DAWN PUTNAM SESSION

Putnam 2024 — Papers A and B

Putnam 120 Archive

Contents

Putnam 2024 — Paper A **5**
2024 A1 5
2024 A2 7
2024 A3 10
2024 A4 13
2024 A5 17
2024 A6 22

Putnam 2024 — Paper B **29**
2024 B1 29
2024 B2 31
2024 B3 34
2024 B4 36
2024 B5 40
2024 B6 46

Putnam 2024 — Paper A

2024 A1

2024 A1. Determine all positive integers n for which there exist positive integers a, b, and c satisfying

$$2a^n + 3b^n = 4c^n.$$

Fields: Number theory (infinite descent via parity, with a residual quadratic-residue argument mod 3) **Position:** A1 – the opener **Answer:** Only $n = 1$ works, e.g. $(a, b, c) = (1, 2, 2)$; every $n \geq 2$ is impossible, by an infinite-descent argument that runs on parity for $n \geq 3$ and needs a separate mod-3 argument to patch the one case, $n = 2$, where the parity descent runs out of room.

Layer 1 THE WHAT — surface analysis

The natural attempt. The equation smells like Fermat: three fixed coefficients $2, 3, 4$ on an n-th power equation, with $n = 1$ trivially solvable ($2a + 3b = 4c$ has solutions everywhere) and every instinct says $n \geq 2$ should fail by an infinite-descent argument. The standard machine for that is: normalise to $\gcd(a, b, c) = 1$ (dividing through by the gcd of all three), then chase parity. From $3b^n = 4c^n - 2a^n$, the right side is even, so b is even. Write $b = 2b'$ and divide the whole equation by 2: $a^n + 2^{n-1} \cdot 3(b')^n = 2c^n$. For $n \geq 2$, 2^{n-1} is even, so this forces a^n even, hence a even; write $a = 2a'$. Substituting and dividing by 2 again: $2^{n-1}(a')^n + 2^{n-2} \cdot 3(b')^n = c^n$. The natural expectation, having gotten b and then a to fall by the same mechanism, is that c falls too, on the same mechanism, for every $n \geq 2$ – giving $\gcd(a, b, c) \geq 2$, contradicting the normalisation, and finishing the problem for all $n \geq 2$ in one uniform argument. Push it through: for this last step to force c even, the coefficient 2^{n-2} must itself be even, i.e. $n - 2 \geq 1$, i.e. $n \geq 3$. Try $n = 2$ specifically and the coefficient is $2^0 = 1$ – odd – so the equation reads $2a'^2 + 3b'^2 = c^2$ with no forced parity on c at all. The uniform descent silently stops working at exactly the value $n = 2$, and no amount of re-massaging the same parity step repairs it, because the obstruction is arithmetic (the power of 2 in the coefficient $3 \cdot 2^{n-2}$), not a mistake in the algebra.

The model chain. Fix $n \geq 2$ and suppose $2a^n + 3b^n = 4c^n$ with a, b, c positive integers; divide through by $\gcd(a, b, c)$ so that $\gcd(a, b, c) = 1$.

Case $n = 2$. From $2a^2 + 3b^2 = 4c^2$, rearrange to $a^2 + c^2 \equiv 0 \pmod 3$. Since the only quadratic residues mod 3 are 0 and 1, a sum of two of them is 0 mod 3 only if both are 0; so $3 \mid a$ and $3 \mid c$. Write $a = 3a''$, $c = 3c''$; substituting and dividing by 9 gives $b^2 = 12(c'')^2 - 6(a'')^2$, so $3 \mid b^2$, hence $3 \mid b$. Then 3 divides all of a, b, c, contradicting $\gcd(a, b, c) = 1$.

Case $n \geq 3$. As in the natural attempt, b is even (from $3b^n = 4c^n - 2a^n$ being even); write $b = 2b'$, divide by 2: $a^n + 3 \cdot 2^{n-1}(b')^n = 2c^n$. Since $n \geq 2$, 2^{n-1} is even, so a^n is even, so a is even; write $a = 2a'$, divide by 2 again: $2^{n-1}(a')^n + 3 \cdot 2^{n-2}(b')^n = c^n$. Now $n \geq 3$ gives $n - 2 \geq 1$, so 2^{n-2} is even, and *both* terms on the left are even, forcing c^n even, hence c even. Then a, b, c are all even, contradicting $\gcd(a, b, c) = 1$.

Both cases contradict the normalisation, so no solution exists for any $n \geq 2$; combined with the explicit $n = 1$ solution, the answer is $n = 1$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains agree, word for word, up to "$2^{n-1}(a')^n + 3 \cdot 2^{n-2}(b')^n = c^n$." At this point the natural attempt expects the same mechanism (a coefficient of 2 forces the next variable even) to fire uniformly for every $n \geq 2$; the model instead checks the exponent $n-2$ on the surviving power of 2 and finds it fails to be ≥ 1 exactly when $n = 2$, forcing a genuine case split rather than a single uniform descent.

> *I got b even and then a even off the same trick – dividing by* 2 *and noticing a coefficient of* 2 *or* 2^{n-1} *forces the next term even – so before I trust that the same trick finishes c, I should actually write out the exponent on the surviving power of* 2 *in this last equation and check it is really* ≥ 1 *for every n I care about, instead of assuming the pattern continues.*

> *The coefficient that just failed me,* 3*, is exactly the one number in the original equation that carries no power of* 2 *at all – so if parity has run out of information at* $n = 2$*, the next place to look for a constraint is arithmetic mod that same* 3*, since it is already sitting in the equation as a coefficient, not an arbitrary prime I have to go hunting for.*

- ▷ **Cue:** an equation like $2a^n + 3b^n = 4c^n$ is being descended on by dividing through by the gcd of a, b, c and chasing parity $\longrightarrow$ **Move:** track, at each division-by-2 step, the exact power of 2 carried by the surviving coefficient (here $3 \cdot 2^{n-2}$), since the descent only continues to force the next variable even while that exponent stays ≥ 1
- ▷ **Cue:** a parity/infinite-descent argument on a family indexed by n silently stops working at one specific value of n $\longrightarrow$ **Move:** do not force the same argument through; isolate that value of n and look for a different modulus already present as a coefficient in the equation (here 3, via quadratic residues mod 3)
- ▷ **Cue:** a Diophantine equation forces $\gcd(a, b, c) |$ every term after normalisation $\longrightarrow$ **Move:** after normalising $\gcd(a, b, c) = 1$, any argument that eventually shows $2 \mid a$, $2 \mid b$, and $2 \mid c$ (or any common prime dividing all three) is a complete contradiction and finishes the problem – no further work is needed once all three are shown divisible by the same prime

Layer 3 THE WHY

Why does the parity descent work uniformly for $n \geq 3$ but need a wholly different modulus at $n = 2$, rather than this being an accident that a cleverer algebraic rearrangement could have avoided? Look at where the powers of 2 in the equation actually live. The original coefficients are $2 = 2^1$, $3 = 3 \cdot 2^0$, $4 = 2^2$. Each division-by-2 step in the descent removes exactly one factor of 2 from every term simultaneously; after two such divisions the a-term has shed its original 2^1 and gained $n-1$ more factors of 2 from raising a''s coefficient to the n-th power (giving 2^{n-1}), while the b-term, which started with a bare 2^0 on its coefficient 3, only ever gains factors of 2 from the same substitution mechanism – and by the second division it is carrying exactly 2^{n-2}. The two division steps are identical operations applied to both terms, so whichever term started with the smaller power of 2 is the one that determines when the whole machine runs out: that term is the $3b^n$ term, and it runs out precisely when $n - 2 < 1$, i.e. $n = 2$. This is not a flaw in the choice of substitution; it is forced by the coefficients $2, 3, 4$ themselves, since 3 is the unique one of the three that is odd. Once parity is exhausted, the next constraint has to come from a prime the equation already distinguishes – and 3 is exactly that prime, both because it appears as a literal coefficient and because $\mathbb{Z}/3\mathbb{Z}$ has the unusually restrictive property that only 0 and 1 occur as squares, which is precisely strong enough to force $3 \mid a$ and $3 \mid c$ from $a^2 + c^2 \equiv 0 \pmod 3$ alone. So the case split at $n = 2$ is not a patch bolted onto the main argument; it is what the equation's own coefficients dictate once you track the power of 2 they carry rather than assume the mechanism generalises unchecked.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the exponent pattern $2 = 2^1$, $3 = 3 \cdot 2^0$, $4 = 2^2$ in the coefficients; on seeing it, write down that 3 is the only odd coefficient, flagging it as the term that will run out of parity information first.
2. That cue suggests a descent by repeatedly dividing the (gcd-normalised) equation by 2, tracking which variable is forced even at each step, and specifically tracking the exact power of 2 surviving on the $3b^n$ term as the step counter, since it is the weakest link.
3. Hesitation occurs exactly after deriving $2^{n-1}(a')^n + 3 \cdot 2^{n-2}(b')^n = c^n$ and being tempted to declare c even by the same pattern that worked for a and b; it is reinforced away by explicitly checking the exponent $n - 2 \geq 1$, which fails only at $n = 2$.
4. The trap is trying to force one single argument to cover all $n \geq 2$ at once, or conversely assuming the mod-3 trick found for $n = 2$ is also needed (or even available in the same form) for larger n – it is not; the two cases are genuinely different arguments, joined only by the shared conclusion.
5. Recognition would be faster next time by treating any coefficient triple like $(2, 3, 4)$ as an instruction to write out each coefficient's power of 2 before starting a parity descent, and to predict in advance exactly which value of the exponent parameter will make the descent stall.
6. The reusable decision rule: in an infinite-descent-by-parity argument on an equation with fixed integer coefficients and a variable exponent n, identify the coefficient with the smallest power of 2 (equivalently, the unique odd coefficient, if there is one) before starting – that coefficient's own exponent structure tells you in advance the exact threshold value of n at which the descent will run out of forced parity, and that threshold case must be handled by a different modulus, typically one already present as a coefficient in the equation.

2024 A2

2024 A2. For which real polynomials p is there a real polynomial q such that

$$p(p(x)) - x = (p(x) - x)^2 q(x)$$

for all real x?

Fields: Algebra (polynomial divisibility via the exact Taylor expansion for polynomials) **Position:** A2 – second in the paper **Answer:** Exactly the polynomials $p(x) = x + c$ and $p(x) = -x + c$, for any real constant c; every other polynomial fails, because a degree ≥ 2 candidate forces $1 + p'(x)$ – which has degree exactly one less than p – to be divisible by $r(x) := p(x) - x$, which has degree exactly equal to that of p, an impossibility unless $r(x) = 0$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The relation is a squared-divisibility condition, so the honest first move is to test it directly on low-degree p. Constant $p(x) = k$ fails outright: $p(p(x)) - x = k - x$ has degree 1, while $(p(x) - x)^2 q(x) = (k - x)^2 q(x)$ has degree ≥ 2 unless $q \equiv 0$, which would force $k = x$ identically – impossible. Linear $p(x) = mx + c$: compute $p(p(x)) - x = (m^2 - 1)x + c(m + 1)$ and $r(x) = (m - 1)x + c$. If $m = 1$: $r(x) = c$ is a nonzero constant (or $r \equiv 0$ if $c = 0$), and $p(p(x)) - x = 2c$, so q can be taken constant – this branch works for every c. If $m = -1$: both $(m^2 - 1)x$ and $c(m + 1)$ vanish, so $p(p(x)) - x \equiv 0$, and $q \equiv 0$ works – this branch also works for every c. For any other m, $p(p(x)) - x$ is a genuine nonzero linear polynomial while $r(x)^2 q(x)$ has degree ≥ 2 unless $q \equiv 0$, which forces $p(p(x)) - x \equiv 0$ identically – but that needs $m^2 = 1$, already covered. So the linear case is fully settled by direct computation: $p(x) = \pm x + c$.

The natural next question is whether some quadratic (or higher) p could also work, and the natural tool to reach for is a degree count on the whole equation. If $\deg p = d \geq 2$, then $\deg p(p(x)) = d^2$, so the left side has degree d^2. On the right, $\deg r(x) = d$ (the leading term of p survives subtracting x once $d \geq 2$), so the right side has degree $2d + \deg q$. Setting these equal gives $\deg q = d^2 - 2d = d(d-2)$. For $d \geq 2$ this is always ≥ 0 – *there is no contradiction here*. Degree-counting on the whole equation is consistent for every $d \geq 2$; it only pins down what degree q would have to be, and gives no reason to rule any d out. A student who expected this comparison to kill $d \geq 2$ outright, the way it killed the genuinely-linear-but-wrong-slope case above, finds instead that the door is still open, and has to go looking for a finer tool.

The model chain. Write $r(x) := p(x) - x$, so the condition is $r(x)^2 \mid p(p(x)) - x$. Since $p(p(x)) = p(x + r(x))$, and for any polynomial p the identity

$$p(x+h) = p(x) + p'(x)\,h + h^2 \cdot (\text{some polynomial in } x, h)$$

holds exactly (it is the polynomial remainder theorem applied twice, or equivalently the binomial expansion of each power $(x+h)^k$ truncated after the linear-in-h term), substituting $h = r(x)$ gives, as an exact polynomial identity in x,

$$p(p(x)) \equiv p(x) + p'(x)\,r(x) \pmod{r(x)^2}.$$

Hence

$$p(p(x)) - x \equiv (p(x) - x) + p'(x)\,r(x) = r(x)\big(1 + p'(x)\big) \pmod{r(x)^2}.$$

The hypothesis says the left side is $\equiv 0 \pmod{r(x)^2}$, so $r(x)\big(1 + p'(x)\big)$ is divisible by $r(x)^2$. If $r(x) \not\equiv 0$, this is equivalent to $r(x) \mid 1 + p'(x)$.

Now compare degrees when $\deg p = d$. If $d = 0$: p constant fails as shown above (in fact $1 + p'(x) = 1$, and $r(x) = p - x$ has degree 1, so 1 cannot be divisible by $r(x)$ unless $r \equiv 0$, impossible for constant $p \neq x$). If $d \geq 2$: $\deg r(x) = d$ exactly (the degree-d leading term of p is untouched by subtracting x), while $\deg(1 + p'(x)) = d - 1$ exactly (the leading coefficient of p' is d times that of p, nonzero). A nonzero polynomial of degree $d-1$ cannot be divisible by a nonzero polynomial of degree d. So $1 + p'(x)$ must be the zero polynomial, i.e. $p'(x) \equiv -1$, i.e. $p(x) = -x + c$ – but that has degree 1, contradicting $d \geq 2$. So no $d \geq 2$ works.

If $d = 1$, $p(x) = mx + c$: $1 + p'(x) = 1 + m$. Either $1 + m = 0$ (giving $p(x) = -x + c$, and then $r(x) = -2x + c$ divides 0 trivially) or $1 + m \neq 0$, a nonzero constant, which is divisible by $r(x)$ only if $r(x)$ is itself a nonzero constant, i.e. $m = 1$ (giving $p(x) = x + c$, any c).

So exactly $p(x) = x + c$ and $p(x) = -x + c$ work, matching (and completing) the linear case found by direct computation, and ruling out every degree ≥ 2 candidate in one stroke.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains agree completely through settling the linear case. The split is at the degree-counting step for $d \geq 2$: "the natural attempt compares only the *top degrees* of the full equation $p(p(x)) - x = r(x)^2 q(x)$, gets $\deg q = d(d-2)$, and finds this consistent for every $d \geq 2$ – a dead end that looks like progress. The model abandons comparing the whole equation's degree and instead extracts the single factor $r(x)$ divides out of $r(x)(1 + p'(x))$, then compares *that* factor's degree, $d - 1$, against $\deg r(x) = d$."

> *Matching the top degrees of both sides only tells me what degree q would need to be – it doesn't force a contradiction for $d \geq 2$, so staring at the whole equation's size is a dead end; I need to know something about the* structure *of $p(p(x)) - x$ near the roots of $r(x)$, not just its overall size.*

> *$p(p(x)) = p(x + r(x))$ is a polynomial in the "increment" $r(x)$, and for polynomials the Taylor expansion in that increment is not an approximation – it's an exact algebraic identity, truncatable mod $r(x)^2$ exactly the way I'd truncate mod h^2 for a small number h. So I can read off a divisibility condition on $1 + p'(x)$ directly, and that is a single polynomial whose degree I can compare against $r(x)$'s degree with no ambiguity left.*

- ▷ **Cue:** a condition requires the *square* of a polynomial difference, $(p(x) - x)^2$, to divide something built by composing p with itself $\longrightarrow$ **Move:** expand the composition as $p(x + r(x))$ and apply the exact polynomial Taylor identity $p(x + h) \equiv p(x) + p'(x)h \pmod{h^2}$, reducing the squared-divisibility condition to a first-derivative divisibility condition
- ▷ **Cue:** comparing the top-level degrees of both sides of a divisibility equation gives no contradiction (the unknown factor's degree just comes out ≥ 0) $\longrightarrow$ **Move:** do not conclude the candidate degree is possible; the degree match only constrains an auxiliary factor – go find a different, more local divisibility condition (e.g. mod the square of a factor) that pins down an actual polynomial, then compare degrees of that specific polynomial instead

Layer 3 THE WHY

Why must the move be "expand mod $r(x)^2$ via the derivative," rather than being a trick pulled from nowhere once degree-counting on the whole equation fails to close the case? Because differentiation and subtraction behave completely predictably with respect to degree, while a bare comparison of the whole equation's degree does not carry enough information once an unknown factor $q(x)$ is present to absorb any degree discrepancy. Differentiating a degree-d polynomial drops its degree by exactly 1 (the leading coefficient survives, scaled by $d \neq 0$); subtracting x from a degree-$d \geq 2$ polynomial changes nothing about its degree (the degree-d leading term cannot be touched by a degree-1 subtraction). These are not properties special to this problem – they are the two most elementary facts about how degree behaves under the operations already sitting in the problem's own definitions ($p'(x)$, and $r(x) = p(x) - x$). Once those two facts are lined up against each other – $\deg(1 + p'(x)) = d - 1$ versus $\deg r(x) = d$ – the third fact that finishes the argument is equally elementary: a nonzero polynomial cannot divide another nonzero polynomial of strictly smaller degree. So the reason the derivative-based expansion succeeds where raw degree-matching on the whole equation fails is that it isolates a divisibility statement between *two specific polynomials whose degrees are pinned down by definitions already in the problem*, rather than a statement involving an unconstrained auxiliary factor $q(x)$ that can silently absorb any degree gap. The squared divisibility in the problem statement is, in retrospect, exactly built to be unpacked this way: a condition mod $r(x)^2$ is custom-made for extracting a first-order (derivative) statement mod $r(x)^1$, the same way a double root of a polynomial is custom-made for being detected by its derivative.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the squared factor $(p(x) - x)^2$ in the divisibility condition; on seeing it, write down $r(x) := p(x) - x$ and restate the problem as "for which p is $p(p(x)) - x$ divisible by $r(x)^2$", since a squared divisor is the standard signal for a derivative-based (Taylor) argument.

2. That cue suggests expanding $p(p(x)) = p(x + r(x))$ using the exact polynomial Taylor identity $p(x + h) \equiv p(x) + p'(x)h \pmod{h^2}$, rather than attempting a direct degree count on the full original equation.

3. Hesitation occurs right after the degree count $\deg q = d(d - 2)$ comes out nonnegative for every $d \geq 2$ and looks, misleadingly, like progress toward ruling out higher degrees; it is reinforced away by recognising that this equation only constrains the unknown factor q, and switching to the derivative expansion, which involves no unknown factor at all.

4. The trap is trusting that a degree comparison on the whole equation will eventually produce a contradiction for $d \geq 2$ if pushed hard enough; it never does; the equation is satisfiable at

the level of pure degree-counting for every $d \geq 2$, and the real obstruction only appears once the specific factor $1 + p'(x)$ is isolated.

5. Recognition would be faster next time by treating any condition of the shape "$g(x)^2$ divides $h(x)$" as an automatic trigger for expanding $h(x)$ via the exact polynomial Taylor identity around wherever g appears as an increment, rather than comparing top degrees of the whole expression.

6. The reusable decision rule: when a problem states a divisibility condition involving the *square* of a polynomial (or more generally the k-th power), do not stop at matching degrees on the two sides of the full equation – that only constrains an auxiliary cofactor. Instead expand the relevant composition as an exact polynomial Taylor series in that factor and extract the divisibility condition one order down (mod the $(k-1)$-th power, here mod $r(x)^1$), which involves specific, degree-determined polynomials (here $1 + p'(x)$ and $r(x)$ itself) whose degrees can be compared directly.

