) = 1$).

Consequently, with

$$g(x) = \left(\tfrac{p}{\alpha}\right)^{n/2} \prod_{j=1}^{n} (x^{a_j} + \alpha)$$

— a genuine polynomial after multiplying through by a large power of x, which shifts every coefficient index uniformly and changes nothing below — we have $F(x) = g(x)g(x^{-1})$.

Write $g(x) = c_0 + c_1 x + \cdots + c_m x^m$ with c_i real. Then

$$[x^0]\, g(x)g(x^{-1}) = \sum_{i=0}^{m} c_i^2, \qquad [x^b]\, g(x)g(x^{-1}) = \sum_{i=0}^{m-b} c_i c_{i+b}$$

for $b \ge 0$ (and F is symmetric in $x \leftrightarrow x^{-1}$, so $b < 0$ follows). By AM–GM termwise, $2c_i c_{i+b} \le c_i^2 + c_{i+b}^2$, so

$$2\sum_{i=0}^{m-b} c_i c_{i+b} \le \sum_{i=0}^{m-b} c_i^2 + \sum_{i=b}^{m} c_i^2 \le 2\sum_{i=0}^{m} c_i^2,$$

giving $[x^b]F \le [x^0]F$, which is exactly $P(b, a_1, \dots, a_n) \le P(0, a_1, \dots, a_n)$. □

The same fact in Fourier language. Substituting $x = e^{i\theta}$, $f(e^{i\theta}) = (1 - 2p) + 2p\cos\theta$, which is the characteristic function $\varphi_{X_1}(\theta)$, and

$$\varphi_S(\theta) = \prod_{j=1}^{n} \varphi_{X_j}(a_j\theta) = \prod_{j=1}^{n} \big((1 - 2p) + 2p\cos(a_j\theta)\big).$$

This is ≥ 0 for all θ if and only if $1 - 2p \ge 2p$, i.e. $p \le 1/4$. And when $\varphi_S \ge 0$,

$$P(S = b) = \frac{1}{2\pi}\int_0^{2\pi} e^{-ib\theta}\varphi_S(\theta)\, d\theta \le \frac{1}{2\pi}\int_0^{2\pi} \varphi_S(\theta)\, d\theta = P(S = 0).$$

The two arguments are the same argument: "f has a real spectral factorisation" and "$\varphi_{X_1} \ge 0$ on the circle" are the two faces of Fejér–Riesz, and both are governed by the single discriminant $1 - 4p$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree at

$$P(b, a_1, \dots, a_n) = [x^b] \prod_{j=1}^{n} \big(px^{a_j} + px^{-a_j} + 1 - 2p\big). \tag{†}$$

At (†) the natural attempt *expands*: it multiplies the product out, term by term, trying to compare two explicit sums of positive quantities. The model chain *factors*: it writes the single factor $f(x)$ as const $\cdot\, (x + \alpha)(x^{-1} + \alpha)$ and thereby writes the whole product as $g(x)g(x^{-1})$.

The split is between reading (†) as a probability computation and reading it as a question about the *multiplicative structure of a Laurent polynomial*. And the payoff is immediate: for a product of the form $g(x)g(1/x)$, the constant term is a sum of squares and every other coefficient is a shifted inner product, so Cauchy–Schwarz finishes in one line — and the only obstruction to being in that form is whether a discriminant is nonnegative. The threshold in the answer *is* that discriminant.

> *The answer is going to be a specific number, and no amount of small-case checking will produce it. So I need a criterion, not an estimate. Where could a sharp threshold in p possibly come from? A sharp threshold is a sign change, and a sign change in a formula built out of p is a discriminant. Which quadratic in this problem involves p?*

> *The distribution is a convolution, so the generating function is a product. The constant term of a product is the hard thing; but if I could write the product as $g(x)g(x^{-1})$, its constant term would be $\sum c_i^2$ and its b-th coefficient $\sum c_i c_{i+b}$ — and then the inequality I want is literally Cauchy–Schwarz. So the whole problem reduces to: when is F of that form?*

