DAWN PUTNAM SESSION

Putnam 2006 — Papers A and B

Putnam 120 Archive

Contents

Putnam 2006 — Paper A **5**
2006 A1 5
2006 A2 7
2006 A3 10
2006 A4 13
2006 A5 15
2006 A6 18

Putnam 2006 — Paper B **23**
2006 B1 23
2006 B2 25
2006 B3 27
2006 B4 30
2006 B5 32
2006 B6 35

Putnam 2006 — Paper A

2006 A1

2006 A1. Find the volume of the region of points (x, y, z) such that

$$(x^2 + y^2 + z^2 + 8)^2 \leq 36(x^2 + y^2).$$

Fields: solid geometry, Pappus's theorem **Position:** A1 — the opener **Answer:** $6\pi^2$: the region is a solid torus, the disc $(r-3)^2 + z^2 \leq 1$ in the half-plane revolved about the z-axis.

Layer 1 THE WHAT — surface analysis

The natural attempt. The inequality is handed over in raw Cartesian form, and nothing in the statement says "use cylindrical coordinates" – that has to be self-issued. Without issuing it, the instinct is to attack $(x^2 + y^2 + z^2 + 8)^2 \leq 36(x^2 + y^2)$ as a quartic inequality in three variables directly: expand the square, get $x^4 + y^4 + z^4 + 64 + 2x^2y^2 + 2x^2z^2 + 2y^2z^2 + 16x^2 + 16y^2 + 16z^2 \leq 36x^2 + 36y^2$, collect terms, and try to complete some square in x and y separately while z sits along for the ride. This is possible in principle – the expression is symmetric enough that it can be forced into a usable shape – but it means carrying six squared-and-cross terms through the algebra by hand, and every one of them has to be re-grouped correctly or the eventual bounding surface comes out wrong. Even once a usable inequality is extracted, the volume still has to be computed by a triple integral over an irregular Cartesian region, typically by slicing at fixed z and integrating the area of the resulting (x, y) cross-section, which is itself an annulus or disc whose boundary equation has to be re-derived from the messy quartic at each slice. The whole route works, but it reconstructs, term by term and slice by slice, machinery that a single substitution would have handed over for free.

The model chain. The right-hand side is $36(x^2 + y^2)$ and nothing else – it does not involve x and y separately, only the combination $x^2 + y^2$, and it does not involve z at all except through the left side. Set $r^2 = x^2 + y^2$ (so $r \geq 0$ is the distance from the z-axis) and rewrite the inequality entirely in r, z:

$$(r^2 + z^2 + 8)^2 \leq 36r^2.$$

Since $r^2 + z^2 + 8 \geq 8 > 0$ always, taking square roots (both sides nonnegative) is equivalent to

$$-6r \leq r^2 + z^2 + 8 \leq 6r.$$

The left inequality is automatic for $r \geq 0$, since its right side is negative or zero while $r^2 + z^2 + 8 \geq 8$. So the region is exactly

$$r^2 + z^2 + 8 \leq 6r \iff (r-3)^2 + z^2 \leq 1.$$

In the (r, z) half-plane ($r \geq 0$) this is a disc of radius 1 centred at $(3, 0)$; since r ranges only over $[2, 4]$ on this disc, it never touches the z-axis, so the region is a genuine, non-self-intersecting solid of revolution when the disc is swept through all $\theta \in [0, 2\pi)$ – a solid torus. By Pappus's theorem, the volume of a solid formed by revolving a planar region of area A, lying entirely on

one side of the axis, through a full turn about that axis, equals 2π times the distance from the axis to the region's centroid, times A. Here $A = \pi \cdot 1^2 = \pi$ and the centroid distance is 3 (by symmetry of the disc about $r = 3$), so

$$V = 2\pi \cdot 3 \cdot \pi = 6\pi^2.$$

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains are staring at the identical quartic inequality before anything happens: $(x^2+y^2+z^2+8)^2 \le 36(x^2+y^2)$. They part at the very first move – whether the solver represents x^2+y^2 as a single new variable before doing anything else, or starts expanding the quartic term by term in x, y, z separately. Everyone eventually needs $r^2 = x^2+y^2$ to exist somewhere in the argument; the difference is whether it is introduced as the very first line, collapsing the problem to two variables immediately, or discovered midway through an expansion that has already produced six cross terms that then have to be reassembled back into r^2.