2024 A3

2024 A3. Let S be the set of bijections

$$T\colon \{1,2,3\} \times \{1,2,\ldots,2024\} \to \{1,2,\ldots,6072\}$$

such that $T(1,j) < T(2,j) < T(3,j)$ for all $j \in \{1,2,\ldots,2024\}$ and $T(i,j) < T(i,j+1)$ for all $i \in \{1,2,3\}$ and $j \in \{1,2,\ldots,2023\}$. Do there exist a and c in $\{1,2,3\}$ and b and d in $\{1,2,\ldots,2024\}$ such that the fraction of elements T in S for which $T(a,b) < T(c,d)$ is at least $1/3$ and at most $2/3$?

Fields: Combinatorics (standard Young tableaux of a rectangular shape, hook length formula) **Position:** A3 – third in the paper, a genuine step up in difficulty from A1/A2 **Answer:** Yes: take $(a,b) = (2,1)$, $(c,d) = (1,2)$. S is exactly the set of standard Young tableaux of the 3×2024 rectangle, and $P\big(T(2,1) < T(1,2)\big) = \frac{n+1}{3n-1}$ with $n = 2024$, which equals $\frac{2025}{6071} \approx 0.3335$ – strictly above $1/3$ and comfortably below $2/3$.

Layer 1 THE WHAT — surface analysis

The natural attempt. S is immediately recognisable: a bijection to $\{1,\ldots,3n\}$ (here $n = 2024$) that increases down each column of 3 and along each row of n is precisely a standard Young tableau (SYT) of the $3 \times n$ rectangular shape. The honest first instinct is to look for a pair of cells related by an obvious symmetry, hoping the answer is forced to be exactly $1/2$ (comfortably in $[1/3, 2/3]$) with no computation at all. But the shape is a short, wide rectangle – 3 rows against 2024 columns – and it has no reflection or rotation symmetry that swaps two distinct cells while fixing the shape (a $3 \times n$ rectangle is symmetric only under transposing rows and columns, which would require $n = 3$). So no free symmetry argument is available, and $T(1,1) = 1$ is forced for every $T \in S$ (it is the unique minimum of a doubly-increasing array), which makes any pair involving cell $(1,1)$ degenerate.

The next natural move is to try small cases by hand and look for a pattern. Take $n = 2$: brute-force enumeration (there turn out to be exactly 5 standard tableaux of the 3×2 rectangle) gives, for the specific comparison $T(2,1)$ vs. $T(1,2)$, exactly 3 of the 5 tableaux with $T(2,1) < T(1,2)$ – a ratio of $3/5$. A student stopping here, without a closed form, has one data point and no confidence it generalises to $n = 2024$; worse, without knowing *why* $3/5$ came out (rather than, say, $1/2$), there's a real risk of guessing the ratio drifts toward $1/2$ (by some vague symmetry-of-the-corner intuition) or toward some other boundary as n grows, with nothing to pin it down. Brute enumeration for $n = 2024$ is obviously not an option, so this approach plateaus exactly where it is most needed: at the general n.

The model chain. Since $T(1,1) = 1$ always, the next-smallest value, 2, must go to one of the two cells that directly "cover" $(1,1)$ in the partial order forced by the increasing conditions: $(2,1)$ (the cell below it) or $(1,2)$ (the cell to its right). No other cell can hold 2, since every other cell has some predecessor besides $(1,1)$ still needing a smaller value. So S splits into exactly two disjoint pieces: $T(2,1) = 2$, or $T(1,2) = 2$; and the event $T(2,1) < T(1,2)$ is *exactly* the event $T(2,1) = 2$ (if $T(1,2) = 2$ instead, then $T(2,1) > T(1,2)$ automatically, since 2 is the globally second-smallest value and cannot appear anywhere else).

Condition on $T(2,1) = 2$. Delete the cells $(1,1)$ and $(2,1)$ (holding values $1,2$) and relabel every remaining value $v \mapsto v-2$; this is an order-preserving bijection between $\{T \in S : T(2,1) = 2\}$ and standard tableaux of the shape left after removing those two cells – row 1 loses its first cell, row 2 loses its first cell, row 3 is untouched, giving row lengths $(n-1, n-1, n)$ after re-sorting into decreasing order, i.e. a skew-shape count equal (via the reflection used in the official solution, reversing row order and relabelling $i \mapsto 3n+1-i$) to $f(n, n-1, n-1)$, the number of SYT with row lengths $n, n-1, n-1$ under the general three-row hook length formula

$$f(n_1,n_2,n_3) = \frac{(n_1-n_2+1)(n_1-n_3+2)(n_2-n_3+1)\,(n_1+n_2+n_3)!}{(n_1+2)!\,(n_2+1)!\,n_3!}.$$

Caveat. This general three-row hook length formula, and its proof by induction on $n_1+n_2+n_3$, is standard combinatorial machinery (stated and proved by induction in the official Kedlaya–Bhargava–Ng solution, citing the Glass–Ng elementary proof of the hook length formula for the general case). It is not rederived from first principles here; instead its consequence is checked directly against hand enumeration below.

Plugging in $(n_1,n_2,n_3) = (n, n-1, n-1)$ and (n,n,n) and simplifying the ratio gives

$$P\big(T(2,1) < T(1,2)\big) = \frac{f(n,n-1,n-1)}{f(n,n,n)} = \frac{(2)(3)(n+1)n}{(1)(2)(3n)(3n-1)} = \frac{n+1}{3n-1}.$$

Direct verification at $n = 2$. By hand: the 5 standard tableaux of the 3×2 rectangle are (writing each as its two columns $(T_{11}, T_{21}, T_{31} \mid T_{12}, T_{22}, T_{32})$)

$$(1,2,5 \mid 3,4,6),\ (1,2,4 \mid 3,5,6),\ (1,2,3 \mid 4,5,6),\ (1,3,5 \mid 2,4,6),\ (1,3,4 \mid 2,5,6);$$

the first three have $T(2,1) = 2$, hence $T(2,1) < T(1,2)$; the last two have $T(1,2) = 2$, hence $T(2,1) > T(1,2)$. That is 3 out of 5, matching $\frac{n+1}{3n-1}\big|_{n=2} = \frac{3}{5}$ exactly. This confirms the formula on a case small enough to check completely by hand, giving confidence in it for $n = 2024$ without needing to trust the general hook length identity blind.

Finally, check the bound: $\frac{n+1}{3n-1} > \frac{1}{3} \iff 3(n+1) > 3n-1 \iff 3 > -1$, always true; and $\frac{n+1}{3n-1} < \frac{2}{3} \iff 3(n+1) < 2(3n-1) \iff 3n+3 < 6n-2 \iff n > 5/3$, true for all $n \geq 2$. At $n = 2024$: $\frac{2025}{6071} \approx 0.33350$, strictly inside $[1/3, 2/3]$. So $(a,b,c,d) = (2,1,1,2)$ works.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains agree on recognising S as SYT of a $3 \times n$ rectangle, and both correctly single out the pair $(2,1), (1,2)$ as the natural place to start, since $(1,1)$ is forced. The split is what happens next with that pair: "the natural attempt tries to *feel out* the probability – by symmetry intuition, or by computing one small case and hoping to extrapolate – while the model *computes* it, by converting the event $T(2,1) = 2$ into an exact sub-counting problem (deleting two cells and relabelling) and evaluating both counts with the same tool, the hook length formula, applied to two different shapes."

> *$T(1,1)$ is always 1, so the value 2 has nowhere else to go except the two cells that cover it – $(2,1)$ and $(1,2)$ – which means "is $T(2,1) < T(1,2)$" is not an arbitrary comparison between two cells; it is exactly the same event as "does the tableau extend downward or rightward from the corner first," a genuinely binary and structurally forced choice that I should recognise as such immediately, rather than treating it as just one comparison among many I could have picked.*

> *I don't need to guess whether this probability is near $1/2$ from the shape of the corner – conditioning on $T(2,1) = 2$ deletes two specific cells and relabels everything else, which is itself a smaller SYT-counting problem of a describable shape, so I should compute the numerator and denominator with the* same *formula (hook length) applied to two different shapes, rather than reasoning about the comparison directly or trusting an intuition about symmetry that this non-square shape doesn't actually have.*

> *Before trusting the closed form $\frac{n+1}{3n-1}$ for $n = 2024$, I should check it against a case small enough to enumerate completely by hand ($n = 2$ gives 5 tableaux, a list short enough to write out and count directly) – this is cheap insurance against a mis-transcribed hook length computation, and it is exactly the kind of check that turns "I trust the formula" into "I verified the formula."*

- ▷ **Cue:** a random object is built by filling a partially-ordered grid (here a Young-diagram-shaped array) with all values from 1 to N, uniformly among valid fillings, and the question compares two specific, order-adjacent cells $\longrightarrow$ **Move:** recognise that the smallest values are forced into a small set of "coverable" cells right after the global minimum, so a comparison between two such cells reduces to conditioning on which one receives the next value, which becomes an exact sub-counting problem after deleting those cells and relabelling
- ▷ **Cue:** a count of standard Young tableaux (or similar combinatorial fillings) needs to be compared as a ratio between a full shape and a shape with a few cells removed $\longrightarrow$ **Move:** apply the same closed-form counting formula (here, the hook length formula) to both shapes and simplify the ratio algebraically, rather than reasoning about the probability directly
- ▷ **Cue:** a closed-form combinatorial formula is being trusted for a large parameter (here $n = 2024$) $\longrightarrow$ **Move:** before committing to it, evaluate it at the smallest nontrivial parameter value where the object can be enumerated completely by hand, and check the formula's output against the direct count

Layer 3 THE WHY

Why does conditioning on "where does the value 2 go" convert this probability into an exact, computable ratio, rather than this being a lucky trick specific to this shape? Because of what a standard Young tableau *is*: a linear extension of the partial order on the cells of the shape. The first k values placed by any linear extension always occupy a sub-shape that is itself a valid (smaller) Young diagram nested inside the original – this is not a special property of the $3 \times n$ rectangle, it is what "linear extension of a partial order defined by a grid" means. So conditioning on the position of an early value is always, in principle, equivalent to conditioning on which smaller sub-shape the first few values trace out, and once that sub-shape is fixed, the remaining problem is counting linear extensions (tableaux) of the complementary region – exactly the kind of count the hook length formula (or its skew generalisation) is built to produce. The reason $(1,1) \to \{(2,1),(1,2)\}$ specifically is the right place to enter this machinery is that it is the *first* point in the tableau where the sub-shape genuinely branches – with only one cell filled, there are exactly two candidate cells for the next one, the minimal possible branching, which is what keeps the resulting sub-counting problem small enough to have a closed form at all (row lengths $(n, n-1, n-1)$, differing from the original (n,n,n) in only two cells).

Why is the resulting probability biased toward $1/3$, not $1/2$? Because the shape's aspect ratio – 3 rows against 2024 columns – is extreme. A tableau has vastly more freedom to extend a row

(there are 2024 possible row-positions to grow into) than to extend a column (only 3 rows exist, so a column can only be extended twice past its first cell). The hook length formula compresses this asymmetry into the rational function $\frac{n+1}{3n-1} \to \frac{1}{3}$ as $n \to \infty$: the "3" in the denominator's leading term is literally the row count, and the limiting bias of $1/3$ (with $2/3$ its complement) is exactly the value the problem's stated bounds $[1/3, 2/3]$ sit against – confirmation that the problem is testing precisely whether the solver can locate a pair whose probability sits strictly on the correct side of those two limiting values, not arbitrarily close to 0 or 1.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that $T(1,1) = 1$ is forced by the increasing conditions, and that the value 2 can only go to one of the (at most two) cells covering $(1,1)$ – write this down the instant the shape is recognised as an SYT-counting problem.

2. That cue suggests treating "does $T(2,1) < T(1,2)$" as identical to the event "$T(2,1) = 2$", and computing its probability by deleting the two smallest cells and relabelling, turning it into a ratio of two hook-length-formula evaluations on two related shapes.

3. Hesitation occurs at translating "$T(2,1) = 2$" into the residual shape's row lengths $(n-1, n-1, n)$ correctly (getting the reflection/relabelling right); it is reinforced away by writing out the $n = 2$ case fully by hand and checking the residual-shape count against direct enumeration before trusting the general formula for $n = 2024$.

4. The trap is assuming, from the "symmetry" of a corner having two neighbouring cells, that the probability should land near $1/2$; the shape is a 3×2024 rectangle, not square, so no such symmetry exists, and chasing an exact-$1/2$ answer (or trying to prove one via a bijective symmetry argument) wastes time on something the shape's aspect ratio already rules out.

5. Recognition would be faster next time by treating any probability question over uniformly random linear extensions of a partially ordered grid as an automatic trigger for "the first few values trace out a nested sub-shape; condition on that sub-shape and count both pieces with the same closed-form counting tool."

6. The reusable decision rule: for a uniformly random filling of a partial order (a Young diagram, a tree, any poset) with $1, \ldots, N$ respecting the order, a question about which of two order-adjacent, "next available" cells receives an early value reduces to a ratio of counts of the structure with those early cells deleted and relabelled; evaluate both counts with the same closed-form formula the problem's total count already requires, and sanity-check the resulting closed form against the smallest case you can enumerate completely by hand before trusting it at the problem's actual (large) parameter.

2024 A4

2024 A4. Find all primes $p > 5$ for which there exists an integer a and an integer r satisfying $1 \le r \le p-1$ with the following property: the sequence $1, a, a^2, \ldots, a^{p-5}$ can be rearranged to form a sequence $b_0, b_1, b_2, \ldots, b_{p-5}$ such that $b_n - b_{n-1} - r$ is divisible by p for $1 \le n \le p-5$.

Fields: number theory: multiplicative order, primitive roots, arithmetic progressions in $\mathbb{Z}/p$ **Position:** A4 — the first problem of the paper where the winning move is a change of viewpoint rather than a technique, and where a careless argument loses the single prime that actually works **Answer:** $p = 7$, and no other prime.

Layer 1 THE WHAT — surface analysis

The natural attempt. Strip the wording first. The condition $b_n - b_{n-1} \equiv r \pmod{p}$ for every n says that $b_0, \ldots, b_{p-5}$ is an arithmetic progression modulo p with common difference r:

$b_n \equiv b_0 + nr$. And the b_n are a rearrangement of $1, a, a^2, \ldots, a^{p-5}$, a geometric progression. So the question is: for which p can a geometric progression of length $p-4$ in $\mathbb{Z}/p$ be reordered into an arithmetic progression?

The honest default now is to attack the two progressions with symmetric functions, because that is the standard way to say "these two sets are equal" without knowing the matching. Sum both sides:

$$\sum_{k=0}^{p-5} a^k = \frac{a^{p-4}-1}{a-1}, \qquad \sum_{n=0}^{p-5}(b_0+nr) = (p-4)b_0 + r\binom{p-4}{2},$$

and equate mod p. That is one equation in the three unknowns a, b_0, r. Take products next: $\prod a^k = a^{(p-5)(p-4)/2}$, and $\prod(b_0+nr)$ — and here the wall arrives. That product is r^{p-4} times a product of $p-4$ consecutive integers in a shifted scale, which is a truncated Wilson-type product with no closed form you can use. You can push to the second power sum; it also closes on nothing. After twenty minutes you have two true equations with three unknowns and no mechanism for generating more.

The alternative default is a search. Fix $p = 11$: the sequence has 7 terms, and you would need to test each a and each r. Doing it by hand for $p = 11$ takes several minutes and returns nothing; $p = 13$ takes longer and also returns nothing; and you are no closer to a proof for all p. This is zombie maths in its number-theoretic form: manufacturing instances of a universally quantified statement.

The model chain. *Step 1: the terms are distinct.* Since $1 \le r \le p-1$, r is invertible mod p, so $b_0, b_0+r, \ldots, b_0+(p-5)r$ are $p-4$ *distinct* residues. Hence $1, a, \ldots, a^{p-5}$ are pairwise distinct mod p. In particular $p \nmid a$, and the multiplicative order d of a satisfies $d > p-5$.

Step 2: a is a primitive root when $p \ge 11$. We have $d \mid p-1$. Any proper divisor of $p-1$ is at most $(p-1)/2$, and $(p-1)/2 \le p-5$ exactly when $p \ge 9$. So for $p \ge 11$, $d > p-5$ forces $d = p-1$: a is a primitive root mod p. (For $p = 7$ this step is unavailable, and that is exactly where the answer hides.)

Step 3: compare complements. Take $p \ge 11$. Since a is a primitive root, $a^0, a^1, \ldots, a^{p-2}$ runs through all of $\mathbb{F}_p^\times$ once. Our list stops at a^{p-5}, so it omits precisely $a^{p-4}, a^{p-3}, a^{p-2}$, that is a^{-3}, a^{-2}, a^{-1}. As a subset of the whole of $\mathbb{Z}/p$ it therefore omits exactly

$$\{\, 0,\ a^{-1},\ a^{-2},\ a^{-3} \,\}.$$

On the other side, $b_0 + nr$ for $n = 0, 1, \ldots, p-1$ runs through all of $\mathbb{Z}/p$ once (again because r is invertible). Our list stops at $n = p-5$, so it omits exactly $b_0+(p-4)r, \ldots, b_0+(p-1)r$, that is

$$\{\, b_0-4r,\ b_0-3r,\ b_0-2r,\ b_0-r \,\}.$$

The two lists are equal as sets, so the two complements are equal:

$$\{b_0-4r, b_0-3r, b_0-2r, b_0-r\} = \{0, c, c^2, c^3\}, \qquad c := a^{-1}.$$

The left-hand set is a four-term arithmetic progression with common difference r. So $0, c, c^2, c^3$ can be rearranged into a four-term arithmetic progression in $\mathbb{Z}/p$.

Step 4: normalise. Multiplying an arithmetic progression by the invertible constant c^{-1} gives another arithmetic progression, so $0, 1, c, c^2$ can be rearranged into a four-term arithmetic progression.

Step 5: 0 is an endpoint. If 0 were an interior term, its two neighbours u, v would satisfy $u + v = 2 \cdot 0 = 0$, so two elements of $\{1, c, c^2\}$ would be additive inverses. The three possibilities are $1 + c = 0$, $1 + c^2 = 0$, $c + c^2 = 0$, giving $c = -1$, $c^2 = -1$, $c = -1$ respectively; these force $\mathrm{ord}(c) \mid 2$ or $\mathrm{ord}(c) \mid 4$. But $\mathrm{ord}(c) = \mathrm{ord}(a) = p-1 \ge 10$. Contradiction.

Step 6: six cases. Reversing the progression if necessary, it is $0, x, 2x, 3x$ with $\{x, 2x, 3x\} = \{1, c, c^2\}$.

- $x = 1$: then $\{c, c^2\} = \{2, 3\}$. If $c = 2$ then $c^2 = 4 \neq 3$. If $c = 3$ then $c^2 = 9 \equiv 2$ forces $p \mid 7$.
- $x = c$: then $\{2c, 3c\} = \{1, c^2\}$. If $2c = 1, 3c = c^2$ then $c = 3$ and $6 \equiv 1$, so $p \mid 5$. If $2c = c^2, 3c = 1$ then $c = 2$ and $6 \equiv 1$, so $p \mid 5$.
- $x = c^2$: then $\{2c^2, 3c^2\} = \{1, c\}$. If $2c^2 = 1, 3c^2 = c$ then $3c = 1$ and $2c^2 = 1$, giving $2 \equiv 9$, so $p \mid 7$. If $2c^2 = c, 3c^2 = 1$ then $2c = 1$ and $3c^2 = 1$, giving $3 \equiv 4$, so $p \mid 1$: impossible.

Every case forces $p \in \{5, 7\}$, both excluded by $p \geq 11$. So no prime $p > 7$ works.

Step 7: $p = 7$ works. Here $p - 5 = 2$, so the sequence is $1, a, a^2$. Take $a = 3$: modulo 7 these are $1, 3, 2$. Rearranged as $b_0, b_1, b_2 = 1, 2, 3$ they form an arithmetic progression with $r = 1$, and $1 \leq r \leq 6$. Hence $p = 7$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree through Step 1: the $p - 4$ residues are distinct, because r is invertible. They split immediately after. At (†) — "I now know the geometric progression consists of $p - 4$ distinct residues" — the natural attempt *describes the set*, by summing it, multiplying it, taking power sums. The model *counts what is absent*. Four residues are missing from each description, and a set of size four is an object you can hold in your hand; a set of size $p - 4$ is not.