> *F is a product over j, so it is enough to factor one factor. Set $px + px^{-1} + 1 - 2p = \frac{p}{\alpha}(x+\alpha)(x^{-1}+\alpha)$ and solve for α: I get $p\alpha^2 - (1-2p)\alpha + p = 0$, discriminant $(1-2p)^2 - 4p^2 = 1 - 4p$. There it is. The threshold $p = 1/4$ is not a probabilistic fact at all — it is the point where a quadratic loses its real roots.*

> *For necessity I need one bad family, and I want the sparsest possible target: weights for which 0 is attainable in* only one *way, so that $P(0)$ is as small as it can be. Binary weights $1, 2, 4, \dots, 2^{n-1}$ do exactly that, because binary representation is unique. Then $P(0) = (1-2p)^n$ decays like a pure geometric, while $P(1)$ picks up n terms — a competition I can actually resolve asymptotically.*

- ▷ **Cue:** a probability that a weighted sum of independent lattice variables hits a given value ⟶ **Move:** write it as a coefficient of a Laurent polynomial $\prod_j f(x^{a_j})$; coefficient extraction, not conditioning, is now the whole problem
- ▷ **Cue:** "is the coefficient at 0 the largest coefficient?" for a symmetric Laurent polynomial $F(x) = F(1/x)$ ⟶ **Move:** try to write $F(x) = g(x)g(1/x)$; then $[x^0]F = \sum c_i^2$ and $[x^b]F = \sum c_i c_{i+b}$, and Cauchy–Schwarz gives the inequality for free
- ▷ **Cue:** a sharp threshold in a parameter is expected in the answer ⟶ **Move:** look for a quadratic in that parameter somewhere in the structure and compute its discriminant; sharp thresholds are sign changes, and sign changes of real solvability are discriminants
- ▷ **Cue:** need a counterexample family for a "for all n, all a_j" statement ⟶ **Move:** choose weights making the target value uniquely representable — binary weights 2^{j-1} — so the probability of that target is a bare geometric term, then let $n \to \infty$
- ▷ **Cue:** nonnegative-coefficient identity needed for a characteristic function ⟶ **Move:** φ of a symmetric three-point law is $(1-2p) + 2p\cos\theta$; nonnegativity on the whole circle is a comparison of the constant against the amplitude, $1 - 2p \geq 2p$

Layer 3 THE WHY

Why must the factorisation be thought of? Because of what independence *is*.

Independence means the law of the sum is the convolution of the laws. Convolution is a product in one and only one place: the transform. So the moment you accept the hypothesis of independence, the object you are actually holding is a *product*, and $P(b, \dots)$ is a coefficient of that product. That is the axiom, and it is unavoidable: any argument about this problem that works term by term in probability space is fighting the structure, because probability space is

where the convolution is a sum over many paths and the transform is where it is a product over few factors.

Once you accept that you hold a product, ask what products are good for. Not for extracting coefficients — expanding a product is exactly as expensive as summing over paths, which is the thing you were trying to escape. Products are good for *factoring*. And the question "is $[x^0]F$ the largest coefficient of F?" has one classical sufficient condition attached to it: that $F = g \cdot \bar{g}$, because then the coefficients of F are inner products of a fixed vector with its own shifts, and a vector's inner product with itself beats its inner product with any shift of itself. That is Cauchy–Schwarz, and it is the reason the autocorrelation of any signal peaks at zero lag. This problem is that statement and nothing else: *is the law of S an autocorrelation?*

Now the discriminant. The factor $f(x) = px + px^{-1} + (1-2p)$ is a single lattice step. Asking whether f is an autocorrelation — whether $f(x) = c(x+\alpha)(x^{-1}+\alpha)$ — is asking whether the quadratic $p\alpha^2 - (1-2p)\alpha + p$ has a real root, and its discriminant is

$$(1-2p)^2 - 4p^2 = 1 - 4p.$$

So $p = 1/4$ is not a probabilistic threshold that happens to be a nice number. It is the exact point at which a single step of this random walk stops being an autocorrelation of anything. On the Fourier side the same statement reads $\varphi_{X_1}(\theta) = (1-2p) + 2p\cos\theta \geq 0$, i.e. the constant part dominates the oscillating part: $1 - 2p \geq 2p$. Fejér–Riesz says these are equivalent — a nonnegative trigonometric polynomial is $|g(e^{i\theta})|^2$ — and the archive's point is that you do not need to know Fejér–Riesz by name. You need to know that "coefficient at 0 is largest" smells like autocorrelation, and then try to factor.