> *The right side is 36 times x^2+y^2 and nothing else – it doesn't care about z, and it doesn't care about x and y individually, only their sum of squares. That is the signature of a region symmetric under rotation about the z-axis, so before I expand a single square I should name $r^2 = x^2+y^2$ and rewrite the whole inequality in r and z – two variables instead of three, and a shape I can actually picture in a half-plane.*

Layer 3 THE WHY

Why does noticing that the right side depends on x, y only through x^2+y^2 earn an immediate switch of variables, rather than being a pleasant coincidence to exploit later? Because "depends on x, y only through x^2+y^2" is exactly the algebraic statement that the defining inequality is invariant under every rotation $(x, y) \mapsto (x\cos\varphi - y\sin\varphi,\ x\sin\varphi + y\cos\varphi)$ about the z-axis – rotating x, y by any angle φ leaves x^2+y^2 fixed, and the left side already only involves x^2+y^2 and z. A region invariant under every rotation about an axis is, by definition, a solid of revolution about that axis. Once that is recognised, cylindrical coordinates are not a convenient trick, they are the coordinates the region was already defined in; writing it in Cartesian coordinates was the detour. And Pappus's theorem is not an external fact borrowed from a formula sheet – it is the cylindrical change-of-variables Jacobian $r\,dr\,d\theta\,dz$ applied to a region whose defining inequality has no θ in it at all, so the θ-integral peels off as a bare factor of 2π:

$$V = \int_0^{2\pi} \iint_{(r-3)^2+z^2\le 1} r\,dr\,dz\,d\theta = 2\pi \iint_{(r-3)^2+z^2\le 1} r\,dr\,dz = 2\pi\,\bar{r}\,A,$$

which is exactly the statement of Pappus's theorem, re-derived rather than invoked. The reusable habit: whenever a three-variable inequality's right-hand side collapses two of the variables into a single combination like x^2+y^2, that collapse is the invitation to a coordinate system with one fewer effective variable, and it should be taken on the first line, before any expansion, not discovered by expanding and then noticing the mess re-groups.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that $36(x^2+y^2)$ on the right involves x and y only through the combination x^2+y^2, and not at all through z. On seeing it, write down: "this region is symmetric under rotation about the z-axis; switch to $r^2 = x^2+y^2$ before expanding anything."
2. That cue suggests collapsing the problem from three Cartesian variables to two, (r, z), turning a quartic surface in space into a disc (or annulus) in a half-plane, and then applying Pappus's theorem to convert that planar area directly into a volume of revolution.

3. Hesitation sets in right after squaring the inequality, when the solver has $(r^2+z^2+8)^2 \le 36r^2$ and has to decide whether the constraint $r^2+z^2+8 \ge -6r$ needs separate treatment. What reinforces moving past it quickly is checking the sign directly: $r \ge 0$ forces $-6r \le 0$, while $r^2+z^2+8 \ge 8 > 0$, so that inequality is never in doubt and can be discarded in one line rather than carried through the rest of the argument as a live case.

4. The trap is treating the squared inequality as requiring two-sided case analysis (as it would for a general expression that could be negative), when here one side is structurally impossible to violate. Carrying an always-true inequality forward as an open case wastes the rest of the solution re-checking something already settled.

5. Recognise it faster by scanning any multivariable inequality's right-hand side first, before touching the left: if it is a bare multiple of x^2+y^2 (or any two-variable combination), that is the cue to substitute that combination as a single symbol immediately, independent of how complicated the left side looks.

6. Reusable rule: when a region's defining inequality in $\mathbb{R}^3$ depends on two of the coordinates only through their sum of squares, it is a solid of revolution about the third axis; convert to cylindrical coordinates before any expansion, reduce to a planar region in the half-plane, and get the volume via Pappus's theorem ($V = 2\pi\bar{r}A$) rather than a direct triple integral.

▷ **Cue:** a 3D inequality's right side is a bare multiple of x^2+y^2 (no z, no separate x, y) $\longrightarrow$ **Move:** substitute $r^2 = x^2+y^2$ immediately and work in the (r, z) half-plane; the region is a solid of revolution about the z-axis

▷ **Cue:** a solid of revolution's cross-section is identified as a disc (or other planar region) of known area and known centroid distance from the axis $\longrightarrow$ **Move:** apply Pappus's theorem, $V = 2\pi\bar{r}A$, instead of setting up a direct triple integral

▷ **Cue:** a squared expression $A^2 \le B$ needs to be un-squared into $-\sqrt{B} \le A \le \sqrt{B}$ $\longrightarrow$ **Move:** check whether one of the two resulting inequalities is automatically true from the sign of A and the nonnegativity of the other side, before carrying both as live cases

2006 A2

2006 A2. Alice and Bob play a game in which they take turns removing stones from a heap that initially has n stones. The number of stones removed at each turn must be one less than a prime number. The winner is the player who takes the last stone. Alice plays first. Prove that there are infinitely many n such that Bob has a winning strategy. (For example, if $n = 17$, then Alice might take 6 leaving 11; then Bob might take 1 leaving 10; then Alice can take the remaining stones to win.)