> *$p - 4$ out of p. That is not a set I should describe — it is a set I should describe by what it is missing, and what it is missing has exactly four elements. Let me write down the four absentees on each side and set them equal.*

> *An arithmetic progression with invertible common difference does not stop; it wraps. If I let n run to $p - 1$ instead of $p - 5$ it covers all of $\mathbb{Z}/p$. So a progression of length $p - 4$ is literally the complement of a progression of length 4 — and that four-term progression is the thing I am actually being asked about.*

> *The order of a is bigger than $p - 5$ and divides $p - 1$; the only divisor of $p - 1$ that large is $p - 1$ itself once $p \geq 11$. So a is a primitive root, which means I know exactly which residues are absent: three consecutive negative powers, plus 0, which is absent for free because it is never a power.*

> *I may divide the whole configuration by c: scaling an arithmetic progression by a nonzero constant leaves it an arithmetic progression. That removes one parameter at no cost, so I should do it before the case analysis rather than after.*

> *$\mathrm{ord}(a) = p - 1$ is large, so a satisfies no short multiplicative relation. Every case I am about to write down produces a short relation like $c^2 = 2c$ or $c = -1$. Each one is therefore a bound on p, not a case to be solved.*

Layer 3 THE WHY

Why must the complement thought be thought? Because of what an arithmetic progression *is* in $\mathbb{Z}/p$. An arithmetic progression with invertible difference r is an initial segment of the cyclic ordering $b_0, b_0 + r, b_0 + 2r, \ldots$ of the whole group; after p steps it returns to the start, having visited every residue once. That is not a property of progressions; it is the definition of r being invertible. So the phrase "an arithmetic progression of length $p - 4$ in $\mathbb{Z}/p$" does not name a long list. It names a short list — the four terms you did not take — which is itself an

arithmetic progression with the same difference. There is nothing to discover here: the moment you have written "r is invertible", the complement is the only compression available, and it compresses $p-4$ unknowns into 4.

The multiplicative side does the same thing for the same reason. If a is a primitive root, $k \mapsto a^k$ is a bijection $\mathbb{Z}/(p-1) \to \mathbb{F}_p^\times$, so the powers $a^0, \dots, a^{p-5}$ are an initial segment of a cyclic ordering too, and they omit the three wrap-around terms a^{-1}, a^{-2}, a^{-3}. Adding the compulsory absentee 0 gives four. *Both sides of the problem are cyclic orderings of $\mathbb{Z}/p$ truncated near the end, and truncation near the end is the same as deletion of a short block.* Once said aloud, the reduction to "can $0, c, c^2, c^3$ be an arithmetic progression?" is not a step; it is a translation.

This also explains the problem's numerology, which otherwise looks arbitrary. Why a^{p-5}? Why not a^{p-2} or a^{p-8}? Because the setter is choosing the size of the complement. At a^{p-2} the complement is $\{0\}$ and the condition is empty. At a^{p-3} it is $\{0, a^{-1}\}$: any two residues are an arithmetic progression, so still empty. At a^{p-4} it is $\{0, a^{-1}, a^{-2}\}$, and three points are an arithmetic progression as soon as one is the average of the other two — a real but weak condition, satisfiable for infinitely many p. At a^{p-5} the complement has four elements, and four is the first length at which "is an arithmetic progression" is rigid enough to pin p down to a finite list. The exponent in the problem statement is the setter telling you the size of the object you are supposed to build.

Then the endgame runs on one principle, used twice. A primitive root has order $p-1$, so it satisfies *no* multiplicative relation of small height: $c^i = -c^j$ forces $\mathrm{ord}(c) \mid 2(i-j)$, and $c^i = \lambda c^j$ with $\lambda \in \{2, 3\}$ forces an explicit congruence in p. Step 5 is the first use (ruling out 0 in the interior), Step 6 is the second (six orderings, each yielding two such relations). Neither is case-bashing for its own sake: each case is one line because the compression in Steps 3 and 4 already did the work. *Any short relation among powers of a primitive root is a bound on p.* That sentence is the whole of the second half.

Finally, the reason $p = 7$ survives is worth saying plainly, because it is where most attempts die. Step 2 needs $(p-1)/2 \le p-5$, i.e. $p \ge 9$. At $p = 7$ the sequence has only three terms and a need not be a primitive root — indeed $a = 3$ has order 6 but $a = 2$ has order 3 and also works, since $1, 2, 4$ is an arithmetic progression with $r = 2$. An argument that proves "no primitive root can do this" and concludes "no prime works" has thrown away the answer. The hypothesis $p > 5$ in the statement is a warning label: the setter has already removed the small primes where the argument breaks, except for the one they want you to find.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the pair of numbers $p-5$ and $1 \le r \le p-1$. On seeing them, write down two lines: (i) r is invertible mod p, so the $p-4$ progression terms are distinct residues; (ii) $p-4$ terms out of p means a complement of size 4. Do not write down a single symmetric function.

2. The cue suggests complement counting inside $\mathbb{Z}/p$, backed by order arithmetic: bound $\mathrm{ord}(a)$ below by the number of distinct powers, then use $\mathrm{ord}(a) \mid p-1$ to force a primitive root, then read off exactly which residues are absent.

3. Hesitation occurs at two places. First, after proving a is a primitive root, you have a fact and no use for it; the reinforcement is that primitivity is precisely a statement about *which residues are missing*, and missing residues are what you are comparing. Second, at the six-case enumeration, which feels like defeat. The reinforcement is arithmetic: six cases at one line each is cheaper than any structural dodge, because Steps 3–5 have already reduced four unknowns to one.

4. Three traps. (i) The big one: proving that no prime with a primitive root argument works, and concluding that no prime works — Step 2 requires $p \ge 11$, and the answer lives at $p = 7$. Always ask which inequality your "general" argument used, and check the primes below it by hand. (ii) Forgetting that 0 belongs to the complement of the powers; the complement then has three elements and the condition evaporates. (iii) Treating the b_n as residues from

the start — they are the actual integers $1, a, \ldots, a^{p-5}$, and the distinctness in Step 1 is a deduction, not an assumption.

5. Recognisable faster next time via the slogan: *a set of size $N-k$ inside a structured universe of size N is a set of size k*. Whenever two structured families are asserted to be equal and both are nearly everything, stop comparing the families and compare their complements. The same instinct handles "all but three residues are quadratic residues", "a permutation fixes all but four points", "a polynomial agrees with another at all but k points".

6. The reusable decision rule: **when a problem hands you a long list drawn from a finite structure, first ask how much of the structure is left out; if the answer is a bounded number, work with the omissions.** A second rule for $\mathbb{F}_p$ specifically: **an element of large multiplicative order satisfies no short relation, so every relation you derive among a few of its powers is a congruence bounding p.** Together they turn "find all primes" into "list the finitely many relations, read off the finitely many primes, then check the small primes the general argument excluded".

▷ **Cue:** a list of $p-k$ residues mod p is claimed to be an arithmetic progression, with k small $\longrightarrow$ **Move:** let the progression run the full p steps: it then covers $\mathbb{Z}/p$ exactly once, so the given list is the complement of a k-term arithmetic progression with the same common difference

▷ **Cue:** the powers $1, a, \ldots, a^m$ are forced to be distinct mod p with $m > (p-1)/2$ $\longrightarrow$ **Move:** the order of a exceeds $(p-1)/2$ and divides $p-1$, hence equals $p-1$: a is a primitive root, and you now know exactly which residues the list omits

▷ **Cue:** the problem specifies $1 \le r \le p-1$ rather than just "an integer r" $\longrightarrow$ **Move:** read it as "r is invertible mod p" and immediately deduce that the progression has distinct terms; this is the hypothesis the whole argument runs on

▷ **Cue:** a small set must be simultaneously a geometric progression and an arithmetic progression $\longrightarrow$ **Move:** scale by the inverse of one of its elements — an arithmetic progression stays one — to normalise a power to 1 and cut the number of free parameters

▷ **Cue:** you derive a relation such as $c^i = -c^j$ or $c^i = \lambda c^j$ for a primitive root c and a small integer λ $\longrightarrow$ **Move:** convert it into a divisibility statement about $\mathrm{ord}(c) = p-1$ or an explicit congruence in p; short relations kill all large p at once

▷ **Cue:** a general argument used an inequality like $p \ge 11$ or $n \ge 5$ $\longrightarrow$ **Move:** test every case below the threshold by hand before answering "none" — the setter usually put the answer there

2024 A5

2024 A5. Consider a circle Ω with radius 9 and center at the origin $(0,0)$, and a disc Δ with radius 1 and center at $(r,0)$, where $0 \le r \le 8$. Two points P and Q are chosen independently and uniformly at random on Ω. Which value(s) of r minimize the probability that the chord $\overline{PQ}$ intersects Δ?

Fields: geometric probability: random chords, the direction–offset parametrisation, the arcsine law
Position: A5 — a problem that is short to state, has a one-word answer, and punishes the single most common wrong instinct in geometric probability **Answer:** $r = 0$, and no other value: the probability is strictly increasing in r on $[0,8]$.

Layer 1 THE WHAT — surface analysis

The natural attempt. Set up coordinates. Write $P = (9\cos\alpha, 9\sin\alpha)$ and $Q = (9\cos\beta, 9\sin\beta)$ with α, β independent and uniform on $[0, 2\pi)$. The chord meets Δ when the distance from $(r,0)$

to the segment $\overline{PQ}$ is at most 1. So the answer is

$$\Pr(r) = \frac{1}{4\pi^2} \iint \mathbf{1}\big[\operatorname{dist}\big((r,0), \overline{PQ}\big) \le 1\big]\, d\alpha\, d\beta,$$

and now you must evaluate a double integral of an indicator function whose region is described by a distance-to-a-segment formula. That is not a calculation anybody wants; and differentiating in r under the integral sign does nothing useful, because the integrand is an indicator and its r-derivative is a measure on the boundary of the region, which you would then have to parametrise anyway.

The well-prepared student does better than this and reparametrises the chord. A chord of a circle is determined by its perpendicular direction and its signed distance from the centre, so write the chord as "direction θ, offset u". Now you must decide the joint law of (θ, u), and here the natural instinct fires: *θ is uniform and u is uniform on* $(-9, 9)$. And with that assumption the calculation is over in one line and gives the wrong answer. The chord meets Δ exactly when u lies within 1 of the projection of $(r, 0)$ onto the offset axis, i.e. within an interval of length 2; a length-2 interval has the same uniform measure wherever it sits; so $\Pr(r)$ is *constant*. Every r would be a minimiser, and the problem would have no content. That is the wall, and it is a good one: the contradiction is with the existence of the question, not with any computation.

The student who avoids that trap — because they know Bertrand's paradox, or because they check — gets the correct offset law and then faces the official solution's route: a difference of two arccosines, integrated over θ, differentiated in r under the integral sign. It works. It also costs a page of algebra with $(80 \pm 2r\cos t - r^2\cos^2 t)^{-1/2}$ in it, and one sign slip loses the problem.

The model chain. Step 0: segment or line? Since $r \le 8$, the disc Δ lies inside Ω, in fact $\Delta \subseteq \{x^2 + y^2 \le 81\}$. If the line through P and Q meets Δ, the intersection points lie inside Ω, and every point of that line inside Ω lies on the chord $\overline{PQ}$. So "the chord meets Δ" and "the line PQ meets Δ" are the same event, and we may work with lines.

Step 1: the right coordinates. Put $\theta = \frac{1}{2}(\alpha + \beta)$ and $\varphi = \frac{1}{2}(\alpha - \beta)$. This is a linear change of variables with constant Jacobian, so θ and φ are independent and uniform on their ranges; folding by the symmetries $\varphi \mapsto -\varphi$ and $\theta \mapsto \theta + \pi$, we may take θ uniform on an interval of length π and φ uniform on $(0, \pi)$. Rotating the whole picture by $-\theta$ about the origin sends P, Q to $(9\cos\varphi, \pm 9\sin\varphi)$ — so the chord becomes the vertical line $x = 9\cos\varphi$ — and sends the centre of Δ to $(r\cos\theta, -r\sin\theta)$.

Step 2: the one-dimensional problem. In the rotated picture the chord meets Δ precisely when $|9\cos\varphi - r\cos\theta| \le 1$. Write $s := r\cos\theta$ for the *shadow* of the centre of Δ on the offset axis, and let $u := 9\cos\varphi$ be the chord's signed offset. Since φ is uniform on $(0, \pi)$, substituting $u = 9\cos\varphi$ gives u the density

$$w(u) = \frac{1}{\pi\sqrt{81 - u^2}}, \qquad -9 < u < 9,$$

the arcsine law on $(-9, 9)$. Hence, conditionally on θ,

$$\Pr\big(\text{chord meets } \Delta \mid \theta\big) \;=\; G(s) := \int_{s-1}^{s+1} w(u)\, du, \qquad s = r\cos\theta,$$

and $\Pr(r) = \mathbb{E}_\theta\big[G(r\cos\theta)\big]$. (Expanding the integral recovers the official solution's $\frac{1}{\pi}\big[\cos^{-1}\frac{s-1}{9} - \cos^{-1}\frac{s+1}{9}\big]$, but we will not need the closed form.)

Step 3: where should a window of width 2 *sit?* The density w is even, and on $(-9, 9)$

$$w'(u) = \frac{u}{\pi(81 - u^2)^{3/2}},$$

so w is strictly increasing in $|u|$: the arcsine density piles up at the rim. Therefore for $0 \le s < 8$,

$$G'(s) = w(s+1) - w(s-1) > 0,$$

because $|s+1| > |s-1|$ when $s > 0$, and $G'(0) = 0$. So G is even and strictly increasing on $[0,8)$; by continuity of G on $[0,8]$ it is strictly increasing on all of $[0,8]$, and $G(s) > G(0)$ for every s with $0 < |s| \le 8$.

Step 4: average over θ. For $r > 0$ we have $|r\cos\theta| \le r \le 8$, with $r\cos\theta \ne 0$ for all but two values of θ. Hence $G(r\cos\theta) \ge G(0)$ pointwise, with strict inequality except on a null set, and so

$$\Pr(r) = \mathbb{E}_\theta\big[G(r\cos\theta)\big] > G(0) = \Pr(0)$$

for every $r \in (0,8]$. The minimum is attained at $r = 0$ and nowhere else.

Numerical confirmation. $\Pr(0) \approx 0.07088$, $\Pr(1) \approx 0.07111$, $\Pr(4) \approx 0.07492$, $\Pr(8) \approx 0.10801$; a direct Monte Carlo simulation of the original two-endpoint model reproduces these to three decimal places.

The same fact without any integral. Step 3 has a purely geometric reading, which is the official second solution: two parallel lines at distance 2 cut two congruent arcs off a circle of radius 9, and the arc length is minimised when the strip is centred on the centre of the circle. Arc length is an increasing function of the chord joining the arc's endpoints; that chord joins a point on one line to a point on the other, so it is shortest when perpendicular to them, which happens exactly in the centred position.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree up to the reparametrisation: a chord is (direction, offset), and by rotational symmetry the direction is uniform and independent of everything else. They split at (†), the moment you must name the law of the offset. The natural attempt *assumes the offset is uniform*, because that is what "uniformly at random" has been about for the past two lines. The model *derives* the offset law from the model that the problem actually specifies — two uniform endpoints — and finds the arcsine density $1/(\pi\sqrt{81-u^2})$, which is not uniform and whose non-uniformity is the entire problem.

There is a second, smaller bifurcation further down. Both paths reach $f(r,\theta) = \cos^{-1}\frac{s-1}{9} - \cos^{-1}\frac{s+1}{9}$. The natural attempt differentiates that expression in r; the model reads it as $\int_{s-1}^{s+1} w$, a fixed-width window against a fixed density, and then only ever differentiates the two endpoints.

> *If the offset were uniform then the answer would not depend on r at all, and the problem would be asking me to choose among equals. A competition does not ask a vacuous question. So the offset is not uniform, and I had better work out what it actually is before I do anything else.*

> *The two arccosines are not two functions to be differentiated. They are the two endpoints of one integral:* $\cos^{-1}\frac{s-1}{9} - \cos^{-1}\frac{s+1}{9} = \int_{s-1}^{s+1}\frac{du}{\sqrt{81-u^2}}$. *Written that way the question is not "what is the derivative of this messy expression" but "where should a window of width* 2 *sit to collect the least mass".*

> *A window of fixed width moving across a density only cares about its two ends:* $\frac{d}{ds}\int_{s-1}^{s+1} w = w(s+1) - w(s-1)$. *So I need one fact about w and nothing else — whether it increases with* $|u|$.

> *Once I have fixed the chord's direction, the disc contributes nothing to the problem except its shadow on the offset axis: an interval of length* 2 *centred at* $r\cos\theta$. *The disc's shape is irrelevant, only its*

> *width is, and the only thing r controls is how far that shadow sits from the origin.*

> *Whatever the θ-average does, it cannot rescue $r > 0$: $|r\cos\theta| \geq 0$ with equality only when $r = 0$. If the conditional probability is an increasing function of $|r\cos\theta|$ then the comparison is pointwise in θ and I never have to perform the outer integral.*

Layer 3 THE WHY

Why must the offset thought be thought? Because of what a chord *is*. A chord of Ω is the intersection of Ω's disc with a line, and a line in the plane is exactly two numbers: a direction and a signed distance from the origin. So the sample space of this problem is a space of lines, and any description of it in terms of two endpoints is a description in the wrong coordinates — one that carries a redundancy (which endpoint is P) and hides the only structure that matters. The moment you say "a chord is a line, and a line is (direction, offset)", the disc Δ stops being a disc: for each fixed direction, the only thing a convex set does to a line of that direction is occupy an interval of offsets, namely its shadow, of length equal to its width. Δ has width 2 in every direction. So the entire problem is: *place an interval of length* 2 *on the offset axis so as to collect as little probability as possible.* The centre of Δ enters only through $r\cos\theta$, the position of that shadow. That is the level of "of course" the analysis must reach, and it is reached by definitions alone.

Then the second question answers itself: what is the law of the offset? This is where the problem does its damage, and the reason is a fact about the map $\varphi \mapsto 9\cos\varphi$, not about probability. That map is flat at $\varphi = 0$ and $\varphi = \pi$ — its derivative $-9\sin\varphi$ vanishes there — so a fixed amount of angle near the ends is compressed into a tiny range of offsets near ± 9. Equal angle, unequal offset: the density of the offset must blow up at the rim. Writing it out, $w(u) = 1/(\pi\sqrt{81 - u^2})$, and $\sqrt{81 - u^2}$ is nothing but the half-length of the chord at offset u. So the statement "short chords are over-represented" is the statement "w is large where the chord is short", and both are the same statement as "the endpoints, not the offsets, are what is uniform". This is the content of Bertrand's paradox, and the problem has resolved the paradox for you in its own statement — "P and Q are chosen independently and uniformly on Ω" — which is why the ambiguity is not an excuse.

Now the answer is forced. The least likely offsets are the ones near $u = 0$, because that is where w is smallest. A disc of radius 1 centred at the origin is hit exactly by the chords with $|u| \leq 1$ — the least likely band there is. Move the disc outwards and, for each direction, its shadow slides towards the rim where the density is higher; the probability can only go up. Hence $r = 0$.

This is worth pressing, because it fights an intuition that feels unassailable: *surely the centre is where all the chords go*. What is true is that every chord, in every direction, passes somewhere along the offset axis, and the centre is the one point every direction's axis contains. What is false is the inference that the offset is therefore likely to be near zero. The correct picture is that most chords are near-tangent: pick two uniform points on the circle and they are, more often than not, close together on the circumference, giving a short chord far from the centre. The intuition confuses "the centre is geometrically central" with "the centre is probabilistically central", and the arcsine density is the precise statement that it is not.

Finally, the official solution's opening — "first restrict P, Q to points on Ω such that the segment $\overline{PQ}$ makes an angle of θ with the y axis" — can now be read as inevitable rather than inspired. Kedlaya conditions on the direction because the problem is invariant under rotation in everything except the position of Δ; conditioning on the one coordinate that the symmetry does *not* fix reduces a two-dimensional problem to a one-dimensional one with a parameter, and the parameter enters only as $r\cos\theta$. Here is where that comes from: it is the standard move of quotienting by the symmetry group of everything except the obstacle. And the reason

the archive's version is shorter than the official one is that Kedlaya then differentiates the closed form, while the window reading differentiates only the endpoints — the same derivative, computed before the antiderivative is taken rather than after.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is "two points chosen independently and uniformly on Ω" together with "the chord intersects Δ". On seeing it, write down two lines. (i) A chord is a line, and a line is (direction, offset); by rotational symmetry the direction is uniform and independent of the offset. (ii) Derive the offset law from the stated model: with P, Q at angles α, β, the offset is $9\cos\frac{\alpha-\beta}{2}$, so it is arcsine-distributed with density $1/(\pi\sqrt{81-u^2})$ — *not* uniform.

2. That cue suggests conditioning on the direction and reducing to a one-dimensional problem: a fixed-width window (the shadow of the convex obstacle) placed against a fixed density, with the parameter controlling only the window's position.

3. Hesitation occurs at exactly two places. First, at the offset law, where Bertrand's paradox has trained people to believe the question is ill-posed; the reinforcement is that the problem *specifies* the model, so derive the law from it and move on. Second, at the difference of arccosines, where the r-derivative looks unpleasant; the reinforcement is to rewrite the difference as a single definite integral before differentiating, after which the derivative is two values of the integrand.