Why does the failure above $1/4$ take the shape it does? Because once φ_{X_1} dips negative on part of the circle, you may *amplify* that dip by composing with $a_j\theta$ for cleverly chosen a_j and multiplying many copies together. Binary weights are the extreme case: they make the n factors oscillate at frequencies $1, 2, 4, \ldots$, which is exactly the recipe for a product that is large where you do not want it. In probability language, binary weights make 0 uniquely representable, so $P(0)$ has no help from multiplicity while $P(1)$ collects n distinct paths. Two descriptions, one phenomenon.

Kedlaya's solution introduces $\alpha = \frac{1-2p+\sqrt{1-4p}}{2p}$ with no comment on where it comes from, and Peres's alternative introduces the characteristic function as a tool. Here is where both come from: they are the two standard ways of asking whether the step distribution is a nonnegative-definite object, and $1-4p$ is the single number governing the answer. Set the discriminant to zero and the problem's answer falls out before any probability is computed.

Caveat. I derived the necessity family and both sufficiency arguments independently and checked the threshold numerically: by exact rational computation over all $(X_1, \ldots, X_n) \in \{0, \pm 1\}^n$ with $a_j = 2^{j-1}$, the inequality first fails at $n = 4$ for $p = 0.26$, at $n = 2$ for $p \in \{0.28, 0.3, 1/3\}$, at $n = 1$ for $p \geq 0.4$, and does not fail for $n \leq 9$ at $p \in \{0.1, 0.2, 0.25\}$ — consistent with the threshold $1/4$. Two errata in the published solution, which I have silently corrected above and flag here so a reader comparing the two is not confused: it states $\varphi_{X+Y}(\theta) = \varphi_X(\theta) + \varphi_Y(\theta)$ where the correct (and used) statement is the *product*, and it writes the inversion formula with $\frac{1}{2}\int_0^{2\pi}$ where the constant is $\frac{1}{2\pi}$. The published necessity argument also splits into three cases ($p > 1/3$, $p = 1/3$, $1/4 < p < 1/3$); the binary family handles all three at once, as shown above, and I have verified that unification directly.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is "independent" next to "probability that a weighted sum equals b". On seeing it, write $P(b, a_1, \ldots, a_n) = [x^b]\prod_j (px^{a_j} + px^{-a_j} + 1 - 2p)$ immediately. Independence

means convolution means product; coefficient extraction is now the entire problem and probability language can be put away.

2. The cue suggests generating functions, and then a second, sharper cue takes over: the question "is the coefficient at 0 the largest?" is the autocorrelation question. That suggests factoring $F(x) = g(x)g(1/x)$, at which point Cauchy–Schwarz is the method.

3. Hesitation occurs when you notice that the product has no cancellation — all coefficients are positive — so nothing you know about comparing sums applies. What reinforces away from it: stop looking at the expanded product and look at the factor. The product is a product of n copies of one shape, so any structural property of that one shape propagates. Ask what you want to be true of the product, then ask what that requires of a single factor.

4. The trap is unimodality. "Symmetric unimodal sums are maximised at the centre" is a true theorem on $\mathbb{R}$ and a false one on $\mathbb{Z}$ under scaling, and the problem is built precisely on that gap — the weights a_j are there to spread the lattice out. A second trap: believing small cases, since every small case with p slightly above $1/4$ still satisfies the inequality.

5. Recognisable faster next time by this pairing: *a sharp threshold is asked for in a parameter* and *the structure is a product depending on that parameter*. Sharp thresholds in such problems are where a real factorisation dies, so compute the discriminant of the relevant quadratic first, before any estimation. Here that single computation, $(1-2p)^2 - 4p^2 = 1 - 4p$, gives the answer in one line and the proof is then reverse-engineered from it.

6. Reusable decision rule: *when asked whether a distribution's mode is at* 0, *ask whether the distribution is an autocorrelation.* Concretely: write the law as a Laurent polynomial, attempt $F(x) = g(x)g(1/x)$, and if it succeeds the inequality is Cauchy–Schwarz. Equivalently, check whether the characteristic function is nonnegative on the whole circle. This rule transfers wholesale to any "peak at the origin" question — lattice walks, correlation of signals, coefficient bounds for products of symmetric polynomials — and it tells you where the counterexamples live too: exactly where the factorisation fails.