Fields: combinatorial game theory, primes, contradiction via consecutive composites **Position:** A2
Answer: No explicit n is produced or needed: assuming the set of Bob-win values is finite forces a contradiction, so it must be infinite.

Layer 1 THE WHAT — surface analysis

The natural attempt. The instinct on a "prove infinitely many n" problem about a two-player subtraction game is to go and find the n's – compute the P-positions (positions where the player to move loses) by hand from the bottom, hoping a clean periodic rule falls out, the way it does for ordinary Nim-subtraction games with a fixed finite move set. The legal move sizes here are $p-1$ for prime p: $1, 2, 4, 6, 10, 12, 16, 18, \ldots$. Calling 0 a P-position (no stones, the player to move has already lost), direct computation gives $1, 2$ as N-positions (move straight to 0), then 3: the only moves are 1 and 2 (since $4 > 3$), landing on 2 or 1, both N – so 3 is the first genuine P-position, matching nothing the mover can do. Continuing, $4, 5, 6, 7$ are all N (each has a move to 0 or to 3), and 8: moves $1, 2, 4, 6$ land on $7, 6, 4, 2$, all N – so 8 is P. Then $9, 10$ are N (moves to 8 or 0), but 11: moves $1, 2, 4, 6, 10$ land on $10, 9, 7, 5, 1$, all N – so 11 is P. The P-positions found

so far are $0, 3, 8, 11, \ldots$, with gaps $3, 5, 3$, no periodicity visible yet, and no way to know from four data points whether one ever will appear. Worse: the classical guarantee that a subtraction game's P-positions are eventually periodic requires the move set to be *finite*. Here the move set is $\{p - 1 : p \text{ prime}\}$, which is infinite, so that theorem does not even apply – the natural attempt isn't just slow, it has walked into a gap in its own toolkit, and continuing to compute cases by hand has no guaranteed payoff at all.

The model chain. Suppose, for contradiction, that the set B of n for which Bob wins (the P-positions, including $n = 0$ by convention) is finite; write $B = \{b_1, \ldots, b_m\}$ and let $M = \max B$. For every nonnegative integer $n \notin B$, n is an N-position, so by definition Alice (the player to move) has *some* legal move from n landing on a P-position: there exist $b \in B$ and a prime p with $n = b + (p - 1)$, i.e. $n - b + 1 = p$ is prime. So every $n \notin B$ satisfies: $n + 1 - b$ is prime for at least one $b \in B$.

Now produce one n for which this fails for *every* $b \in B$ simultaneously. Take $t = M + 1$ and consider the t consecutive integers $(t+1)! + 2,\ (t+1)! + 3,\ \ldots,\ (t+1)! + (t+1)$: the j-th of these is divisible by j (for $2 \le j \le t + 1$) and strictly greater than j, so each one is composite. Set

$$n = (t+1)! + t.$$

For each $b \in B$ (so $0 \le b \le M = t - 1$), $n + 1 - b = (t+1)! + (t + 1 - b)$, and $t + 1 - b$ ranges over $\{2, 3, \ldots, t + 1\}$ as b ranges over $\{0, \ldots, t - 1\}$ – exactly the block of consecutive composites just built. So $n + 1 - b$ is composite for every $b \in B$, meaning n has no legal move to any $b \in B$. Since M was chosen to exceed every element of B, $n > M$ so $n \notin B$; but n has no move to a P-position, which by definition makes n itself a P-position – contradicting $n \notin B$. This contradiction shows B cannot be finite, so infinitely many n are Bob-win positions.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both approaches start from the same definition: n is a Bob-win (P-)position exactly when every legal move from n lands on an N-position, i.e. n has *no* legal move to a P-position. They part on what to do with a set defined this negatively. The natural attempt treats "prove infinitely many n are P-positions" as an instruction to construct or characterise B explicitly, computing it position by position. The model chain instead never tries to pin down B's contents at all – it assumes B is finite, purely as a hypothesis to break, and asks what that hypothesis alone forces on every n outside it. The bifurcating point is the line "the P-positions found so far are $0, 3, 8, 11, \ldots$, with gaps $3, 5, 3$, no periodicity visible yet": that is exactly where reading the problem as "characterise B" should be abandoned for reading it as "assume B is finite and contradict that."