4. The trap is the uniform-offset assumption, and it is set deliberately: the word "uniformly" appears in the statement, attached to the endpoints, and the reader transfers it to the offset. The tell that you have fallen in is that the answer becomes independent of r — the problem becomes vacuous. A second, smaller trap is worrying about segment versus line; check once that $r+1 \le 9$ makes them equivalent, record it, and stop thinking about it. A third is answering "$r = 8$", by the intuition that the disc should hide near the rim; the numbers go the other way.

5. Recognisable faster next time via the slogan: *a random line meeting a convex set depends on the set only through its width in each direction*. Any question of the form "how likely is a random line/chord/needle to meet this object" should be immediately rewritten in (direction, offset) coordinates, with the object replaced by its shadow. Buffon's needle, Cauchy's formula for the perimeter of a convex body, and this problem are the same computation.

6. The reusable decision rule: **to optimise the position of a fixed-width window against a fixed density, never differentiate the cumulative function — compare the density at the two moving endpoints.** The derivative of $\int_{s-a}^{s+a} w$ is $w(s+a) - w(s-a)$, so the whole optimisation reduces to knowing how w varies, and for an even density increasing in $|u|$ the answer is always "centre the window". A companion rule: **when a modelling assumption makes the answer independent of the problem's parameter, the assumption is wrong** — a well-posed question does not have a vacuous answer, so use the vacuity as a diagnostic.

▷ **Cue:** two points are chosen independently and uniformly on a circle, and a property of the chord is asked for $\longrightarrow$ **Move:** reparametrise as (direction, offset): the direction is uniform and independent, and the offset is $R\cos\varphi$ with φ uniform, hence arcsine-distributed with density $1/(\pi\sqrt{R^2-u^2})$

▷ **Cue:** a random line or chord must meet a small convex set $\longrightarrow$ **Move:** replace the set by its shadow on the offset axis for each direction — an interval whose length is the set's width in that direction; the set's shape then drops out of the problem

▷ **Cue:** an expression is a difference of two values of the same inverse function, such as $\cos^{-1}A - \cos^{-1}B$ $\longrightarrow$ **Move:** rewrite it as one definite integral $\int_B^A$ of the derivative; a difference of antiderivatives is a window against a density, which is a far easier object to optimise

▷ **Cue:** a window of fixed width must be placed to minimise or maximise the mass it collects $\longrightarrow$ **Move:** differentiate only the endpoints: $\frac{d}{ds}\int_{s-a}^{s+a} w = w(s+a) - w(s-a)$, and read off

the answer from the monotonicity of w

- ▷ **Cue:** your modelling assumption makes the answer independent of the parameter the question asks you to optimise $\longrightarrow$ **Move:** reject the assumption — the vacuity is evidence, not a coincidence — and rederive the law from the sampling procedure the problem actually states
- ▷ **Cue:** the problem is invariant under a symmetry group that moves everything except one obstacle $\longrightarrow$ **Move:** condition on the coordinate the symmetry does not fix (here the chord's direction); the problem drops a dimension and the obstacle enters only through one scalar

2024 A6

2024 A6. Let $c_0, c_1, c_2, \ldots$ be the sequence defined so that

$$\frac{1-3x-\sqrt{1-14x+9x^2}}{4} = \sum_{k=0}^{\infty} c_k x^k$$

for sufficiently small x. For a positive integer n, let A be the n-by-n matrix with i,j-entry c_{i+j-1} for i and j in $\{1,\ldots,n\}$. Find the determinant of A.

Fields: algebraic combinatorics: Hankel determinants, orthogonal polynomials, continued fractions of quadratic functions **Position:** A6 — the hardest problem on the paper, and the one where the entire difficulty is knowing that a certain classical machine exists; the computation, once the machine is named, is half a page **Answer:** $\det A = 10^{n(n-1)/2}$.

Layer 1 THE WHAT — surface analysis

The natural attempt. Compute. That instinct is correct here and should be obeyed immediately, because the answer is not guessable from the statement. Rationalise the square root first: writing $f(x)$ for the left-hand side, $\sqrt{1-14x+9x^2} = 1-3x-4f$, and squaring gives

$$f = x + 3xf + 2f^2.$$

So $c_0 = 0$, $c_1 = 1$, and

$$c_n = 3c_{n-1} + 2\sum_{i=1}^{n-1} c_i c_{n-i} \qquad (n \geq 2),$$

a Catalan-type quadratic recursion. It gives $c_1, \cdots = 1, 5, 35, 295, 2765, 27705, \ldots$ (the large Schröder-like family). Note $c_0 = 0$, so the matrix A has top-left entry $c_1 = 1$ and involves only $c_1, \ldots, c_{2n-1}$.

Now compute determinants:

$$n=1:\ \det = 1; \qquad n=2:\ \det\begin{pmatrix} 1 & 5 \\ 5 & 35 \end{pmatrix} = 10; \qquad n=3:\ \det\begin{pmatrix} 1 & 5 & 35 \\ 5 & 35 & 295 \\ 35 & 295 & 2765 \end{pmatrix} = 1000.$$

So the guess is $10^0, 10^1, 10^3, \ldots$, that is $10^{n(n-1)/2}$; $n = 4$ confirms 10^6. Good. Now prove it — and this is the wall.

The natural proof attempt is row reduction with the recursion. It fails, and the reason it fails is structural rather than arithmetic: the entry of A in position (i,j) depends on i and j only through $i+j$. The matrix therefore contains *no information attached to a single row*. Any row operation you perform has an identical effect on some column operation, and the recursion $c_n = 3c_{n-1} + 2\sum c_i c_{n-i}$ is a convolution, so subtracting a multiple of one row from another never clears an entry — it produces a partial convolution, which is not a smaller instance of

anything. You can do this for an hour. For $n = 3$ you can brute-force it; the general step never arrives.

The second natural attempt is induction on n with a cofactor expansion, hoping for a Dodgson/Desnanot–Jacobi identity relating $\det A_n$, $\det A_{n-1}$ and the bordered minors. This is closer to correct — Hankel determinants do satisfy such identities — but the minors that appear are Hankel determinants of *shifted* sequences, and you now have two families to control instead of one, with no supply of relations between them.

The model chain. *Step 1: shift so the matrix is honestly Hankel.* Put

$$h(x) := \frac{f(x)}{x} = \sum_{k\ge 0} \mu_k x^k, \qquad \mu_k := c_{k+1},$$

so $\mu_0 = 1$, and $A = (\mu_{i+j})_{0\le i,j\le n-1}$, the Hankel matrix of h.

Step 2: name the machine. For a power series $h = \sum_{k\ge 0} \mu_k x^k$ with $\mu_0 = 1$ there is a unique *J-fraction* (Jacobi continued fraction) expansion

$$h(x) = \cfrac{1}{1 - b_0 x - \cfrac{\lambda_1 x^2}{1 - b_1 x - \cfrac{\lambda_2 x^2}{1 - \cdots}}},$$

and then

$$\boxed{\det(\mu_{i+j})_{0\le i,j\le n-1} \;=\; \lambda_1^{n-1}\lambda_2^{n-2}\cdots\lambda_{n-1}^{1}.}$$

This is classical (Wall, *Analytic Theory of Continued Fractions*, Theorems 50.1 and 51.1; equivalently Viennot's theory of orthogonal polynomials and Motzkin paths). Layer 3 explains why it is true and why it is the only sensible thing to look for.

Caveat. The boxed theorem is quoted, not reproved here; a full proof is sketched in Layer 3 via Gram–Schmidt, and the resulting formula has been verified numerically against the direct determinant for $n \le 7$.

Step 3: strip one layer of the continued fraction. Rationalising, with $S := \sqrt{1 - 14x + 9x^2}$,

$$\frac{1}{h} = \frac{x}{f} = \frac{4x}{1 - 3x - S} = \frac{4x(1 - 3x + S)}{(1 - 3x)^2 - S^2} = \frac{4x(1 - 3x + S)}{8x} = \frac{1 - 3x + S}{2}.$$

Hence

$$1 - \frac{1}{h} = \frac{1 + 3x - S}{2}.$$

Expanding $S = 1 - 7x - 20x^2 - 140x^3 - \cdots$ gives $1 - 1/h = 5x + 10x^2 + \cdots$, so $b_0 = 5$, and we must have $\lambda_1 x^2 h_1 = \frac{1+3x-S}{2} - 5x = \frac{1-7x-S}{2}$. Taking $\lambda_1 = 10$ (forced, since $h_1(0) = 1$),

$$h_1(x) = \frac{1 - 7x - S}{20x^2}.$$

Step 4: the continued fraction is periodic from here. Claim $h_1 = \dfrac{1}{1 - 7x - 10x^2 h_1}$, i.e. $10x^2 h_1^2 - (1 - 7x)h_1 + 1 = 0$. Writing $u := 1 - 7x - S$, so $h_1 = u/(20x^2)$, multiply the claimed equation by $40x^2$: it becomes $u^2 - 2(1 - 7x)u + 40x^2 = 0$. Now

$$u^2 = (1 - 7x)^2 - 2(1 - 7x)S + S^2, \qquad 2(1 - 7x)u = 2(1 - 7x)^2 - 2(1 - 7x)S,$$

so

$$u^2 - 2(1-7x)u = S^2 - (1-7x)^2 = (1-14x+9x^2) - (1-14x+49x^2) = -40x^2,$$

which is the claim. So h_1 is a fixed point of the same layer-stripping operation: $b_i = 7$ and $\lambda_i = 10$ for every $i \geq 1$, and

$$h(x) = \cfrac{1}{1 - 5x - \cfrac{10x^2}{1 - 7x - \cfrac{10x^2}{1 - 7x - \cdots}}}.$$

Step 5: read off the determinant. All $\lambda_i = 10$, so

$$\det A = \prod_{k=1}^{n-1} \lambda_k^{n-k} = 10^{(n-1)+(n-2)+\cdots+1} = 10^{n(n-1)/2}.$$

The value $b_0 = 5$ never enters; nor do any of the b_i. Checks: $n = 1, 2, 3, 4$ give $1, 10, 1000, 10^6$, matching the direct computation, and the agreement continues to $n = 7$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree through the small cases and the functional equation $f = x + 3xf + 2f^2$. They split at (†): "I have the numbers $1, 10, 1000$ and a quadratic recursion; now prove the pattern." The natural attempt *returns to the matrix* and tries to reduce it with the recursion. The model *never touches the matrix again*. It observes that a Hankel matrix carries no per-row data, so the determinant must be a function of the generating series alone, and it goes looking for the normal form of a series in which Hankel determinants are visible. That normal form is the continued fraction.

> *This matrix has no rows. Its (i, j) entry depends on i and j only through $i + j$, so there is nothing I can do to row i that is not the same as doing something to column i — the matrix is a disguise for a single object, the linear functional $L(x^k) = \mu_k$. I should stop manipulating the disguise and find the right normal form for the functional.*

> *$1, 10, 1000, 10^6$: the exponents are $0, 1, 3, 6$, which is $n(n-1)/2 = (n-1) + (n-2) + \cdots + 1$. That is not a formula, it is a* product with descending multiplicities. *So there must be a sequence of quantities $\lambda_1, \lambda_2, \ldots$ attached to something like levels, with λ_k appearing $n - k$ times — and in this problem they must all happen to equal* 10.

> *The problem hands me a square root of a quadratic. That means f is a quadratic irrationality over $\mathbb{Q}(x)$ — the exact analogue of $\sqrt{d}$ over $\mathbb{Q}$. Lagrange's theorem says quadratic irrationals have eventually periodic continued fractions. So if I write a continued fraction for f, I should expect it to repeat almost immediately, which is precisely the "all the λ_k are equal" that the numbers are telling me to look for.*

> *To strip a layer I write $1 - 1/h = b_0 x + \lambda_1 x^2 h_1$ and identify h_1. The fear is an infinite regress of new unknown series. The test that ends it in one step: does h_1 satisfy the same equation as h_1 one level down? If yes, I have a fixed point, not a regress.*

> *Before I expand anything: the singularities of h are the roots of $1 - 14x + 9x^2$. Factor it as $(1 - \alpha x)(1 - \beta x)$ with $\alpha, \beta = 7 \pm 2\sqrt{10}$. Those two numbers are $7 \pm 2\sqrt{10}$, so their mean is 7 and a quarter of their difference is $\sqrt{10}$ — and 7 and 10 are exactly the constants I am about to find in the continued fraction. The answer is written on the face of the problem.*

Layer 3 THE WHY

Why is a Hankel determinant a continued-fraction quantity? Not because somebody noticed a coincidence. Drive it to the definitions.

A Hankel matrix (μ_{i+j}) is the *Gram matrix* of the monomials. Define the linear functional L on $\mathbb{R}[x]$ by $L(x^k) = \mu_k$, and the bilinear form $\langle P, Q\rangle := L(PQ)$. Then $\langle x^i, x^j\rangle = L(x^{i+j}) = \mu_{i+j}$, so

$$A = \Big(\langle x^i, x^j\rangle\Big)_{0\le i,j\le n-1},$$

the Gram matrix of $1, x, \dots, x^{n-1}$ in that form. And the determinant of a Gram matrix is unchanged by any unitriangular change of basis, so it equals the determinant of the Gram matrix of *any* basis obtained from $1, x, \dots, x^{n-1}$ by Gram–Schmidt. Apply Gram–Schmidt: let $B_0 = 1$ and let B_i be the monic degree-i polynomial orthogonal to all lower degrees. Then the Gram matrix in the basis $B_0, \dots, B_{n-1}$ is diagonal, and

$$\det A = \prod_{i=0}^{n-1} L\big(B_i^2\big).$$

That is the entire theorem, and every step of it is a definition plus the fact that det is invariant under $M \mapsto UMU^{\mathsf{T}}$ for unitriangular U. *A Hankel determinant is a product of squared norms.* The triangular exponent $n(n-1)/2$ that the small cases produced is now explained before any computation: $L(B_i^2)$ will turn out to be $\lambda_1 \cdots \lambda_i$, so the product over $i < n$ contains λ_1 exactly $n - 1$ times, λ_2 exactly $n - 2$ times, and so on.

Why do continued fractions appear? Because orthogonal polynomials for *any* functional satisfy a three-term recurrence. The reason is one line: $\langle xB_i, B_j\rangle = L(xB_iB_j) = \langle B_i, xB_j\rangle$, and xB_j has degree $j + 1$, so this vanishes whenever $j + 1 < i$. Hence xB_i, expanded in the orthogonal basis, has no components below B_{i-1}:

$$B_{i+1} = (x - b_i)B_i - \lambda_i B_{i-1}.$$

And the denominators of the convergents of a continued fraction satisfy exactly this recurrence. So "orthogonal polynomials for L", "the three-term recurrence", and "the J-fraction of $\sum \mu_k x^k$" are three names for one object; the b_i and λ_i are the same numbers in all three. Kedlaya's solution opens by positing a continued fraction with unknown coefficients a_n, b_n and then "reinterprets in the language of orthogonal polynomials". Here is where that comes from: it is Gram–Schmidt on $1, x, x^2, \dots$, written in a different alphabet. The continued fraction is not a device imported from analysis; it is the only normal form a linear functional on polynomials has.

There is a second picture, worth carrying because it makes the exponent $n - k$ obvious. Expanding the J-fraction by first-return decomposition shows that μ_m counts weighted Motzkin paths of length m from height 0 to height 0 that never go below 0, with weight b_i for a level step at height i, λ_i for a down-step from height i, and 1 for an up-step: indeed if M_i is the generating function for such paths at level i, then a nonempty path either starts with a level step or rises, returns, and falls, giving $M_i = 1 + b_i x M_i + \lambda_{i+1} x^2 M_{i+1} M_i$, i.e. $M_i = 1/(1 - b_i x - \lambda_{i+1} x^2 M_{i+1})$. The Hankel determinant then counts families of n non-intersecting such paths, and in any such family one path must stay above height 1, another above height 2, and so on — which is where λ_k appears with multiplicity $n - k$. *The triangular exponent is a statement about how many paths are forced to be simultaneously high.*

Finally, why is the answer a clean power of a single number? Because f is quadratic over $\mathbb{Q}(x)$, and that is announced on the face of the problem by the square root. Over $\mathbb{Q}$, Lagrange's theorem says the continued fraction of a real number is eventually periodic if and only if the number is a quadratic irrational; the same proof runs over $\mathbb{Q}(x)$, because it only uses that the tails of the continued fraction of a quadratic satisfy quadratics with bounded coefficients, so finitely many occur. Here the period is 1, so every λ_k after the first coincides — and the first coincides too, as it happens. A periodic continued fraction means a constant λ, and a constant λ means $\det A = \lambda^{n(n-1)/2}$.

And the two constants can be read off without any expansion, which is the sentence to remember. Factor the quadratic under the root as

$$1 - 14x + 9x^2 = (1 - \alpha x)(1 - \beta x), \qquad \alpha, \beta = 7 \pm 2\sqrt{10},$$

so $\alpha + \beta = 14$, $\alpha\beta = 9$. Then the periodic part of the J-fraction has

$$b = \frac{\alpha + \beta}{2} = 7, \qquad \lambda = \left(\frac{\alpha - \beta}{4}\right)^2 = (\sqrt{10})^2 = 10,$$

because a J-fraction with constant b and λ is the one attached to the interval $[\,b - 2\sqrt{\lambda},\ b + 2\sqrt{\lambda}\,] = [\beta, \alpha]$ — the interval whose endpoints are the reciprocals of the two singularities of h. So the 10 in $10^{n(n-1)/2}$ is a quarter of the gap between the roots of $1 - 14x + 9x^2$, squared. The only thing the first layer contributes is $b_0 = 5$, and the b's do not appear in the determinant at all. In the general form (the official remark's $c_n = ac_{n-1} + b\sum c_i c_{n-i}$, here $a = 3$, $b = 2$), this reads $\lambda = ab + b^2$ and $b_0 = a + b$, $b_i = a + 2b$; with $a = 3, b = 2$ that is $\lambda = 10$, $b_0 = 5$, $b_i = 7$, exactly as computed. The Catalan numbers are $a = 0, b = 1$, giving $\lambda = 1$ and the familiar Hankel determinant 1.

Layer $+\alpha$ BACKTESTING

1. Two cues fire at once. (i) "i, j-entry c_{i+j-1}" — write down: *Hankel matrix*, therefore Gram matrix of the monomials for the functional $L(x^k) = c_{k+1}$, therefore the determinant is a product of squared norms of orthogonal polynomials. (ii) "$\sqrt{1 - 14x + 9x^2}$" — write down: the generating function is quadratic over $\mathbb{Q}(x)$, so rationalise to get a quadratic functional equation, and expect an eventually periodic continued fraction. Then, before theorising further, compute $\det A$ for $n = 1, 2, 3$.

2. The cues suggest the J-fraction machine: expand h as $1/(1 - b_0 x - \lambda_1 x^2/(1 - b_1 x - \cdots))$ and use $\det = \prod_k \lambda_k^{n-k}$. The quadratic equation supplies the λ's in one layer-stripping step because the tail is a fixed point.

3. Hesitation occurs at two places. First, immediately after obtaining $c_n = 3c_{n-1} + 2\sum c_i c_{n-i}$: you have a recursion and a matrix and you want to combine them. The reinforcement is the observation that the matrix entry depends only on $i + j$, so there is no per-row information to exploit and any progress must come from the series. Second, at the layer-stripping, where you fear an unbounded regress of new series $h_1, h_2, \ldots$; the reinforcement is Lagrange — a quadratic has a periodic continued fraction, so test for a fixed point after *one* step, and it will be there.

4. Three traps. (i) $c_0 = 0$. The series begins at x^1, so the matrix is built from $c_1, \ldots, c_{2n-1}$, and an off-by-one in the indexing produces a different (and much less pleasant) determinant. (ii) Guessing the pattern from $n = 1, 2$ alone: $1, 10$ is consistent with 10^{n-1}, and only $n = 3$ distinguishes 10^{n-1} from $10^{n(n-1)/2}$. Always compute one case past the first apparent pattern. (iii) Believing that the quadratic recursion for c_n is the useful form of the functional equation. It is the useful form for *computing* c_n and the useless form for the determinant; for the determinant you want $1 - 1/h$, not h itself.

5. Recognisable faster next time via two slogans. *An entry depending only on $i + j$ means a Gram matrix, and a Gram determinant is a product of squared norms.* And: *a square root of a quadratic is*

an announcement of periodicity — over $\mathbb{Q}$ it gives periodic continued fractions of $\sqrt{d}$, over $\mathbb{Q}(x)$ it gives constant recurrence coefficients, and in both cases it means the structure constants you are hunting are all equal.