2022 B6

2022 B6. Find all continuous functions $f : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$f(xf(y)) + f(yf(x)) = 1 + f(x+y)$$

for all $x, y > 0$.

Fields: analysis (functional equations with a continuity hypothesis); asymptotics as the extraction mechanism **Position:** B6 — the last problem of the paper and one of the hardest functional equations the Putnam has set; there is no substitution that finishes it, and the whole solution is a controlled study of f near 0 and near ∞ **Answer:** $f(x) = \dfrac{1}{1+cx}$ for some constant $c \geq 0$, the case $c = 0$ being the constant function 1.

Layer 1 THE WHAT — surface analysis

The natural attempt. You do what you always do with a functional equation: substitute. $y = x$ is the only symmetric substitution available and it gives something real:

$$2f(xf(x)) = 1 + f(2x). \tag{$*$}$$

That is genuinely useful and will stay with us. After that the well runs dry fast. There is no $f(0)$ to plug in — the domain is $\mathbb{R}^+$, open at 0 — so the usual opening move of a functional-equation problem is unavailable, and that absence is deliberate. There is no obvious fixed point to chase: if $f(a) = 1$ for some a then setting $y = a$ gives $f(x) + f(af(x)) = 1 + f(x+a)$, which relates three unknown values and closes on none of them.

So you guess. $f(x) = 1$ works. Trying $f(x) = 1/(1+cx)$:

$$f(xf(y)) = \frac{1}{1+\frac{cx}{1+cy}} = \frac{1+cy}{1+cx+cy},$$

and adding the mirror term gives $\frac{2+cx+cy}{1+c(x+y)} = 1 + \frac{1}{1+c(x+y)} = 1 + f(x+y)$. So the whole family works, and the answer is presumably exactly this family. Now you must prove there is nothing else, and the substitution toolkit has nothing left to offer.

The wall is sharp and worth naming precisely. Every standard technique for functional equations — injectivity, surjectivity, Cauchy's equation, a clever double substitution — extracts *algebraic* information. This equation has almost none: it mixes $f \circ f$ (in $f(xf(y))$) with f of a sum, and those two structures do not commute with anything. What the problem does give you, and what no other Putnam functional equation that year gives you, is the word **continuous**. That word is not there to exclude pathologies at the end. It is the hypothesis the entire solution runs on.

The model chain. Throughout write $g(x) := xf(x)$, which is continuous and positive. The plan: pin down the behaviour of f at 0^+, then at ∞, then read off f from the asymptotics.

Step 1: $f(x) \to 1$ as $x \to 0^+$. Fix any $y > 0$. From the equation, $f(xf(y)) \le 1 + f(x+y)$ for all x. As $x \to 0^+$ the argument $xf(y)$ runs over a punctured right-neighbourhood of 0 (a fixed positive multiple of x), so writing $L_\pm$ for the lim sup and lim inf of f at 0^+,

$$L_+ \le 1 + \lim_{x\to 0^+} f(x+y) = 1 + f(y) < \infty,$$

using continuity of f at y. So f is bounded near 0, hence $g(x) = xf(x) \to 0$ as $x \to 0^+$.

Now feed $(*)$. First a surjectivity remark that will be used three times: g is continuous on $(0,\infty)$ with $g(0^+) = 0$, so for every $\delta > 0$ and every $t \in (0, g(\delta))$ there is $x \in (0,\delta)$ with $g(x) = t$ — by the intermediate value theorem on $(0,\delta]$. Consequently $\limsup_{x\to 0^+} f(g(x)) = L_+$ and likewise for lim inf: the values $g(x)$ for small x sweep out all small values of the argument. Then $(*)$ gives

$$2L_+ = \limsup_{x\to 0^+} \big(1 + f(2x)\big) = 1 + L_+, \qquad 2L_- = 1 + L_-,$$

so $L_+ = L_- = 1$. Hence $\lim_{x\to 0^+} f(x) = 1$.

Step 2: if $f \ge 1$ everywhere, then $f \equiv 1$. Suppose $f(x) \ge 1$ for all x. Put $S_c = \sup\{f(x) : 0 < x \le c\}$, finite for finite c (continuity on $[\varepsilon, c]$ plus Step 1 near 0), and $S_\infty = \sup_{x>0} f(x) \in [1,\infty]$.