6. The reusable decision rule: **when a matrix's entries depend on the indices only through $i+j$ (or only through $i-j$), stop performing row operations — there is no per-row information — and convert the matrix into the single object it encodes: a linear functional, a convolution operator, a generating function. Then find that object's normal form.** For $i+j$ the normal form is the sequence of orthogonal polynomials, equivalently the continued fraction, and $\det = \prod \lambda_k^{n-k}$. For $i-j$ (Toeplitz) the analogous normal form is the Wiener–Hopf factorisation and the analogous answer is Szegő's theorem. The general habit: *a determinant whose answer is a product with descending multiplicities is telling you there is a sequence of level constants; find what is indexed by the level.*

▷ **Cue:** a determinant is asked for and the (i,j) entry depends on i and j only through $i+j$ $\longrightarrow$ **Move:** name it Hankel; form the linear functional $L(x^k) = \mu_k$, recognise the matrix as the Gram matrix of $1, x, \ldots, x^{n-1}$ for $\langle P, Q\rangle = L(PQ)$, and compute $\det = \prod_{i<n} L(B_i^2)$ by Gram–Schmidt

▷ **Cue:** a generating function contains $\sqrt{\text{quadratic in } x}$ $\longrightarrow$ **Move:** rationalise to a quadratic functional equation, and expect the continued fraction to be eventually periodic (Lagrange over $\mathbb{Q}(x)$), hence expect all the structure constants to be equal

▷ **Cue:** the small cases of a determinant are $1, \lambda, \lambda^3, \lambda^6, \ldots$ $\longrightarrow$ **Move:** the exponent $n(n-1)/2$ is $(n-1)+(n-2)+\cdots+1$, a product with descending multiplicities; look for level constants λ_k contributing λ_k^{n-k}, not for a single closed formula

▷ **Cue:** you have written an object as $1/(1-bx-\lambda x^2 \cdot (\text{new object}))$ $\longrightarrow$ **Move:** test at once whether the new object satisfies the same equation; a fixed point ends the computation in one step instead of an infinite regress

▷ **Cue:** a Hankel generating function has $\sqrt{(1-\alpha x)(1-\beta x)}$ in it $\longrightarrow$ **Move:** read the periodic J-fraction constants straight off the roots: $b = (\alpha+\beta)/2$ and $\lambda = ((\alpha-\beta)/4)^2$, so the Hankel determinant is $\lambda^{n(n-1)/2}$

▷ **Cue:** a functional equation of the form $f = x + axf + bf^2$ (a Catalan/Schröder-type recursion) $\longrightarrow$ **Move:** its Hankel determinants are $(ab+b^2)^{n(n-1)/2}$; the special cases $a=0, b=1$ (Catalan) and $a=1, b=1$ (Schröder) are worth memorising as anchors

Putnam 2024 — Paper B

2024 B1

2024 B1. Let n and k be positive integers. The square in the ith row and jth column of an n-by-n grid contains the number $i + j - k$. For which n and k is it possible to select n squares from the grid, no two in the same row or column, such that the numbers contained in the selected squares are exactly $1, 2, \ldots, n$?

Fields: combinatorics (invariants), light algebra **Position:** B1 – opener of the second half, deliberately gentle: a two-line necessary condition plus one explicit construction **Answer:** Possible exactly when n is odd and $k = (n+1)/2$.

Layer 1 THE WHAT — surface analysis

The natural attempt. "No two in the same row or column" is a permutation in disguise, but the natural first move is positional, not algebraic: find where the value 1 can even go, since it is the smallest entry available. The entry $i + j - k$ is smallest when i, j are both as small as possible, so a student starts by pinning down which square can hold 1, then asks what square can hold 2 without clashing rows or columns with the first choice, and so on, effectively trying to build the selection square by square, checking clashes as they go, perhaps for $n = 3, 4, 5$ by hand to find the pattern. This can succeed – for $n = 5, k = 3$ the diagonal-ish selection $(1,3), (2,4), (3,5), (4,1), (5,2)$ does give $1, 3, 5, 2, 4$ – but it takes several small cases before the pattern "n odd, $k = (n+1)/2$" is even guessed, and having guessed it, the student still has no proof that other (n, k) are impossible; they would have to argue, case by case or by some induction on n, that no clever alternative placement can rescue an even n or a mismatched k. That is real work with no guarantee of termination in exam time: it is zombie maths, checking positions instead of asking what is forced no matter which positions are chosen.

The model chain. "No two in the same row or column" among n selected squares means precisely that the selection is described by a permutation σ of $\{1, \ldots, n\}$: the selected squares are $(i, \sigma(i))$ for $i = 1, \ldots, n$, and the corresponding values are $i + \sigma(i) - k$. Sum this value over all $i = 1, \ldots, n$:

$$\sum_{i=1}^{n} \big(i + \sigma(i) - k\big) = \sum_{i=1}^{n} i + \sum_{i=1}^{n} \sigma(i) - nk.$$

Because σ is a permutation of $\{1, \ldots, n\}$, both sums on the right equal $\frac{n(n+1)}{2}$ regardless of which permutation σ is: this identity holds before a single square has actually been chosen. So the total of the selected values is always $n(n+1) - nk$, no matter which valid selection is made. If the selected values are to be exactly $1, 2, \ldots, n$, their total must also equal $\frac{n(n+1)}{2}$. Equating,

$$n(n+1) - nk = \frac{n(n+1)}{2} \implies k = \frac{n+1}{2},$$

which forces n to be odd, since k is a positive integer. This is necessity, obtained with no construction at all.

For sufficiency, exhibit one permutation for n odd and $k = \frac{n+1}{2} =: m$. Take

$$(1,m),(2,m+1),\ldots,(m,n),\ (m+1,1),(m+2,2),\ldots,(n,m-1),$$

i.e. $\sigma(i) = i+m-1$ if $i \le m$, and $\sigma(i) = i+m-1-n$ if $i > m$ – a cyclic shift of the column index by $m-1$. For $i \le m$ the value is $i+\sigma(i)-m = i+(i+m-1)-m = 2i-1$, giving the odd numbers $1,3,\ldots,n$ as i runs from 1 to m. For $i > m$ the value is $i+\sigma(i)-m = i+(i+m-1-n)-m = 2i-1-n$; as i runs from $m+1$ to n this gives $2,4,\ldots,n-1$, the even numbers. Together, exactly $\{1,2,\ldots,n\}$ is hit, once each. (Verified directly for $n=5, m=3$: values $1,3,5,2,4$ – matches.)

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains agree that the selection is a permutation. They split immediately after that observation. The natural attempt asks a *local* question: "which specific square can hold the value 1, and which square can then hold 2 without a clash?" – a question about individual positions that has to be answered by trial. The model asks a *global* question instead: "what is the total of the values over the whole selection, and does that total depend on which permutation was chosen at all?" – a question answerable in one line, before any square has been picked.

> *I don't need to find, or even think about, which squares are selected in order to extract information about k – I only need one number, the sum of the selected values, and that number is already pinned down the instant I know the selection is a permutation, because the row indices and the column indices are each some rearrangement of $1,\ldots,n$ and rearranging a fixed set of numbers never changes what they add up to.*

Layer 3 THE WHY

The value in square (i,j) is $i+j-k$: linear in i and linear in j, separately. Any selection with no two squares sharing a row or column is, by definition, a bijection between row indices and column indices – a permutation σ. Summing a linear functional of i and $\sigma(i)$ over $i = 1,\ldots,n$ therefore always splits into $\sum i$ plus $\sum\sigma(i)$ minus a constant term repeated n times, and the two sums $\sum i$, $\sum\sigma(i)$ are each just $1+2+\cdots+n$ reordered – and reordering a finite sum of fixed numbers cannot change its value. That is the entire content of "no two in the same row or column": it guarantees the column indices form some permutation, and a permutation's values sum to a fixed total no matter its order. Once this is seen, the necessary condition on k falls out with zero casework, because the problem's own phrasing ("$i+j-k$", "no two in the same row or column") already hands the solver both a linear functional and a permutation structure – exactly the two ingredients a summing argument needs. This is the general lesson behind invariant arguments: whenever a constraint forces an unknown arrangement to be a bijection between two fixed index sets, any functional that is linear (or otherwise symmetric) in the two coordinates can be evaluated in total *before* the arrangement is known, and that evaluation is a free necessary condition. Sufficiency, by contrast, genuinely requires a construction – an invariant only ever rules things out, it never on its own proves something is achievable, which is why the model chain still needs the explicit cyclic-shift permutation to close the argument.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the phrase "no two in the same row or column" applied to a full selection of n squares in an n-by-n grid: on seeing it, write down immediately that this means the selection is a permutation σ of $\{1,\ldots,n\}$, before thinking about which squares specifically.
2. That cue suggests summing the (linear) entry formula over the whole selection, since summing over a permutation of a fixed set is invariant to the choice of permutation – this converts a positional search into a one-line algebraic identity.

3. Hesitation occurs right after guessing the pattern from small cases: there is a temptation to try to prove necessity by another round of case-by-case placement arguments for general n. The reinforcement is to notice the entry formula is linear in each index and ask "what happens if I just add up all n selected values, symbolically, instead of tracking positions?"
4. The trap is treating this as a search-and-construct problem throughout, which finds the answer by experimentation but leaves necessity unproved except by an open-ended case analysis that does not obviously terminate.
5. Recognisable faster next time by naming the mechanism: "selection with no repeated row or column = permutation; entry formula linear in the two indices = summable invariant." Any grid problem phrased with those two ingredients should trigger an immediate summation before any positional trial.
6. Reusable rule: when a problem forces an unknown selection to be a bijection between two fixed index sets and asks about a linear (or otherwise symmetric) functional of the two indices, sum that functional over the whole selection first – the total is invariant under the choice of bijection and gives a necessary condition for free; only then hunt for an explicit construction to establish sufficiency.

▷ **Cue:** selection forced to have "no two in the same row/column" over a full grid $\longrightarrow$ **Move:** recognise this as a permutation and sum any coordinate-linear quantity over the whole selection to get a free, arrangement-independent necessary condition

▷ **Cue:** a necessary condition obtained by summation, with a specific surviving case $\longrightarrow$ **Move:** always separately verify sufficiency by an explicit construction – an invariant only rules out, it never proves achievability

2024 B2

2024 B2. Two convex quadrilaterals are called *partners* if they have three vertices in common and they can be labeled $ABCD$ and $ABCE$ so that E is the reflection of D across the perpendicular bisector of the diagonal $\overline{AC}$. Is there an infinite sequence of convex quadrilaterals such that each quadrilateral is a partner of its successor and no two elements of the sequence are congruent?

Fields: Euclidean geometry (isometry invariants), pigeonhole **Position:** B2 – second problem of the second half, geometric, with the whole difficulty in identifying which quantities survive the move **Answer:** No: reflecting D to E across the perpendicular bisector of AC preserves both the quadrilateral's area and its full multiset of side lengths exactly, and only finitely many congruence classes of convex quadrilateral share a given area and side-length multiset, so an infinite pairwise-non-congruent chain of partners cannot exist.

Layer 1 THE WHAT — surface analysis

The natural attempt. Faced with a "does an infinite non-repeating sequence exist" question, the instinct is to try to build one: start from a convenient quadrilateral – a square, say – and generate partner after partner, hoping either to see an obvious way to keep growing it forever or to hit a wall empirically. After a few steps the shapes plainly keep changing, which tempts the guess "yes, infinitely many, since nothing local stops it." Suspecting instead that the answer is "no" (natural for a Putnam B2, where "yes, obviously" is rarely the answer), the next instinct is to hunt for a conserved quantity by testing the easiest candidates first. Perimeter is the obvious one to check: the new quadrilateral $ABCE$ has sides AB, BC, CE, EA, and a short reflection argument (see below) shows $CE = AD$ and $EA = DC$, so the perimeter $AB + BC + CE + EA = AB + BC + AD + DC$ is exactly the same as before. A student who stops here and declares "perimeter is invariant, so only finitely many quadrilaterals can occur" has made an unjustified leap: an enormous, genuinely infinite family of non-congruent convex

quadrilaterals can share one fixed perimeter (think of a family that gets thinner and longer while perimeter stays put). Fixing one scalar total does not fix a four-parameter shape; the natural attempt has found a true invariant that is nonetheless too weak to finish the problem, and does not yet know it.

The model chain. Split the quadrilateral along the shared diagonal AC: because $ABCD$ is convex, $[ABCD] = [ABC] + [ACD]$, where $[\cdot]$ denotes area, with the two triangles meeting exactly along AC and no overlap. Reflecting across the perpendicular bisector of AC is an isometry that swaps A and C and sends $D \mapsto E$; consequently it carries triangle ACD *wholly* onto triangle CAE, so $[ACD] = [ACE]$ exactly (congruent triangles, not merely similar). Hence

$$[ABCE] = [ABC] + [ACE] = [ABC] + [ACD] = [ABCD] :$$

area is preserved exactly at every step. The same reflection, being an isometry, also preserves the lengths AD and DC individually: it sends segment AD to segment CE (since $A \mapsto C$, $D \mapsto E$) and segment DC to segment EA (since $D \mapsto E$, $C \mapsto A$), so $CE = AD$ and $EA = DC$ exactly. The new side lengths AB, BC, CE, EA are therefore literally the same four numbers as AB, BC, CD, DA, just with the last two attached to the opposite vertex. So both the area and the exact multiset $\{AB, BC, CD, DA\}$ of side lengths are preserved forever along the sequence, regardless of which three vertices persist at each step (the argument only used that *some* diagonal is shared and reflected, which is exactly what "partner" guarantees).

Now invoke finiteness: for four fixed side lengths in a *fixed cyclic order* and a fixed area, only finitely many congruence classes of convex quadrilateral exist. (Sketch: the congruence class of a quadrilateral with sides a, b, c, d in cyclic order is determined by the length f of one diagonal, via SSS on the two flanking triangles; as f varies over its valid range, the two angles at the diagonal's endpoints vary monotonically, so by Bretschneider's formula, which expresses the area in terms of a, b, c, d and the cosine of the half-sum of those two angles, fixing the area pins that angle-sum to at most two values, hence f to at most two values.) Since the 4 fixed lengths can occur in at most $4!/4 = 6$ distinct cyclic orders, the whole sequence draws from at most $6 \times 2 = 12$ congruence classes in total. A sequence of pairwise non-congruent quadrilaterals can therefore have at most 12 terms: no infinite such sequence exists.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains reach "perimeter is preserved." The natural attempt stops there and misjudges what it has proved. The model chain asks one more question at exactly that point: "is a fixed SUM of four lengths the same thing as a fixed SHAPE?" – and answers no, because a quadrilateral has more degrees of freedom (effectively, the diagonal length and the individual side lengths) than a single scalar total can pin down. That question is the split: the natural attempt treats "found an invariant" as "found enough invariants," the model treats it as a prompt to go looking for a second, independent one.

> *One invariant is not automatically enough – perimeter being fixed only fixes the SUM of the four side lengths, and a fixed sum leaves an entire continuum of possible shapes, so if I want to actually bound the number of possible congruence classes I need invariants that pin down the individual numbers (the full multiset of side lengths, not their total) together with a second, independent scalar (area) that is not just some linear combination of the first.*

Layer 3 THE WHY

Convexity is what licenses "$[ABCD] = [ABC] + [ACD]$" without correction terms: the diagonal AC of a convex quadrilateral lies entirely inside it, so the two triangles it cuts off tile the quadrilateral exactly, with no overlap and no gap – that is precisely what "convex" buys here, and it is why the problem insists on convex quadrilaterals throughout. An isometry, by definition, preserves every length and hence every area of any figure it carries in full; since the

perpendicular-bisector reflection carries triangle ACD onto triangle CAE as a whole (not just vertex D in isolation), it is illegitimate to only track what happens to D – the entire triangle, area included, goes along for the ride, which is the source of the area invariant. The side lengths transfer as a *set* rather than staying in their original cyclic slots because the reflection swaps $A \leftrightarrow C$: the segment that used to run from A to D now runs from C to E, so the numerical value AD reappears as CE, attached to a different vertex but numerically identical – this is the definition of reflection (it maps segments to congruent segments) doing all the work, not a coincidence to be checked case by case. Finally, the reason two invariants (area and the side-length multiset) are enough to force finiteness, where one (perimeter) was not, is a dimension-counting fact: a convex quadrilateral's congruence class has four continuous degrees of freedom (say, three side lengths and one diagonal, with the fourth side length determined), and fixing the side-length multiset removes essentially all of that freedom except a discrete choice of cyclic order together with the diagonal length, while fixing the area on top of that pins the diagonal to a finite set via the angle-sum argument above.

Caveat. I have verified the area and side-length invariances myself, directly, from the definitions of convexity and isometry, with no gaps. The finiteness sub-lemma – that fixed side lengths (in a fixed cyclic order) and fixed area force the diagonal, and hence the congruence class, to at most two possibilities – rests on a monotonicity claim (the two angles flanking the diagonal vary monotonically as the diagonal length varies) that I have not independently re-derived here; I am relying on the structure of the official Bretschneider-formula argument for that specific step rather than a derivation I checked line by line myself.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is "reflection across the perpendicular bisector of a diagonal, with three vertices shared": on seeing "perpendicular bisector," write down immediately what it does as an isometry – it swaps the two endpoints of the diagonal and preserves every length and area of anything it carries as a whole.

2. That cue suggests looking for a sub-figure wholly carried by the isometry (here, triangle $ACD \to ACE$) and logging every length and area of that sub-figure as an exact invariant of the move, rather than only tracking the one vertex ($D \to E$) that visibly changed.

3. Hesitation occurs right after finding that perimeter is preserved and being tempted to declare the problem solved. The reinforcement is to ask explicitly: does this one preserved scalar actually bound the SHAPE, or only one linear combination of the shape's parameters? Here it plainly does not, since a fixed perimeter alone admits infinitely many non-congruent quadrilaterals.

4. The trap is mistaking a preserved sum (perimeter) for a preserved, shape-determining quantity; a family of shapes can live inside a fixed-perimeter constraint with no upper bound on how many pairwise non-congruent members it contains, so an invariant of that weak a type cannot, by itself, force a sequence to be finite.

5. Recognisable faster next time by naming the mechanism: "a construction defined by reflecting (or rotating) part of a figure while the rest stays fixed" should immediately trigger the question "what is the LARGEST sub-figure that isometry carries wholly?", and every length and area of that sub-figure – not just its visible boundary vertex – should be logged as an exact invariant.

6. Reusable rule: to prove an infinite sequence of pairwise non-congruent objects cannot exist under a repeated generating move, find enough independently-varying invariants preserved by that move (not merely one linear total) to reduce the possible objects to finitely many congruence classes, then invoke pigeonhole; a single scalar invariant is rarely enough on its own, because it usually fixes only one combination of several genuine degrees of

freedom.

- ▷ **Cue:** construction defined by reflecting (or rotating) part of a figure while other vertices stay fixed $\longrightarrow$ **Move:** identify the maximal sub-figure wholly carried by the isometry and log its lengths and area as exact invariants, since an isometry preserves both exactly for anything it carries in full
- ▷ **Cue:** one preserved scalar quantity found (e.g. perimeter) while trying to show an infinite non-repeating sequence is impossible $\longrightarrow$ **Move:** check whether that scalar actually bounds the shape's degrees of freedom, or only one linear combination of them; if not, hunt for a second, independent invariant before concluding finiteness

2024 B3

2024 B3. Let r_n be the nth smallest positive solution to $\tan x = x$, where the argument of tangent is in radians. Prove that

$$0 < r_{n+1} - r_n - \pi < \frac{1}{(n^2+n)\pi}$$

for $n \geq 1$.

Fields: real analysis (monotonicity of inverse functions) **Position:** B3 – third problem of the second half, calculus-heavy but mechanically short once the right substitution is found **Answer:** Setting $t(x) = x - \tan^{-1} x$ (a strictly increasing bijection $\mathbb{R} \to \mathbb{R}$) and writing $r_n = x(n\pi)$ for its inverse function $x(\cdot)$, both bounds fall straight out of the sign of the derivative of $x(t) - t$ and of $x(t) - t + 1/t$.

Layer 1 THE WHAT — surface analysis

The natural attempt. $\tan x = x$ is the given equation, so the natural instinct is to work with $f(x) := \tan x - x$ directly: differentiate it, $f'(x) = \tan^2 x$, note this is nonnegative, and try to squeeze $r_{n+1} - r_n$ from the growth of f between consecutive roots. This is workable – it is essentially the official solution's second route – but it is long. First one has to locate r_n precisely inside the interval $(n\pi, n\pi + \pi/2)$ using the intermediate value theorem, since $\tan x$ only makes sense to compare to x on each branch. Then, to get the lower bound, one shows $f(r_n + \pi) = \tan(r_n) - r_n - \pi = -\pi < 0$ and argues from monotonicity of f on the next branch that $r_{n+1} > r_n + \pi$. For the sharp upper bound one has to go further: differentiate f repeatedly, show by induction that every derivative $f^{(k)}$ is a nonnegative-coefficient polynomial in $\tan x$ (hence positive and increasing on each branch), assemble a Taylor expansion with remainder of f at $r_n + \pi$, and take a limit $k \to \infty$ of a partial sum while being careful that a non-strict inequality does not silently appear in the limiting step. Every piece of this is correct, but there are several places – fixing the interval containing r_n, tracking the sign of each derivative, handling the $k \to \infty$ limit correctly – where an exam-time solver can lose the thread or accidentally weaken a strict inequality to a non-strict one.

The model chain. Rewrite the defining equation before differentiating anything. If $\tan x = x$, then applying the principal-branch arctangent to both sides is not directly legal (tan is not injective), but it says precisely that x and $\tan^{-1}(x)$ differ by an integer multiple of π: $x = \tan^{-1}(x) + n\pi$ for the appropriate integer n. So define

$$t(x) := x - \tan^{-1}(x),$$

and the roots of $\tan x = x$ are exactly the solutions of $t(x) = n\pi$. Now $t(0) = 0$ and $t'(x) = 1 - \frac{1}{1+x^2} = \frac{x^2}{1+x^2} \geq 0$, with equality only at $x = 0$; so t is strictly increasing on $\mathbb{R}$, continuous, and $t(x) \to \pm\infty$ as $x \to \pm\infty$ – a genuine bijection $\mathbb{R} \to \mathbb{R}$. It therefore has a well-defined, differentiable inverse function $x(t)$, and by definition $r_n = x(n\pi)$ for $n \geq 1$: not "the nth root of

a transcendental equation" but simply "the value of a known increasing function at the known input $n\pi$."

By the inverse function rule, $\dfrac{dx}{dt} = \dfrac{1}{t'(x)} = \dfrac{1+x(t)^2}{x(t)^2} = 1 + \dfrac{1}{x(t)^2}$.

Lower bound. Let $\varphi(t) := x(t) - t$. Then $\varphi'(t) = \dfrac{1}{x(t)^2} > 0$ for all $t \neq 0$ (in particular for $t > 0$, where $x(t) > 0$ since t is increasing and $t(0) = 0$). So φ is strictly increasing on $t > 0$, and comparing $t = n\pi$ to $t = (n+1)\pi$,

$$\varphi(n\pi) < \varphi((n+1)\pi) \iff r_n - n\pi < r_{n+1} - (n+1)\pi \iff r_{n+1} - r_n - \pi > 0.$$

Upper bound. For $t > 0$, $x(t) > 0$, and $\tan^{-1}(x(t)) > 0$, so $t = x(t) - \tan^{-1}(x(t)) < x(t)$: that is, $x(t) > t > 0$. Let $\psi(t) := x(t) - t + \dfrac{1}{t}$. Then $\psi'(t) = \dfrac{1}{x(t)^2} - \dfrac{1}{t^2}$, which is negative precisely because $x(t) > t > 0 \Rightarrow x(t)^2 > t^2 \Rightarrow \dfrac{1}{x(t)^2} < \dfrac{1}{t^2}$. So ψ is strictly decreasing on $t > 0$, and

$$\psi(n\pi) > \psi((n+1)\pi) \iff r_n - n\pi + \frac{1}{n\pi} > r_{n+1} - (n+1)\pi + \frac{1}{(n+1)\pi}.$$

Rearranging,

$$r_{n+1} - r_n - \pi < \frac{1}{n\pi} - \frac{1}{(n+1)\pi} = \frac{1}{\pi} \cdot \frac{1}{n(n+1)} = \frac{1}{(n^2+n)\pi},$$

which is exactly the stated upper bound. Both inequalities together give the full claim, with no casework, no induction on derivatives, and no limiting argument.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains eventually need a monotonicity statement to compare r_n and r_{n+1}. They split at the very first move, before any differentiation happens. The natural attempt keeps x as the free variable and studies $f(x) = \tan x - x$ directly, so every subsequent step has to fight the fact that $\tan x$ is unbounded and only piecewise monotone, and r_n itself has to be located inside an interval before anything else can be said about it. The model chain instead swaps which variable is free: it sets $t(x) = x - \tan^{-1} x$ – a function already known, globally, to be smooth and strictly increasing – and treats x as a function *of* t, so that $r_n = x(n\pi)$ is not a root to hunt for at all, but one evaluation of an already-inverted, well-behaved function.

> *I keep trying to control* $\tan x$*, and* $\tan x$ *is the wrong object to control – it blows up and is only piecewise monotone. The actual content of the equation is that* x *and its own arctangent differ by a multiple of* π*, so the quantity that behaves nicely is* $t(x) = x - \tan^{-1} x$*, not* x *itself; if I invert that function,* r_n *stops being "a root I have to locate" and becomes "the value of a known monotone function at the known input* $n\pi$*," and every inequality I want collapses to signing a derivative.*

Layer 3 THE WHY

The definition of the principal-branch arctangent is exactly what licenses the rewrite $x = \tan^{-1}(x) + n\pi$: $\tan^{-1}$ is defined to be the inverse of tan restricted to one period, $(-\pi/2, \pi/2)$, so "$\tan x = x$" and "x minus its arctangent is an integer multiple of π" are the same statement, read through the periodicity of tangent – this is not a trick substitution, it is unpacking what the equation already says. That the derivative $t'(x) = x^2/(1+x^2)$ is a square divided by a positive quantity, and hence never negative, is what licenses inverting t not just locally near one root but globally over all of $\mathbb{R}$ at once (a strictly monotone continuous surjection between two copies of $\mathbb{R}$ has a genuine global inverse, by the definition of monotone bijection) – which is precisely why "nth smallest positive root, in increasing order" becomes, automatically

and without extra argument, "value of the inverse function at the nth multiple of π": the ordering built into "nth root" is inherited for free from t being increasing. Both derivative computations that finish the problem, $\varphi'(t) = 1/x(t)^2$ and $\psi'(t) = 1/x(t)^2 - 1/t^2$, reduce to the single fact $dx/dt = 1 + 1/x(t)^2$, itself nothing more than the inverse function rule applied to $t'(x) = x^2/(1+x^2)$; and the sign of ψ' is settled purely by $x(t) > t$ for $t > 0$, which is itself just $\tan^{-1}(x) > 0$ for $x > 0$ – a direct, unavoidable consequence of the definition of arctangent on the positive reals, not a further insight that needs discovering. Once the substitution is made, nothing in the rest of the argument requires cleverness; it requires only correctly identifying which known, elementary fact each sign depends on.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the equation $\tan x = x$ itself, read through the periodicity of tan: on seeing it, write down $x = \tan^{-1}(x) + n\pi$ rather than differentiating $\tan x - x$ as it stands.

2. That cue suggests defining $t(x) = x - \tan^{-1} x$ and studying its (global, strict) monotonicity, which converts "the nth root" into "the value of an inverse function at $n\pi$."

3. Hesitation occurs at the moment of deciding whether to invert $t(x)$ or to keep pushing forward with $\tan x$ directly, as the natural attempt does. The reinforcement is to compare the two candidate derivatives: $t'(x) = x^2/(1+x^2)$ is a clean, globally sign-definite closed form, while $\tan^2 x$ is unbounded and only piecewise defined – a clean, globally-signed derivative is the signal to invert rather than to push forward with the original equation.

4. The trap is trying to extract both inequalities by bounding $\tan x$ and its higher derivatives near r_n directly; this is provably possible (the official second solution does it) but requires locating r_n's interval, signing every derivative by induction, and handling a $k \to \infty$ limiting step where a strict inequality can silently degrade – all avoidable with the substitution.

5. Recognisable faster next time by naming the mechanism: an equation of the form (periodic function)$(x) = x$ is often better read as "x minus the corresponding inverse trig function equals a multiple of the period" – whenever an equation pins x into an interval indexed by n via a periodic function, look for that inverse-function substitution before differentiating the original equation.

6. Reusable rule: when comparing consecutive terms of a sequence defined implicitly as roots of $f(x) = g(x)$ with g periodic, substitute $t = x -$ (the matching inverse branch of g) to turn "the nth root" into "the value of a genuinely monotone, differentiable inverse function at the nth grid point of the period"; every comparison inequality then reduces to signing a derivative of that inverse function, not to estimating the original equation directly.

▷ **Cue:** equation of the form (periodic function)(x) = x, e.g. $\tan x = x$ $\longrightarrow$ **Move:** substitute $t(x) = x -$ (inverse of the periodic function) and work with its typically clean, sign-definite derivative instead of the original equation

▷ **Cue:** need to compare consecutive values $x(n\pi), x((n+1)\pi)$ of an implicitly-defined increasing sequence $\longrightarrow$ **Move:** examine the sign of the derivative of $x(t)$ minus (or plus) a comparison function such as t or $1/t$, rather than estimating $x(t)$ directly

2024 B4

2024 B4. Let n be a positive integer. Set $a_{n,0} = 1$. For $k \geq 0$, choose an integer $m_{n,k}$ uniformly at random from the set $\{1, \ldots, n\}$, and let

$$a_{n,k+1} = \begin{cases} a_{n,k} + 1, & \text{if } m_{n,k} > a_{n,k}; \\ a_{n,k}, & \text{if } m_{n,k} = a_{n,k}; \\ a_{n,k} - 1, & \text{if } m_{n,k} < a_{n,k}. \end{cases}$$

Let $E(n)$ be the expected value of $a_{n,n}$. Determine $\lim_{n\to\infty} E(n)/n$.

Fields: probability (finite Markov chain); linear difference equations **Position:** B4 — positioned as the step up in difficulty on the B paper, and the problem on this paper with the widest gap between the expensive correct route and the cheap correct route **Answer:** $\dfrac{1-e^{-2}}{2}$, because the conditional expected increment is an affine function of the state, so the expectation obeys $e_{k+1} = (1-\frac{2}{n})e_k + \frac{n+1}{n}$ on its own.

Layer 1 THE WHAT — surface analysis

The natural attempt. The problem hands you a random walk on $\{1,\dots,n\}$ with state-dependent transition probabilities. The default response is to write down the chain in full. Let v_k be the column vector whose i-th entry is $\Pr[a_{n,k} = i]$. Counting the m that land in each of the three cases — there are $n-i$ values with $m > i$, exactly one with $m = i$, and $i-1$ with $m < i$ — gives $v_{k+1} = Av_k$ with

$$A_{ij} = \begin{cases} \frac{1}{n} & i = j, \\ \frac{j-1}{n} & i = j-1, \\ \frac{n-j}{n} & i = j+1, \\ 0 & \text{otherwise,} \end{cases}$$

and $v_0 = (1,0,\dots,0)^t$. With $w = (1,2,\dots,n)$ the answer is the single entry of wA^nv_0.

Now look at what you have asked yourself for. You want the n-th power of an $n \times n$ tridiagonal stochastic matrix, applied to a corner vector, and you want it to an accuracy that survives division by n. The honest route is diagonalisation. A does have clean spectrum — its eigenvalues are $\frac{n-2j}{n}$ for $j = 0,\dots,n-1$ — but finding that, and then expressing the corner vector v_0 in the basis of right eigenvectors, is a substantial piece of work, and the right eigenvectors of a non-symmetric tridiagonal matrix are not pretty. That is the wall: not a contradiction, just a computation you will not finish in the time available.

Two cheaper-looking detours also present themselves and both cost real time. You can simulate: $n = 2$ gives $E = 3/2$, $n = 3$ gives $53/27$, $n = 4$ gives $77/32$. Divide by n and you get $0.75, 0.654, 0.602$ — drifting slowly downward towards something, and $\frac{1-e^{-2}}{2} = 0.4323$ is nowhere in sight at $n = 4$. Numerics do not identify this limit. Or you can find the combinatorial recoding: $a_{n,k}$ is the number of lit bulbs in an apparatus of one permanently lit red bulb and $n-1$ white bulbs, where each step picks a bulb uniformly and toggles it unless it is the red one. That recoding is correct and it does finish the problem, but seeing it requires you to already know the answer has the shape $\frac{1}{2}(1-(1-\frac{2}{n})^n)$.

The model chain. Do not track the distribution. Track only the number you were asked for.

Condition on the state. If $a_{n,k} = i$, then of the n equally likely values of $m_{n,k}$ there are $n-i$ giving $i+1$, one giving i, and $i-1$ giving $i-1$. Hence

$$\mathbb{E}[a_{n,k+1} \mid a_{n,k} = i] = \frac{(n-i)(i+1) + i + (i-1)^2}{n}.$$

Expand the numerator: $(n-i)(i+1) = ni + n - i^2 - i$ and $(i-1)^2 = i^2 - 2i + 1$, so the i^2 terms cancel and the numerator is $ni + n - 2i + 1$. Therefore

$$\mathbb{E}[a_{n,k+1} \mid a_{n,k} = i] = \left(1 - \frac{2}{n}\right) i + \frac{n+1}{n}. \tag{†}$$

The right-hand side is affine in i. That is the whole problem. Writing $e_k := \mathbb{E}[a_{n,k}]$ and taking expectations of (†) — legitimate by the tower property, and requiring no independence and no knowledge of the distribution of $a_{n,k}$ — gives a closed one-dimensional recurrence:

$$e_{k+1} = \left(1 - \frac{2}{n}\right) e_k + \frac{n+1}{n}, \qquad e_0 = 1.$$

Its fixed point is $e^* = \frac{n+1}{2}$, so $e_k - \frac{n+1}{2}$ is geometric with ratio $1 - \frac{2}{n}$:

$$e_k = \frac{n+1}{2} - \frac{n-1}{2}\left(1 - \frac{2}{n}\right)^k.$$

Set $k = n$:

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

and dividing by n,

$$\frac{E(n)}{n} = \frac{n+1}{2n} - \frac{n-1}{2n}\left(1 - \frac{2}{n}\right)^n \longrightarrow \frac{1}{2} - \frac{1}{2}e^{-2} = \frac{1 - e^{-2}}{2}. \qquad \square$$

Two checks worth making before you write this up. First, $n = 2$: the formula gives $\frac{3}{2} - \frac{1}{2} \cdot 0 = \frac{3}{2}$, and by hand the chain on $\{1, 2\}$ from $a = 1$ reaches each of $1, 2$ with probability $\frac{1}{2}$ at both steps, so $E(2) = \frac{3}{2}$. Second, $n = 4$: the formula gives $\frac{5}{2} - \frac{3}{2} \cdot \frac{1}{16} = \frac{77}{32}$, which is what the direct enumeration of the chain gives.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree up to the line that counts cases: *there are $n - i$ values of m above i, one equal to i, and $i - 1$ below*. Call it $(*)$.

At $(*)$ the natural attempt assembles those three counts into a *row of a matrix* — it stores the information as a transition law and then asks for A^n. The model chain assembles the same three counts into a *single number*, the conditional expectation, and looks at it. The split is not about technique; it is about what object you decide the three counts belong to. One reading commits you to the distribution of $a_{n,n}$, which you were never asked for. The other reading commits you to $\mathbb{E}[a_{n,n}]$, which you were.

> *I have been asked for one number, an expectation, not for the distribution. Before I build a transition matrix I should take expectations of the one-step law and see whether it closes — if $\mathbb{E}[a_{k+1} \mid a_k]$ turns out to depend on a_k only through a_k itself and not through a_k^2, then the sequence of expectations satisfies its own recurrence and the distribution never has to be computed.*

> *Why should it close? Because the number of m exceeding i is $n - i$ — linear in i — and the increment in that case is $+1$; the number below is $i - 1$, linear in i, with increment -1. Probability linear in the state times an increment constant in the state gives a drift linear in the state. The quadratic terms are forced to cancel, and I can see they will before I expand.*

> *The contraction factor is $1 - \frac{2}{n}$, and 2 out of n is the size of a two-element set. Which two? Run two copies of the chain from i and $i + 1$ with the* same *m: they merge exactly when $m \in \{i, i + 1\}$ and otherwise stay one apart. So $(1 - \frac{2}{n})^n$ is the probability that two adjacent starting points have not yet merged after n steps, and e^{-2} in the answer is that probability in the limit. The formula is telling me what it means.*

- ▷ **Cue:** a Markov chain where you are asked for an expectation, not a distribution $\longrightarrow$ **Move:** compute $\mathbb{E}[X_{k+1} \mid X_k]$ first; if it is affine in X_k, the expectations satisfy a scalar linear recurrence and the chain is finished
- ▷ **Cue:** transition probabilities that count "how many of $1, \ldots, n$ lie above the current state" $\longrightarrow$ **Move:** the probabilities are linear in the state, so the drift is linear in the state — expect the expectation to close

▷ **Cue:** an answer containing e^{-c} from a problem with n steps $\longrightarrow$ **Move:** look for a per-step event of probability c/n; $(1-c/n)^n$ is the probability it never happens, and identifying that event is the combinatorial meaning of the answer

▷ **Cue:** a scalar affine recurrence $e_{k+1} = \lambda e_k + \mu$ $\longrightarrow$ **Move:** find the fixed point $e^* = \mu/(1-\lambda)$ and use $e_k - e^* = \lambda^k(e_0 - e^*)$; never iterate blindly

Layer 3 THE WHY

Why must the first of those thoughts be thought? Because of what an expectation is, and because of the one property of it that costs nothing.

$\mathbb{E}$ is linear, unconditionally. It does not require independence, it does not require the summands to be identically distributed, it does not care about the joint law. That freeness is the reason expectation is almost always cheaper than distribution, and it is why the instinctive order of operations — first find the distribution, then take its mean — is backwards. The distribution of $a_{n,n}$ is a genuinely complicated object: it is $A^n v_0$ for a tridiagonal A whose eigenvectors are unpleasant. The mean of that distribution is a single number satisfying a three-line recurrence. You are not being asked to compute a hard object and then throw most of it away. You are being asked for the number.

But linearity of expectation alone does not get you a closed recurrence. The extra ingredient is the structure of *this* chain, and this is the part that deserves to be driven to the axiom.

A Markov chain's expectation closes on itself precisely when the drift $\mathbb{E}[X_{k+1} \mid X_k = i]$ is an affine function of i. If the drift were quadratic in i, taking expectations would produce $\mathbb{E}[a_{n,k}^2]$ on the right-hand side, and you would need a recurrence for the second moment, which would drag in the third, and so on — the moment hierarchy would not truncate and you would be back to needing the whole distribution. So the question "does this problem have a cheap route?" is exactly the question "is the drift affine?", and that is a question you can answer by inspection, before any algebra.

Here it is affine, and the reason is the word *uniformly* in the problem statement. $m_{n,k}$ is uniform on $\{1,\dots,n\}$, so the probability that $m > i$ is the *count* $n-i$ divided by n. A count of integers in an interval is a linear function of the endpoint. The increments attached to the three cases are $+1, 0, -1$ — constants. Probability linear in the state, increment constant in the state, so drift linear in the state. Change "uniformly" to any distribution whose tail function $\Pr[m > i]$ is not affine in i and the trick dies instantly. The hypothesis that looks like a piece of routine set-dressing is the entire mechanism.

Kedlaya's first solution begins by constructing the matrix A and then produces two left eigenvectors, $w_0 = (1,\dots,1)$ with eigenvalue 1 and $w_1 = (n-2j+1)_j$ with eigenvalue $\frac{n-2}{n}$, observing that the target vector $w = (1,\dots,n)$ is the combination $\frac{n+1}{2}w_0 - \frac{1}{2}w_1$. Here is where that comes from. Those two eigenvectors are not discovered by solving an eigenvalue problem; they are the same computation as (†) written in matrix language. w_0 being a left 1-eigenvector is the statement that A is stochastic — total probability is conserved. w_1 being an eigenvector with eigenvalue $1-\frac{2}{n}$ is the statement that the drift is affine with slope $1-\frac{2}{n}$, once you recentre the state at the symmetry point. There is nothing to find. If you have written (†) you have already written both eigenvectors, and if you go looking for eigenvectors first you are searching a space in which the two you need happen to be the only two you can name.

The symmetry point deserves its own sentence, because it explains the fixed point $\frac{n+1}{2}$ without any algebra. The map $i \mapsto n+1-i$ on states, together with $m \mapsto n+1-m$ on the input, sends the case $m > i$ to the case $m < i$ and reverses the sign of every increment. So the chain is genuinely symmetric about $\frac{n+1}{2}$, and a symmetric chain cannot have a drift fixed point anywhere else. That is not a check on the arithmetic — it is a reason the arithmetic had to come out that way, and it is worth doing before the arithmetic, because it tells you the answer will be $\frac{1}{2}$ minus a correction rather than anything else.