First, $S_\infty < \infty$. If $f \equiv 1$ there is nothing to prove, so take y with $f(y) > 1$. The equation rearranges to $f(x+y) - f(xf(y)) = f(yf(x)) - 1 \ge 0$, so $f(xf(y)) \le f(x+y)$. Letting x run over $(0, c-y]$, the left argument runs over $(0,(c-y)f(y)]$ and the right argument over $(y,c]$, whence $S_{(c-y)f(y)} \le S_c$. Put $c_0 = \frac{yf(y)}{f(y)-1}$; for $c \ge c_0$ one checks $(c-y)f(y) \ge c$, so in fact $S_{(c-y)f(y)} = S_c$. Since $(c-y)f(y) - c_0 = f(y)(c-c_0)$ and $f(y) > 1$, iterating from any $c > c_0$ sends the cutoff to infinity while S never changes. Hence $S_\infty = S_c < \infty$.

Second, $S_\infty \le 1$. Since $g(0^+) = 0$ and $g(x) \ge x \to \infty$, g is onto $\mathbb{R}^+$ by the intermediate value theorem. So by $(*)$,

$$S_\infty = \sup_{t>0} f(t) = \sup_{x>0} f(g(x)) = \sup_{x>0} \tfrac{1}{2}\big(1 + f(2x)\big) = \tfrac{1+S_\infty}{2},$$

giving $S_\infty = 1$. Combined with $f \ge 1$: $f \equiv 1$. This is the $c = 0$ solution.

Step 3: otherwise $f(x) \to 0$ as $x \to \infty$. Assume from now on that $f(x_0) < 1$ for some x_0, and let $I = \inf_{x>0} f(x) < 1$. Fix $\varepsilon \in \big(0, \frac{1-I}{2}\big)$ and y with $f(y) < I + \varepsilon$. Then $g(x) \ne y$ for every x: otherwise $(*)$ would give

$$1 + I \le 1 + f(2x) = 2f(g(x)) = 2f(y) < 2I + 2\varepsilon < 1 + I,$$

a contradiction. Since g is continuous with $g(0^+) = 0 < y$ and never equals y, the intermediate value theorem forces $g(x) < y$ for all x. Thus $xf(x) < y$, so $f(x) < y/x \to 0$.

Step 4: normalise. By Steps 1 and 3 and continuity, f attains the value $\frac{1}{2}$, and $f^{-1}(\frac{1}{2})$ is closed and bounded away from both 0 and ∞, hence compact and nonempty. If f solves the equation so does $x \mapsto f(cx)$ for any $c > 0$ (substitute $u = cx$, $v = cy$ throughout). Rescale so that the *least* element of $f^{-1}(\frac{1}{2})$ is 1; so $f(1) = \frac{1}{2}$ and $f(t) \neq \frac{1}{2}$ for $0 < t < 1$.

Step 5: $xf(x) \to 1$ as $x \to \infty$. By $(*)$, $f(g(x)) = \frac{1}{2}(1 + f(2x)) > \frac{1}{2} = f(1)$, so $g(x) \neq 1$ for every x; as in Step 3, $g(0^+) = 0$ forces $g(x) < 1$ for all x. Also $f(2x) \to 0$, so $f(g(x)) \to \frac{1}{2}$.

Now suppose $g(x_n) \leq 1 - \delta$ along some $x_n \to \infty$. On $(0, 1-\delta]$ the function f never takes the value $\frac{1}{2}$ (minimality of 1) and tends to 1 at 0^+, so by the intermediate value theorem $f > \frac{1}{2}$ throughout $(0, 1-\delta]$; and $\inf_{(0,1-\delta]} f > \frac{1}{2}$ strictly, since $f \geq \min_{[\eta,1-\delta]} f > \frac{1}{2}$ by compactness and $f > \frac{3}{4}$ on $(0, \eta]$ for small η by Step 1. That contradicts $f(g(x_n)) \to \frac{1}{2}$. Hence $g(x) \to 1$: $f(x) \sim 1/x$.