Finally, the shape of the correction. $\frac{1}{2}e^{-2}$ is not an artefact. Couple two copies of the chain started at i and $i+1$ using the same m at every step. If $m \in \{i, i+1\}$ they coalesce — check the

four cases and the two states become equal — and if not, they remain exactly one apart. So each step coalesces with probability $\frac{2}{n}$, independently, and the expected value of $a_{n,n}$ increases by exactly $(1-\frac{2}{n})^n$ when you move the start point up by one. Chasing that from $a_{n,0} = 1$ to $a_{n,0} = n$ and using the symmetry above recovers the formula. The 2 in e^{-2} is the size of the set $\{i, i+1\}$. Once you know that, the answer is not a limit you computed; it is a probability you can name.

> **Caveat.** I verified the closed form $E(n) = \frac{n+1}{2} - \frac{n-1}{2}(1-\frac{2}{n})^n$ against exact rational enumeration of the chain for $n = 1, \ldots, 8$; it agrees in every case. The coalescence argument in the last paragraph I verified by checking the four cases $m \leq i-1$, $m = i$, $m = i+1$, $m \geq i+2$ by hand, but I have presented it as an interpretation of the answer rather than as the proof given above, which stands on (†) alone.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the phrase "chosen uniformly at random from $\{1, \ldots, n\}$" sitting next to a three-case rule split at the current state. On seeing it, write down the three counts $n-i$, 1, $i-1$ — and then write down, on the next line, $\mathbb{E}[a_{k+1} \mid a_k = i]$ rather than a matrix row. Also circle the word *expected* in "Let $E(n)$ be the expected value".

2. The cue suggests conditional expectation followed by the tower property. The counts are linear in i and the increments are constant, so the drift is affine and the expectations close into a scalar recurrence. No distribution is needed at any point.

3. Hesitation occurs at the moment you realise the walk is not a simple random walk — the step probabilities move with the state, and the instinct is that this makes the problem harder, so you reach for machinery. What reinforces it away is the observation that state-dependence is only fatal if it is *non-linearly* state-dependent. Test the drift for affineness before deciding the problem is hard. Here the i^2 terms cancel, and you can see they will as soon as you notice that $(n-i)(i+1)$ and $(i-1)^2$ both carry i^2 with opposite signs.

4. The trap is A^n. The matrix is easy to write down, the eigenvalues $\frac{n-2j}{n}$ are findable, and the whole route is visibly "the correct professional method" — and it will take most of the paper, because expressing the corner vector v_0 in the right-eigenvector basis of a non-symmetric tridiagonal matrix is not a five-minute job. This is zombie maths in its most respectable disguise: not brute-force arithmetic, but brute-force *structure*, invoked without first asking whether the question needs it. A secondary trap is numerical: $E(n)/n$ at $n = 4$ is 0.602, and the limit is 0.432, so small cases actively mislead about the answer.

5. Recognisable faster next time by a two-word test on the question: *which functional?* If the problem asks for an expectation, a sum, a trace, or any other linear functional of an evolving object, compute how that functional alone evolves before computing how the object evolves. Most of the time the functional's evolution is self-contained, and when it is not you have lost one line finding out.

6. Reusable decision rule: *a moment closes if and only if its drift is a polynomial of the same degree in the state.* Before setting up any machinery for a stochastic recursion, write $\mathbb{E}[X_{k+1} \mid X_k]$ and read off its degree in X_k. Degree ≤ 1 means the first moment satisfies a scalar affine recurrence and you are done in three lines; degree 2 means you need the second moment too and should check whether *that* closes; higher degree means the moment hierarchy does not truncate and you should look for a different invariant altogether. The same test governs branching processes, urn models and Markov chains on the integers, none of which resemble each other.

2024 B5

2024 B5. Let k and m be positive integers. For a positive integer n, let $f(n)$ be the number of integer sequences $x_1, \dots, x_k, y_1, \dots, y_m, z$ satisfying $1 \le x_1 \le \cdots \le x_k \le z \le n$ and $1 \le y_1 \le \cdots \le y_m \le z \le n$. Show that $f(n)$ can be expressed as a polynomial in n with nonnegative coefficients.

Fields: combinatorics (counting with repetition); polynomial algebra — root location **Position:** B5 — the hard half of the B paper, and the one whose difficulty is entirely in choosing what to prove rather than in any single step **Answer:** $f(n) = \sum_{z=1}^{n} \binom{z+k-1}{k}\binom{z+m-1}{m}$, and it has nonnegative coefficients because *all* of its roots are real and non-positive: $0, -1, \dots, -m$ from the vanishing of the summand, plus one root of Q in each interval $(-j-1, -j)$ for $j = 0, \dots, k-1$, located by a sign alternation.

Layer 1 THE WHAT — surface analysis

The natural attempt. The count itself is immediate and not the difficulty. Condition on z. Given z, the x-sequence is a weakly increasing k-tuple from $\{1, \dots, z\}$, i.e. a multiset of size k drawn from z values, and the y-sequence is independently a multiset of size m from the same z values. Writing

$$R(z, j) := \binom{z+j-1}{j} = \frac{z(z+1)\cdots(z+j-1)}{j!}$$

for the number of size-j multisets from $\{1, \dots, z\}$ — stars and bars — we get

$$f(n) = \sum_{z=1}^{n} R(z,k)\, R(z,m). \tag{$*$}$$

That took two minutes. Everything after it is where the problem lives.

The natural next move is to *compute* the polynomial and look at it. $R(z,k)R(z,m)$ is a polynomial in z of degree $k+m$ with nonnegative coefficients, so expand it in powers of z and sum each power using Faulhaber's formula. Now look at what you get. $\sum_{z\le n} z = \frac{n^2+n}{2}$, fine. $\sum_{z\le n} z^2 = \frac{2n^3+3n^2+n}{6}$, fine. $\sum_{z\le n} z^3 = \frac{n^4+2n^3+n^2}{4}$, fine. And then

$$\sum_{z\le n} z^4 = \frac{6n^5 + 15n^4 + 10n^3 - n}{30},$$

which has a negative coefficient. So the plan of "sum term by term and observe that each piece is nonnegative" is dead: the pieces are not nonnegative, and the Bernoulli numbers guarantee they will keep failing to be. Any proof by this route has to argue that the negative contributions cancel against positive ones, and cancellation arguments in a two-parameter family indexed by k and m are exactly what you do not want on a Putnam paper.

The second natural move — and it is the better instinct — is to change basis. The sum in $(*)$ telescopes beautifully in the R-basis, because hockey-stick gives

$$\sum_{z=1}^{n} R(z,j) = \binom{n+j}{j+1} = R(n, j+1),$$

and every $R(n,i)$ has nonnegative coefficients. So it would be enough to write $R(z,k)R(z,m)$ as a nonnegative combination of the $R(z,i)$. Test it on the smallest case, $k = m = 1$: we need z^2 in terms of $1, z, \frac{z(z+1)}{2}$, and the answer is

$$z^2 = 2R(z,2) - R(z,1).$$

A negative coefficient, in the very first case. So the linearisation coefficients are *not* nonnegative and this route dies too — even though the conclusion survives, since $2R(n,3) - R(n,2) =$

$\frac{n(n+1)(2n+1)}{6}$ does have nonnegative coefficients. That near-miss is the most useful thing on the page: it tells you that nonnegativity here is not inherited term by term from any natural decomposition. It is a property of the finished polynomial, and it will have to be proved as one.

A third instinct is to look for a bijection — to write $f(n) = \sum_i a_i n^i$ with a_i a count of something. That cannot work either, and you can see it cannot in five seconds: for $k = m = 1$, $f(n) = \frac{n(n+1)(2n+1)}{6} = \frac{n}{6} + \frac{n^2}{2} + \frac{n^3}{3}$. The coefficients are not integers. There is no set to biject with.

The model chain. Prove something stronger and more structural: *every root of f is real and non-positive.* Then $f(x) = c\prod_j (x + t_j)$ with $c > 0$ and all $t_j \geq 0$, and expanding that product gives nonnegative coefficients with nothing left to check.

Since f is symmetric in k and m, assume without loss of generality that $m \geq k$.

Step 1: f is a polynomial, of degree $k + m + 1$, satisfying a difference equation. $R(z,k)R(z,m)$ is a polynomial in z of degree $k + m$, and the sum of a polynomial over $z = 1, \ldots, n$ is a polynomial in n of one higher degree. So f is a polynomial of degree $k + m + 1$, and for every integer $n \geq 1$,

$$f(n) - f(n-1) = R(n,k)R(n,m),$$

which, holding at infinitely many points, is an identity of polynomials. Also $f(1) = R(1,k)R(1,m) = 1$, so $f(0) = 0$.

Step 2: f vanishes at $0, -1, \ldots, -m$. The polynomial $R(x,m) = \frac{x(x+1)\cdots(x+m-1)}{m!}$ vanishes at $x = 0, -1, \ldots, -(m-1)$. Feed this into the difference equation run backwards: $f(-j) = f(-j+1) - R(-j+1,k)R(-j+1,m)$, and for $j = 1, \ldots, m$ the argument $-j+1$ lies in $\{0, -1, \ldots, -(m-1)\}$, so the subtracted term is 0. Starting from $f(0) = 0$ and inducting,

$$f(0) = f(-1) = \cdots = f(-m) = 0.$$

Hence

$$f(x) = x(x+1)\cdots(x+m)\,Q(x), \qquad \deg Q = (k+m+1) - (m+1) = k.$$

Step 3: a difference equation for Q. Substitute into Step 1. Since $f(x-1) = (x-1)x(x+1)\cdots(x+m-1)Q(x-1)$ and $R(x,m) = \frac{x(x+1)\cdots(x+m-1)}{m!}$, every term carries the common factor $x(x+1)\cdots(x+m-1)$. Cancelling it,

$$(x+m)\,Q(x) - (x-1)\,Q(x-1) = \frac{R(x,k)}{m!}. \tag{†}$$

Step 4: sign alternation of Q at the non-positive integers. $R(x,k) = \frac{x(x+1)\cdots(x+k-1)}{k!}$ vanishes at $x = 0, -1, \ldots, -(k-1)$. Putting $x = -j$ in (†) for $j = 0, 1, \ldots, k-1$ kills the right-hand side and leaves

$$(m-j)\,Q(-j) + (j+1)\,Q(-j-1) = 0, \qquad \text{i.e.} \qquad Q(-j) = -\frac{j+1}{m-j}\,Q(-j-1).$$

This is where $m \geq k$ is used: $j \leq k-1 \leq m-1$, so $m - j \geq 1$ and we never divide by zero.

The multiplier $-\frac{j+1}{m-j}$ is a strictly negative number. So the $k+1$ values $Q(0), Q(-1), \ldots, Q(-k)$ are linked by nonzero factors: if one of them were zero they would all be zero, giving a degree-k polynomial with $k+1$ roots, forcing $Q \equiv 0$ and hence $f \equiv 0$, which is false since $f(1) = 1$. So all $k+1$ values are nonzero, and consecutive ones have opposite signs.

Step 5: locate the roots and finish. By the intermediate value theorem, Q has a root in each of the k open intervals

$$(-1, 0),\ (-2, -1),\ \ldots,\ (-k, -k+1),$$

and these are disjoint, so they are k distinct real roots of a degree-k polynomial — all of them, and all strictly negative. Combined with Step 2,

$$f(x) = c\,x(x+1)\cdots(x+m)\prod_{j=1}^{k}(x+t_j), \qquad t_j \in (j-1,\, j) \subset (0,\infty).$$

The leading coefficient c is positive because $f(n) > 0$ for large n. Every factor is a polynomial with nonnegative coefficients, so the product is too. □

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree exactly up to

$$f(n) = \sum_{z=1}^{n} R(z,k)R(z,m). \tag{$*$}$$

Nobody fails to reach $(*)$; it is the stars-and-bars count and it takes two minutes.

At $(*)$ the natural attempt tries to *build* the coefficients — expand, sum, change basis, look for the pieces that are individually nonnegative. The model chain gives up on building them and asks instead what a polynomial with nonnegative coefficients *is*, structurally, and goes after that structure. The split is between constructing the object and characterising it.

> *Two failed decompositions in a row is not bad luck. Faulhaber gives negative coefficients and the R-basis linearisation gives negative coefficients, yet the total comes out nonnegative both times. So nonnegativity is not a property this polynomial inherits from any of its natural pieces; it is a property of the finished object. I should stop decomposing and start characterising.*

> *What are the polynomials with nonnegative coefficients that I can actually certify? A product of factors $(x+t)$ with $t \geq 0$ is one, and that is the only sufficient condition I can check without knowing the coefficients. So I should try to prove the stronger statement that every root of f is real and non-positive. Strengthening to a root statement is the move, because roots I can find and coefficients I cannot.*

> *Where would I find non-positive real roots of f? I know f only through the recurrence $f(n) - f(n-1) = R(n,k)R(n,m)$ at positive integers. But a polynomial identity holds everywhere, so I can run that recurrence* backwards *through $n = 0, -1, -2, \ldots$ — and the right-hand side $R(n,m)$ is exactly zero at $n = 0, -1, \ldots, -(m-1)$. The recurrence hands me $m+1$ roots for free.*

> *That accounts for $m+1$ of the $k+m+1$ roots and leaves k. I cannot compute them, but I do not need to: I only need them real and negative, and the intermediate value theorem gives that from a sign pattern. The same backward recurrence, applied to the quotient Q, evaluates Q at $0, -1, \ldots, -k$ up to a known multiplier, and the multiplier $-\frac{j+1}{m-j}$ is negative. Alternating signs at $k+1$ consecutive integers is exactly k roots by IVT — and* $\deg Q = k$*, so that is all of them.*

- ▷ **Cue:** "prove this polynomial has nonnegative coefficients" with no explicit formula in sight ⟶ **Move:** strengthen to "all roots are real and non-positive"; a product of $(x+t)$ with $t \geq 0$ has nonnegative coefficients, and roots are locatable when coefficients are not
- ▷ **Cue:** a polynomial defined as $\sum_{z=1}^{n} g(z)$ for a polynomial g ⟶ **Move:** it satisfies $f(n) - f(n-1) = g(n)$ as a polynomial identity, so run the recurrence backwards past $n = 0$ — the zeros of g at non-positive integers become roots of f

▷ **Cue:** a recurrence linking $Q(-j)$ to $Q(-j-1)$ by a factor of constant sign $\longrightarrow$ **Move:** if the factor is negative you have a sign alternation, hence a root of Q strictly between each consecutive pair, by IVT

▷ **Cue:** a two-parameter symmetric problem where a step needs an inequality between the parameters $\longrightarrow$ **Move:** invoke the symmetry and assume WLOG; here $m \geq k$ is what keeps $m - j$ away from zero

Layer 3 THE WHY

Why must the strengthening be thought? Because of what a coefficient is, and what a root is, and which of the two this problem is actually able to give you.

Start from the definition. A polynomial with nonnegative coefficients is a nonnegative combination of monomials. Nothing in $(*)$ is a monomial. The summand $R(z,k)R(z,m)$ has nonnegative coefficients in z, and that fact is genuinely true and genuinely useless, because the operation applied to it — summation over $z = 1$ to n — does not preserve nonnegativity of coefficients. It cannot: $\sum_{z\leq n} z^4$ has a negative coefficient, so the operator $g \mapsto \sum_{z\leq n} g(z)$ carries a monomial with nonnegative coefficients to a polynomial without them. The single most tempting sentence in this problem — "each term is nonnegative, hence so is the sum" — is false about coefficients, and it is false for a reason you can name in one line.

So coefficients are the wrong currency. What is the right one?

Here is the axiom. There are exactly two ways to certify that a real polynomial has nonnegative coefficients: know them, or know the roots. If $f(x) = c\prod(x - r_i)$ with every r_i real and ≤ 0, then every factor is $(x + |r_i|)$ with $|r_i| \geq 0$, and a product of polynomials with nonnegative coefficients has nonnegative coefficients — because the coefficients of a product are sums of products of the coefficients, with no subtraction anywhere. Nonnegative coefficients is not a fragile analytic property here; it is closed under multiplication, and factorisation is the only lever that exploits that closure. Since you cannot know the coefficients, you must go for the roots. This is not a lucky strengthening. It is the only remaining option, and two failed decompositions are what tell you the other option has been exhausted.

Then the second question: how do you find roots of a polynomial you have never written down?

You have exactly one piece of information about f away from the positive integers, and it is the difference equation

$$f(n) - f(n-1) = R(n,k)R(n,m).$$

Say precisely why that is information. f was *defined* as a sum over $z = 1, \ldots, n$, an object with no meaning at $n = 0$ or $n = -3$. But the problem asserts, and Step 1 establishes, that f agrees with a polynomial at every positive integer — and a polynomial is determined by finitely many values. So the moment you know f is a polynomial, the difference equation stops being a statement about sums and becomes an identity of polynomials, valid at every real number. That licence — an identity proved at infinitely many integers holds identically — is what lets you evaluate a counting function at $n = -5$, where it counts nothing. This is the real conceptual content of B5 and it is worth saying out loud, because the instinct that $f(-5)$ is meaningless is correct about the count and wrong about the polynomial.

Once you accept the licence, the roots appear without being hunted. Running the recurrence backwards from $f(0) = 0$, each backward step subtracts $R(n,k)R(n,m)$, and $R(n,m)$ vanishes identically at $n = 0, -1, \ldots, -(m-1)$ because it is $\frac{n(n+1)\cdots(n+m-1)}{m!}$. A multiset count of size m from n objects is, as a polynomial, a product of m consecutive linear factors, and those factors are the roots. So f inherits $m + 1$ roots from the vanishing of its own increment. You did not find them; the combinatorial formula for R handed them over.

And the remaining k roots come from the same mechanism applied one level down. Ked-

laya's second solution states the relation

$$Q(-j) = -\frac{j+1}{m-j}\,Q(-j-1) \qquad (j = 0, \ldots, k-1)$$

with no motivation at all. Here is where it comes from. Having stripped off the $m+1$ known roots, you are left with Q of degree k, and you still have the same difference equation — now in the form (†) — and the right-hand side $R(x,k)/m!$ still vanishes at k non-positive integers, for the identical reason. Evaluating (†) at those zeros is not a trick; it is the only place where (†) simplifies, and evaluating an identity where one side is zero is the first thing anyone does with an identity. What comes out is a two-term relation with a coefficient of definite sign, and a definite negative sign at k consecutive integers is a sign alternation, and a sign alternation over $k+1$ points is k roots by the intermediate value theorem. Q has degree k. The count is exact, which is the signal that nothing has been wasted.

One last piece of bookkeeping earns its place. The hypothesis $m \geq k$ is needed to keep $m-j$ positive, and it is available for free because f is symmetric under exchanging the roles of x and y. Notice that without the symmetry the argument would have a genuine hole at $j = m$, and the hole would not be repairable by working harder — it would need the other choice of which $m+1$ roots to strip off first. A "without loss of generality" that does real work is worth flagging when you write it.

Caveat. I verified the root structure symbolically for all $1 \leq k, m \leq 4$: in every case f has degree $k+m+1$, all roots real and non-positive, integer roots at $0, -1, \ldots, -\max(k,m)$, and $\min(k,m)$ further roots lying one in each interval $(-j-1, -j)$ for $j = 0, \ldots, \min(k,m)-1$, exactly as the argument predicts; all coefficients came out nonnegative. I have not verified whether the roots t_j have any closed form, and none is claimed — the proof deliberately locates rather than computes them.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the pair of constraint chains $1 \leq x_1 \leq \cdots \leq x_k \leq z$ and $1 \leq y_1 \leq \cdots \leq y_m \leq z$ sharing the single upper bound z. On seeing it, write down: condition on z, count each chain by stars and bars, multiply, sum. That is $f(n) = \sum_{z=1}^{n} R(z,k)R(z,m)$ and it should be on the page within two minutes. The second cue — the real one — is the phrase "polynomial in n with nonnegative coefficients", which is a structural request, not a computational one.

2. The first cue suggests direct expansion and Faulhaber summation. Follow it far enough to see it fail: $\sum_{z \leq n} z^4$ has a negative coefficient. Then the R-basis, which also fails at $k = m = 1$ where $z^2 = 2R(z,2) - R(z,1)$. Those two failures are the method: they establish that nonnegativity is not inherited from any decomposition, which is what licenses the switch to a root argument.

3. Hesitation occurs at the moment you propose evaluating a counting function at a negative argument. $f(-3)$ counts nothing; the instinct that this is illegitimate is strong and it is correct about the count. What reinforces it away is stating the licence explicitly: f has been proved to agree with a polynomial at all positive integers, a polynomial is determined by finitely many values, and therefore the difference equation is a polynomial identity valid on all of $\mathbb{R}$. Write that sentence down before you use it; it costs one line and it is the load-bearing step.