Step 6: the reflection $f(1/y) + f(y) = 1$. Replace y by xy in the original equation:

$$f\big(xf(xy)\big) + f\big(xy\, f(x)\big) = 1 + f\big(x(1+y)\big).$$

Let $x \to \infty$ with y fixed. Then $xf(xy) = \frac{1}{y} \cdot g(xy) \to \frac{1}{y}$ and $xyf(x) = y\, g(x) \to y$, so by continuity of f the left side tends to $f(1/y) + f(y)$; and $f\big(x(1+y)\big) \to 0$. Hence

$$f(1/y) + f(y) = 1 \qquad (y > 0).$$

Step 7: finish. Using the reflection to rewrite $f(yf(x)) = 1 - f\big(\frac{1}{yf(x)}\big)$, the original equation becomes

$$f(xf(y)) = f(x+y) + f\left(\frac{1}{yf(x)}\right).$$

Multiply through by $xf(y)$ and let $x \to \infty$ with y fixed, using $tf(t) \to 1$ three times:

- left side: with $u = xf(y) \to \infty$, $xf(y)\, f(xf(y)) = uf(u) \to 1$;
- first right term: $xf(y)f(x+y) = f(y) \cdot \frac{x}{x+y} \cdot (x+y)f(x+y) \to f(y)$;
- second right term: with $w = \frac{1}{yf(x)} = \frac{x}{y\, g(x)} \to \infty$, we have $f(w) \sim 1/w = yf(x)$, so $xf(y)f(w) \sim f(y)\, y\, xf(x) \to yf(y)$.

Therefore $1 = f(y) + yf(y)$, i.e. $f(y) = \dfrac{1}{1+y}$ for all $y > 0$. Undoing the rescaling of Step 4 gives $f(x) = \dfrac{1}{1+cx}$ with $c > 0$, and Step 2 supplies $c = 0$. All of these were checked to work at the start. □

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree through $(*)$: $2f(xf(x)) = 1 + f(2x)$. At $(*)$ the natural attempt reads it as an *equation* — an identity relating unknown values of f, to be attacked by further substitution. The model chain reads it as a *limit statement*: the left side controls f near wherever $xf(x)$ goes, and the right side controls f near $2x$, so $(*)$ is a transport rule that carries information about f on one part of $\mathbb{R}^+$ to information about f on another.

The consequence of that reading is a specific object: $g(x) = xf(x)$. Once you decide to read $(*)$ as transport, g is not a clever auxiliary function; it is literally the argument on the left-hand side, and everything you need to know is where g sends things. Every one of the six steps above is a statement about g: $g(0^+) = 0$, g surjective, g misses a value, $g < 1$, $g \to 1$.

> *There is no $f(0)$ to substitute, and the problem has handed me the word "continuous". Those two facts point the same way: the information in this problem lives at the* boundary *of the domain, at 0^+ and at ∞, and continuity is the instrument for reading it. I should stop substituting and start taking limits.*

> *$(*)$ says $2f(xf(x)) = 1 + f(2x)$. The argument on the left is $xf(x)$ and on the right is $2x$. As $x \to 0^+$ both are small; as $x \to \infty$ the right one is large but I have no idea about the left one. So the entire content of $(*)$ is:* where does $xf(x)$ go? *Name it g and answer that question.*

> *g is continuous. A continuous function on an interval that starts near 0 cannot jump over a value — so if I can show g never* equals *some particular y, I have shown g stays below y forever, and that is a bound on f, namely $f(x) < y/x$. Proving a value is missed is much easier than proving a bound directly: I only have to derive a contradiction from a single equation.*

> *Now suppose I know $xf(x) \to 1$. Then f is asymptotically $1/x$ at infinity, so any expression of the form $x \cdot f($something that grows like $\lambda x)$ has a computable limit $1/\lambda$. That means I can take limits of the functional equation and land on an identity in the* finite *variable y alone. Substituting $y \mapsto xy$ first is what makes both inner arguments converge to finite places instead of to 0 or ∞.*

> *The normalisation: the equation is invariant under $f(x) \mapsto f(cx)$, and the answer family is a one-parameter family, so the parameter is exactly this scaling. Kill it by fixing one value of f, and choose the* smallest *preimage of $\frac{1}{2}$ so that I get a one-sided statement I can push against with the intermediate value theorem.*