4. The trap is the explicit formula. There *is* one — Kedlaya's first solution derives $f_{k,m}(n) = \frac{R(n,m+1)}{R(m+2,k)} \sum_{i=0}^{k} R(m+1,i)R(n,i)$ — and it is reachable by computing $f_{0,m}$, $f_{1,m}$, $f_{2,m}$, spotting the pattern, and inducting via a summation-by-parts identity. It is correct and it will take you most of the remaining time, because the pattern is not guessable from fewer than three

cases and the induction step is a page of binomial algebra. This is zombie maths wearing the costume of diligence: computing the object when the question only asked for one of its properties.

5. Recognisable faster next time by reading the *verb* in the question. "Show that $f(n)$ can be expressed as a polynomial with nonnegative coefficients" asks for a property, not a value. Whenever a question asks for a property of an object rather than the object, ask first which certificates of that property exist, and pick the certificate you have a chance of establishing. For nonnegative coefficients the certificates are: an explicit formula, or a factorisation into $(x+t)$ with $t \geq 0$. For a polynomial defined by a sum you will never have the first, so aim at the second before writing anything.

6. Reusable decision rule: *a counting function that is secretly a polynomial should be evaluated outside its combinatorial range.* Establish polynomiality, then treat every recurrence it satisfies as an identity on $\mathbb{R}$ and push it backwards past the boundary of meaning. The zeros of the increment become roots of the function, sign patterns at consecutive integers become roots by IVT, and the polynomial's structure is read off from a region where it counts nothing at all. This applies far beyond B5 — to Ehrhart polynomials, chromatic polynomials, binomial-coefficient identities extended to negative upper index, and any "show this count has property P" where P is a property of polynomials rather than of counts.

2024 B6

2024 B6. For a real number a, let $F_a(x) = \sum_{n\geq 1} n^a e^{2n} x^{n^2}$ for $0 \leq x < 1$. Find a real number c such that

$$\lim_{x\to 1^-} F_a(x)e^{-1/(1-x)} = 0 \qquad \text{for all } a < c, \text{ and}$$
$$\lim_{x\to 1^-} F_a(x)e^{-1/(1-x)} = \infty \qquad \text{for all } a > c.$$

Fields: analysis (asymptotics of series; Laplace's method in discrete form) **Position:** B6 — the last problem on the paper and genuinely hard; the estimate is not deep but the bookkeeping is unforgiving **Answer:** $c = -\frac{1}{2}$. With $t := 1/(1-x)$ the sum is dominated by the terms near $n = t$, each of size about $t^a e^t$, and about $\sqrt{t}$ of them contribute, so $F_a(x)e^{-t} \asymp t^{a+1/2}$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The series is a theta-like object and the trained reflex is to compare it with an integral. Write the summand as an exponential,

$$n^a e^{2n} x^{n^2} = \exp\big(a\log n + 2n + n^2 \log x\big),$$

and set $u := -\log x > 0$, so $u \to 0^+$ as $x \to 1^-$. The exponent is $-un^2 + 2n + a\log n$, a downward parabola in n with vertex at $n = 1/u$ and vertex value $1/u$. So

$$F_a(x) \approx \int_0^\infty s^a e^{-us^2+2s}\, ds = e^{1/u} \int_0^\infty s^a e^{-u(s-1/u)^2}\, ds,$$

and the Gaussian integral has width $u^{-1/2}$ centred at $s = 1/u$, giving

$$F_a(x) \approx e^{1/u} \cdot u^{-a} \cdot u^{-1/2} \cdot (\text{constant}).$$

Now divide by $e^{1/(1-x)}$ — and stop, because $1/u = 1/(-\log x)$ is not $1/(1-x)$. The two differ, and they differ by an amount that matters: for $x = 1 - 1/t$,

$$\frac{1}{-\log x} = \frac{1}{\frac{1}{t} + \frac{1}{2t^2} + \cdots} = t - \tfrac{1}{2} + O(1/t),$$

so $e^{1/u} = e^t \cdot e^{-1/2} \cdot (1 + o(1))$. Here the discrepancy happens to be a bounded factor and the comparison survives. But you cannot know that until you have done the expansion, and the expansion is the thing the naive substitution silently skipped. If the exponent had been n^3 instead of n^2, or the prefactor e^{2n} had been $e^{2n \log n}$, the same carelessness would have changed the answer rather than a constant. That is the wall: not that the method fails, but that it produces a number you cannot certify, and B6 is marked on the certification.

The second thing the natural attempt skips is the sum-to-integral step itself. $\sum_n \approx \int ds$ is exactly right here, but it is right *because* the summand varies slowly on the scale of one step near the peak (the peak has width $\sqrt{t} \gg 1$), and a full Euler–Maclaurin justification of that is more work than the problem needs. Worse, the integral route delivers an asymptotic constant — a genuine limit at $a = -\frac{1}{2}$ — which is more than was asked for, at more than the price.

The model chain. Substitute $t := 1/(1-x)$, so $x = 1 - 1/t$ and $t \to \infty$ as $x \to 1^-$. The single estimate that controls everything is the two-sided bound on $\log x$:

$$-\frac{1}{t} - \frac{1}{t^2} \;\le\; \log x \;\le\; -\frac{1}{t} \qquad (t \ge 2). \tag{†}$$

The right-hand inequality is $\log(1-v) \le -v$; the left is $-\log(1-v) = v + \frac{v^2}{2} + \frac{v^3}{3} + \cdots \le v + v^2$ for $0 < v \le \frac{1}{2}$. Throughout, c denotes a positive constant independent of n and t, possibly different at each appearance.

The identity that makes the whole problem transparent is the completion of the square in the exponent:

$$2n - t - \frac{n^2}{t} = -\frac{(t-n)^2}{t} = -t\left(1 - \frac{n}{t}\right)^2. \tag{‡}$$

So after dividing by e^t, the summand is a Gaussian in n centred at $n = t$ with standard deviation of order $\sqrt{t}$, and the whole question is how many terms sit under that Gaussian and how big each is.

Case $a > -\frac{1}{2}$: the limit is ∞. Using the left half of (†), $x^{n^2} = e^{n^2 \log x} \ge e^{-n^2/t - n^2/t^2}$, so by (‡),

$$F_a(x)e^{-t} \;\ge\; \sum_{n \ge 1} n^a e^{-n^2/t^2}\, e^{-t(1-n/t)^2}.$$

Every term is positive, so we may keep only $t < n < t + \sqrt{t}$. On that range: $n^a \ge ct^a$ (take $c = 1$ if $a \ge 0$ and $c = 2^a$ if $a < 0$, since $n < 2t$); $e^{-n^2/t^2} \ge e^{-4}$; and $|n - t| < \sqrt{t}$ gives $t(1-n/t)^2 < 1$, so $e^{-t(1-n/t)^2} > e^{-1}$. The range contains at least $\sqrt{t} - 1$ integers. Hence

$$F_a(x)e^{-t} \;\ge\; c\, t^a \sqrt{t} = c\, t^{a+1/2} \longrightarrow \infty$$

since $a + \frac{1}{2} > 0$.

Case $a < -\frac{1}{2}$: the limit is 0. Using the right half of (†), $x^{n^2} \le e^{-n^2/t}$, so by (‡),

$$F_a(x)e^{-t} \;\le\; \sum_{n \ge 1} n^a e^{-t(1-n/t)^2} = \sum_{n \ge 1} n^a e^{-(n-t)^2/t}.$$

Fix $\varepsilon > 0$ with $a + \varepsilon < -\frac{1}{2}$ and split at $|n - t| = t^{1/2+\varepsilon}$.

Central range, $|n - t| \le t^{1/2+\varepsilon}$. For t large this forces $n \ge t/2$, so $n^a \le (t/2)^a = ct^a$ (recall $a < 0$), and the exponential is at most 1. The range holds at most $2t^{1/2+\varepsilon} + 1$ integers, so its contribution is at most $c\, t^{a+1/2+\varepsilon} \to 0$.

Left tail, $1 \le n < t - t^{1/2+\varepsilon}$. Here $n^a \le 1$ (as $a < 0$, $n \ge 1$) and $(n-t)^2/t > t^{2\varepsilon}$, so each term is at most $e^{-t^{2\varepsilon}}$; there are fewer than t such terms, giving at most $t\, e^{-t^{2\varepsilon}} \to 0$.

Right tail, $n > t + t^{1/2+\varepsilon}$. Here $n^a \le 1$ again, and we split the exponent in half:

$$\frac{(n-t)^2}{t} = \frac{(n-t)^2}{2t} + \frac{(n-t)^2}{2t} \;\ge\; \frac{t^{2\varepsilon}}{2} + \frac{(n-t)^2}{2t}.$$

Therefore the tail is at most $e^{-t^{2\varepsilon}/2} \sum_{j\ge 1} e^{-j^2/(2t)} \le e^{-t^{2\varepsilon}/2} \cdot \sqrt{\pi t/2} \to 0$, the sum being bounded by $\int_0^\infty e^{-v^2/(2t)} dv$ since the summand decreases in j.

All three ranges tend to 0, so $F_a(x)e^{-t} \to 0$.

Hence $c = -\frac{1}{2}$. □

(The two bounds together say $F_a(x)e^{-1/(1-x)} \asymp t^{a+1/2}$, which is the honest content: at $a = -\frac{1}{2}$ the quantity tends to a positive constant, and numerically that constant is about 1.075.)

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree through the first act: rewrite the summand as a single exponential, recognise the exponent $2n + n^2 \log x$ as a downward parabola in n, and locate its vertex. Call that line $(\star)$.

At $(\star)$ the natural attempt replaces the sum by an integral and the parabola by its exact vertex form in the variable $u = -\log x$ — it goes for an *asymptotic equality*, and inherits the obligation to reconcile $1/u$ with $1/(1-x)$ and to justify the sum-integral swap. The model chain replaces $\log x$ by the two inequalities in (†) at the outset, changing variables to $t = 1/(1-x)$ *before* doing any analysis, and thereafter proves only *two one-sided bounds* — one from below for $a > -\frac{1}{2}$, one from above for $a < -\frac{1}{2}$. It never computes the sum. The split is between computing an asymptotic and bracketing one.

> *The question asks for a threshold c, not for the asymptotic value. A threshold is a statement about two inequalities, not about an equality. So I should never try to evaluate this sum — I should try to squeeze it between $c\,t^\alpha$ and $C\,t^\alpha$, and I get to throw away whatever I like in whichever direction I am going.*

> *The problem divides by $e^{1/(1-x)}$, not by $e^{1/(-\log x)}$. The competition setter chose $1/(1-x)$. That choice is a hint about which variable to work in: put $t = 1/(1-x)$ first and force $\log x$ to be an* estimate *in terms of t, rather than putting $u = -\log x$ and hoping the reconciliation is harmless.*

> *$2n - t - n^2/t$ is $-(n-t)^2/t$. So after dividing by e^t the summand is a Gaussian bump in n, centred at $n = t$, with width $\sqrt{t}$. The answer is therefore going to be (height) × (number of terms under the bump) $= t^a \times \sqrt{t}$, and the threshold is $a + \frac{1}{2} = 0$. I now know the answer; the rest of the work is making the two inequalities honest.*

> *For the lower bound I only need* one *slab of terms, not all of them, because every term is positive. Taking $t < n < t + \sqrt{t}$ gives me $\sqrt{t}$ terms each of size at least ct^a, and that is the whole of the divergent case. Discarding almost the entire sum is not laziness; it is the correct move, because the discarded part is nonnegative.*

- ▷ **Cue:** a limit that asks for a *threshold* value of a parameter rather than a value of the limit ⟶ **Move:** prove two one-sided bounds of the same order, not one asymptotic; you may discard freely in the favourable direction
- ▷ **Cue:** a sum $\sum_n e^{\phi(n)}$ with ϕ concave and a sharp interior maximum ⟶ **Move:** complete the square; the answer is (peak value) × (width of the peak in integer steps), and the width is where ϕ drops by $O(1)$

▷ **Cue:** the limiting expression is normalised by $e^{g(x)}$ for a specific g ⟶ **Move:** change variables so that g becomes the variable; the setter's choice of normalisation names the right coordinate

▷ **Cue:** $\log x$ appearing as $x \to 1^-$ ⟶ **Move:** replace it immediately by the two-sided bound $-v - v^2 \le \log(1-v) \le -v$; never carry $\log x$ through an asymptotic calculation

▷ **Cue:** a lower bound needed for a sum of positive terms ⟶ **Move:** keep a single well-chosen block of terms and throw the rest away; the block should be the one where the summand is within a constant factor of its maximum

Layer 3 THE WHY

Why must the first of those thoughts be thought? Read the question again, with attention to its logical shape.

You are asked to *find a real number c* such that the limit is 0 below it and ∞ above it. Nothing is asked about $a = c$. Nothing is asked about the rate. This is a statement purely about which of two divergent behaviours occurs, and a statement of that kind is proved by inequalities in both directions — one per case — and by nothing else. Every asymptotic constant you compute is work the question did not commission, and on a B6 with forty minutes left, unrequested work is the whole difference between a solution and a page of correct fragments. That is why the model chain proves two one-sided bounds and never once writes "∼".

That licence has a concrete consequence that beginners systematically refuse to use: in the divergence case you are allowed to throw away nearly the entire sum. Every term $n^a e^{2n} x^{n^2}$ is positive. So the sum is at least any subsum. Choosing the subsum $t < n < t + \sqrt{t}$ and bounding each of its $\sqrt{t}$ terms below by ct^a is a two-line argument that settles half the problem. The reluctance to do it comes from a feeling that discarding information is sloppy. It is the opposite: it is the recognition that the discarded terms cannot hurt you, and any effort spent on them is effort spent proving something true but irrelevant.

Now the deeper structural question: where does $\sqrt{t}$ come from, and why is the threshold a half?

Drive it to the definition. The summand, after dividing by e^t, has logarithm

$$a \log n + 2n - t - \frac{n^2}{t} + O\left(\frac{n^2}{t^2}\right) = a \log n - \frac{(n-t)^2}{t} + O(1)$$

near the peak. The exponent is a quadratic in n with second derivative $-2/t$. That is the entire mechanism. A quadratic with curvature κ drops by 1 when you move $\kappa^{-1/2}$ from its vertex; here $\kappa = 2/t$, so the peak has width of order $\sqrt{t}$. A sum of positive terms whose logarithm is a concave function with a sharp interior maximum is, up to constants, *the peak value times the number of integers within one unit of the peak in the exponent* — that is Laplace's method, and in its discrete form it is not an approximation theorem but an obvious two-sided bound: at most (peak) × (everything, controlled by the Gaussian tail), at least (peak/e) × (the terms within the width).

So the answer factorises as

$$\underbrace{t^a}_{\text{height, from } n^a \text{ at } n \approx t} \times \underbrace{\sqrt{t}}_{\text{width, from the curvature } 2/t},$$

and the threshold is exactly the value of a at which the height's decay cancels the width's growth: $a + \frac{1}{2} = 0$. The $\frac{1}{2}$ in the answer is the $\frac{1}{2}$ in "square root", which is the $\frac{1}{2}$ in "a quadratic". It has nothing to do with the particular constants 2 and 9 and everything to do with the fact that the exponent $2n + n^2 \log x$ is a *quadratic* in n. Change x^{n^2} to x^{n^3} and e^{2n} to $e^{3n^2/\cdots}$ appropriately and the width becomes $t^{1/3}$-ish and the threshold moves; the structure of the argument does not.

Now the piece of the official solution that is stated without motivation. Kedlaya's write-up opens by setting $t := 1/(1-x)$ and immediately asserting

$$-\frac{1}{t} - \frac{1}{t^2} \le \log x \le -\frac{1}{t}.$$

Here is where that comes from, and why it is the first line rather than a lemma retrieved later.

The quantity being normalised by is $e^{1/(1-x)}$. That is the setter's choice, and it is a choice: $1/(1-x)$ and $1/(-\log x)$ agree to leading order but differ by $\frac{1}{2} + o(1)$, which is a factor of $e^{-1/2}$ in the answer. In a problem about *thresholds* that bounded factor is harmless; but you do not get to know it is harmless until you have compared the two, and the comparison is exactly (†). So (†) is not a technical lemma. It is the translation dictionary between the variable the problem is phrased in and the variable the summand is naturally expressed in, and writing it down first is what stops you from spending twenty minutes on an asymptotic in the wrong coordinate. The general principle is worth having: *when a limit is normalised by $e^{g(x)}$, work in the variable g.* The setter told you which coordinate to use by choosing the normalisation.

Finally, why two-sided? The upper bound $\log x \le -1/t$ is used in the convergence case (it makes x^{n^2} small, which is the direction you want when bounding above) and the lower bound $\log x \ge -1/t - 1/t^2$ in the divergence case (it makes x^{n^2} large). Each case uses exactly one half of (†), and the error term $-1/t^2$ costs only the harmless factor $e^{-n^2/t^2} \ge e^{-4}$ on the range $n < 2t$ where the lower bound is applied. Notice that this is the second time the argument survives because the range was restricted first: restricting n to a slab makes a crude estimate of $\log x$ adequate. Choose the range before you choose the estimate.

Caveat. I verified the answer $c = -\frac{1}{2}$ numerically by evaluating $F_a(x)e^{-1/(1-x)}$ in log-space at $t = 50,\ 200,\ 10^3,\ 5\times10^3,\ 2\times10^4$ for $a = -0.8,\ -0.5,\ -0.2$. The computed values scale as $t^{a+1/2}$ to three significant figures in the exponent (e.g. a factor of 20 in t multiplies the value by $20^{\pm0.3}$ for $a = -0.2$ and $a = -0.8$ respectively), and at $a = -\frac{1}{2}$ the quantity is constant at ≈ 1.075. The proof above establishes the two required limits; it does *not* establish the existence of a limit at $a = -\frac{1}{2}$, and the value 1.075 is reported as numerical observation only, not as a proved constant.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the normalisation $e^{-1/(1-x)}$ — not the series. On seeing a limit divided by $e^{g(x)}$, write $t := g(x)$ on the first line and re-express everything in t. The second cue is the pairing of e^{2n} with x^{n^2}: an exponential in n against an exponential in n^2 means the combined exponent is a downward parabola with an interior maximum, so write the summand as $\exp(a\log n + 2n + n^2\log x)$ and find the vertex before anything else.

2. The cue suggests the discrete Laplace method: locate the peak, complete the square, read off height and width. Completing the square gives $2n - t - n^2/t = -(n-t)^2/t$ exactly, so the peak is at $n = t$, the height is t^a and the width is $\sqrt{t}$. Height times width is $t^{a+1/2}$, which is the answer before any rigour has been applied.

3. Hesitation occurs at the substitution $\log x \approx -(1-x)$. It feels like the kind of approximation that is either obviously fine or fatally wrong, and you cannot tell which, so you stall. What reinforces it away is refusing to approximate at all: replace $\log x$ by the two-sided bound $-\frac{1}{t} - \frac{1}{t^2} \le \log x \le -\frac{1}{t}$ and carry both halves. One half is used in each case, the discarded error contributes a bounded factor on the range that matters, and there is nothing left to worry about. A second hesitation occurs at "may I really throw away all but $\sqrt{t}$ terms?" — yes, in the lower bound, because every term is positive.

4. The trap is the integral. $\int_0^\infty s^a e^{-us^2+2s}ds$ is exactly computable in terms of the error function and gives the sharp asymptotic including the constant — and it costs you a justification of

the sum-to-integral comparison, a reconciliation of $1/(-\log x)$ with $1/(1-x)$, and a great deal more than the question wanted. This is zombie maths in its analytic dialect: computing the quantity when you were asked only where it changes behaviour. The related trap is the n^a factor: it is tempting to treat it as negligible next to e^{2n}, and it is negligible *pointwise*, but it is precisely the factor whose exponent the threshold is about.

5. Recognisable faster next time by the shape of the statement, not the shape of the series. "Find c such that the limit is 0 below and ∞ above" is a request for a scaling exponent. Every such problem is answered by (size of the largest term) $\times$ (number of terms of comparable size), each estimated to within a constant factor. Write those two quantities down first, multiply, set the exponent to zero, and only then decide how to make it rigorous.

6. Reusable decision rule: *when asked for a threshold, prove two inequalities, never an asymptotic — and in each inequality restrict the range before estimating.* Bounding below, keep one block of terms near the maximum and discard the rest, which costs nothing because they are positive. Bounding above, split into a central block (count the terms, bound each by the peak) and tails (bound the exponential crudely, but keep half the exponent in reserve to guarantee convergence of the tail sum). The reserve trick — writing $\frac{(n-t)^2}{t} = \frac{(n-t)^2}{2t} + \frac{(n-t)^2}{2t}$ and using one half for smallness and one half for summability — is worth memorising on its own; it appears wherever a tail must be both small and summable.