- ▷ **Cue:** a functional equation on $\mathbb{R}^+$ with no value at the boundary and an explicit continuity hypothesis $\longrightarrow$ **Move:** stop substituting; the intended content is asymptotic. Establish $\lim_{x\to 0^+} f$ and $\lim_{x\to\infty} f$ first, and treat continuity as the main tool rather than a technicality
- ▷ **Cue:** the substitution $y = x$ produces $f(\varphi(x)) =$ something(x) for an explicit φ $\longrightarrow$ **Move:** φ is now the central object; study its range and its limits at both ends, because the equation can only transport information along φ
- ▷ **Cue:** a continuous function φ on an interval with $\varphi(0^+) = 0$, and you want an upper bound on φ $\longrightarrow$ **Move:** show φ never attains the value y; the intermediate value theorem upgrades "never equals" to "always less than" for free
- ▷ **Cue:** a functional equation invariant under $f(x) \mapsto f(cx)$ and an expected one-parameter answer $\longrightarrow$ **Move:** normalise by fixing one value, e.g. the least t with $f(t) = \frac{1}{2}$ set to 1; "least" is worth more than "some", because it gives a one-sided exclusion to push against
- ▷ **Cue:** you know $xf(x) \to L$ at infinity and the equation involves f of a product $\longrightarrow$ **Move:** substitute $y \mapsto xy$ so the inner arguments converge to finite limits, then pass to the limit; the equation collapses to an identity in one variable

Layer 3 THE WHY

Why is this an asymptotic problem and not an algebraic one? Look at what the equation is made of.

The left side contains $f(xf(y))$: a composition, f evaluated at a point that depends on f itself. The right side contains $f(x+y)$: f evaluated at a point that depends only on the additive structure of the domain. There is no algebraic identity relating composition to addition — that is precisely what makes the equation hard, and it is also why every substitution beyond $y = x$ fails to close. Compositions and sums only cooperate *asymptotically*: if $f(t) \approx 1/t$ for large t,

then $f(xf(y)) \approx \frac{1}{x} \cdot \frac{1}{f(y)}$ becomes a product of separate factors, and products do cooperate with everything. So the solution has to earn an asymptotic before it can compute anything, and Steps 1–5 are entirely the earning.

Now the deeper "of course". Why $g(x) = xf(x)$, and why does it converge to a constant?

Because of the scaling symmetry. The equation is invariant under $f(x) \mapsto f(cx)$, so the solution set is closed under dilation of the variable, and the answer therefore *must* be a one-parameter family of dilates of a single function. A one-parameter family of dilates of a function that tends to 0 at infinity is pinned down by a single scale-invariant quantity, and the only scale-invariant combination of x and f available — one that is unchanged when you replace $f(x)$ by $f(cx)$ and rescale x — is the product $xf(x)$. So g is not a clever substitution that somebody found. It is the invariant of the symmetry group of the problem, and the statement $g \to 1$ is the statement that the invariant is asymptotically *constant*, which is the only thing a scale-invariant quantity is allowed to do at infinity in a scale-invariant problem. The moment you notice that $f(x) \mapsto f(cx)$ preserves the equation, $xf(x)$ is the object you must study.

Then the structure of the endgame is forced too. Once $tf(t) \to 1$, the correct move is to make the *inner* arguments converge. $f(xf(y))$ with $x \to \infty$ sends its argument to infinity and tells you nothing; but after $y \mapsto xy$ the inner argument $xf(xy)$ tends to $1/y$, a finite number, and continuity of f converts a limit of arguments into a limit of values. This is why continuity is a load-bearing hypothesis and not a nicety: without it, knowing $xf(xy) \to 1/y$ tells you nothing about $f(xf(xy))$. The reflection identity $f(1/y) + f(y) = 1$ that drops out is the problem's real content — it is the statement that f is symmetric about the point $(1, \frac{1}{2})$ under inversion of the variable, exactly the symmetry $\frac{1}{1+y}$ has.

Kedlaya's solution begins by establishing $\lim_{x\to 0^+} f(x) = 1$ from limsup/liminf bookkeeping, then asserts "since $xf(x) \to 0$ as $x \to 0^+$ and $xf(x) \to \infty$ as $x \to \infty$, $xf(x)$ takes all positive real values by the intermediate value theorem". Those two sentences look like technical lemmas. Here is where they come from: the first says f is normalised at the boundary so that the invariant g starts at 0; the second says g is a continuous surjection, which is what licenses transporting any statement about f at one place to f at another via $(*)$. Every later step in the official solution is an application of one of those two facts, and together they are the reason the whole proof is about g and not about f.

Finally, why does $f \geq 1$ have to be quarantined into its own step? Because the argument of Steps 3–7 opens with "let $I = \inf f < 1$" and simply has nothing to say if f never dips below 1. Structurally: $f \equiv 1$ is the fixed point of the dilation symmetry — the one solution that is its own dilate — so it cannot be reached by an argument that uses the dilation to normalise. Degenerate cases in a problem with a symmetry group are exactly the fixed points of that group, and they always need separate handling. That is worth carrying away from this problem more than any of its analysis.

Caveat. I constructed the seven-step chain above independently and then compared with the published solution; the two agree in architecture and I believe the argument to be complete, but three points deserve flagging because the published write-up passes over them. (i) In Step 1, the identity $\limsup_{x\to 0^+} f(g(x)) = L_+$ is not automatic from $g(x) \to 0$ — it needs g to sweep out all small values near 0, which I have supplied by the intermediate value theorem on $(0, \delta]$. (ii) In Step 5, the contradiction requires $\inf f$ over $(0, 1-\delta]$ to exceed $\frac{1}{2}$ *strictly*, which needs a compactness argument on $[\eta, 1-\delta]$ together with $f \to 1$ at 0^+; the published version does not spell this out. (iii) I have verified the answer family satisfies the equation by direct substitution. The published solution contains a typographical slip — it cites equation (B62), the limit statement, where it means (B61), the functional equation — which I have silently corrected in the chain above.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is what is *absent*: the domain is $\mathbb{R}^+$ with 0 excluded, so there is no $f(0)$ to substitute, and the word "continuous" is stated explicitly. On seeing that pair, write down "this is an asymptotic problem" and set as the first goal $\lim_{x\to 0^+} f(x)$ and $\lim_{x\to\infty} f(x)$. The second cue is $y = x$, giving $2f(xf(x)) = 1 + f(2x)$ — write it down within the first minute.

2. Those cues suggest: name $g(x) = xf(x)$ and treat $(*)$ as a transport rule along g. The method is then limit analysis plus the intermediate value theorem, not substitution algebra. Every subsequent step is a question about g: where does it start, what does it hit, what does it miss, where does it end.

3. Hesitation occurs after $(*)$, when no further substitution produces anything. That is the moment the problem is won or lost. What reinforces away from it: count your hypotheses. You have used the equation and nothing else; you have not used continuity, and you have not used that the domain is open at 0. An unused hypothesis in a B6 is not decoration. Go and use it.

4. The trap is treating continuity as a regularity afterthought — the habit from Cauchy-equation problems, where continuity merely kills pathological solutions at the end. Here it is the engine from the first line. A second trap: forgetting the case $f \geq 1$, which is invisible from inside the main argument because the main argument silently assumes $\inf f < 1$, and which contains a genuine solution, $f \equiv 1$.

5. Recognisable faster next time by this pairing: *the equation is invariant under* $f(x) \mapsto f(cx)$ and *the expected answer is a one-parameter family*. Whenever a functional equation has a continuous symmetry group, find the invariant of that group — here $xf(x)$ — and make it the subject of the investigation. The symmetry also tells you in advance that you may normalise one value of f for free, and that the fixed points of the symmetry are separate cases.

6. Reusable decision rule: *when an equation mixes composition with addition, buy an asymptotic before attempting algebra.* Composition and addition do not interact algebraically, but they do interact asymptotically — an asymptotic $f(t) \sim L/t$ turns $f(\text{product})$ into a product of separate factors and lets you pass to the limit inside the equation. Concretely: establish limits at both ends of the domain, then substitute so that every inner argument converges to a *finite* point, then use continuity to move the limit inside f. This is the standard route through hard functional equations with a continuity hypothesis, and it transfers directly to equations of the form $f(xf(y)) = \cdots$, $f(x)f(y) = f(xy) + \cdots$, and any equation where the unknown appears inside its own argument.