> *I don't need to know which n are Bob-wins – the problem only asks me to prove there are infinitely many, and "prove a set is infinite" is exactly the shape a contradiction argument is built for: assume it's finite, and see what that forces about everything outside it. If B is finite, then every n not in B must have a legal move to some fixed $b \in B$, which means $n + 1 - b$ is prime for some b in a finite, fixed list – that's no longer a statement about the game at all, it's a statement about primality that I can attack with ordinary number theory: find one n that's composite after subtracting each of finitely many fixed constants and adding one. That's the standard consecutive-composite-run construction, and it has nothing to do with knowing the actual shape of B.*

Layer 3 THE WHY

Why does assuming B finite convert a game-theoretic question into an arithmetic one, rather than just being a bookkeeping convenience? Because "n is a P-position" is, by definition, a purely negative statement: no legal move from n reaches another element of B. Its contrapositive – "n is not a P-position" – is a positive, existential statement: *some* legal move from n reaches B, i.e. $n - b$ is a legal move size for some $b \in B$. Legal move sizes are exactly $\{p - 1 : p \text{ prime}\}$,

so this existential statement translates verbatim into "$n+1-b$ is prime for some $b \in B$." That translation is only usable once B is a *concrete, finite* object one can quantify over – with B still open-ended (possibly infinite, possibly unbounded), "$n+1-b$ is prime for some $b \in B$" is not a checkable arithmetic condition, it is just a restatement of "n is not a P-position." Assuming finiteness is what turns the definition into a target a number-theoretic construction can hit: covering "$n+1-b$ prime for some b in a finite list" is a job for the fact that arbitrarily long runs of consecutive composite integers exist (via $(t+1)!+2, \ldots, (t+1)!+(t+1)$, each divisible by its own offset) – primes are sparse enough that no finite list of "is $n+1-b$ prime" conditions can be forced to succeed for every sufficiently structured n. The reusable habit: when a set is defined only by what it *excludes* reaching (a "no move lands here" condition), and the goal is to show it's infinite, try assuming finiteness and converting the resulting universal statement about the complement into a concrete arithmetic (or combinatorial) target — finiteness is frequently the only thing that makes such a target well-defined.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the phrase "prove that there are infinitely many n" applied to a set defined by a losing condition (no legal move reaches a fixed set). On seeing it, write down: "try assuming this set is finite, see what that forces about its complement," rather than reaching for direct construction.

2. That cue suggests a proof by contradiction converting the game-theoretic complement condition ("every $n \notin B$ has a move to B") into an arithmetic statement ("$n+1-b$ is prime for some b in the finite set B"), then defeating it with a consecutive-composite-integers construction.

3. Hesitation sets in after computing $0, 3, 8, 11$ by hand and finding no periodicity: the pull is to keep computing further terms, hoping a pattern appears one or two steps later. What reinforces stopping is recalling that eventual periodicity of subtraction-game P-positions is a theorem about *finite* move sets, and checking that the move set here (all $p-1$ for prime p) is unbounded – so there is no theoretical guarantee a pattern will ever stabilise, which removes the incentive to keep computing.

4. The trap is spending effort computing P-positions by hand in the hope of spotting an explicit rule, when the problem only demands infinitude, not characterisation – a much weaker and structurally different claim that does not require knowing a single specific Bob-win value of n.

5. Recognise it faster by treating any "prove infinitely many objects satisfy [a condition defined by non-reachability, non-membership, or failure of some existential]" as a standing cue to try assume-finite-and-contradict before attempting direct construction or characterisation.

6. Reusable rule: to prove a set B (defined by "$n \in B$ iff no legal move from n reaches B," or any similarly self-referential exclusion condition) is infinite, assume $B = \{b_1, \ldots, b_m\}$ is finite; the complement condition becomes "for every $n \notin B$, some fixed arithmetic relation holds between n and one of the b_i" (here, primality of $n+1-b_i$); defeat this using a construction that can dodge finitely many such relations simultaneously (here, a run of m consecutive composite integers via $(t+1)!+2, \ldots, (t+1)!+(t+1)$).

▷ **Cue:** prove a set defined by a "no move/no path reaches here" condition is infinite $\longrightarrow$ **Move:** assume it is finite; convert the resulting universal statement about the complement into a concrete arithmetic condition indexed by the finite set, then build one witness that violates all of the (finitely many) conditions at once

▷ **Cue:** need m integers, each avoiding primality (or another sparse arithmetic property) after a fixed shift $\longrightarrow$ **Move:** use a block of m consecutive composite integers, e.g. $(t+1)!+2, \ldots, (t+1)!+(t+1)$, each divisible by its own offset from $(t+1)!$

2006 A3

2006 A3. Let $1, 2, 3, \ldots, 2005, 2006, 2007, 2009, 2012, 2016, \ldots$ be a sequence defined by $x_k = k$ for $k = 1, 2, \ldots, 2006$ and $x_{k+1} = x_k + x_{k-2005}$ for $k \geq 2006$. Show that the sequence has 2005 consecutive terms each divisible by 2006.

Fields: linear recurrences, periodicity modulo N, pigeonhole **Position:** A3 **Answer:** Existence proof: extending the recursion backward past index 1 forces 2005 consecutive zeros, $x_{-4009}, \ldots, x_{-2005} = 0$; since the recursion is reversible, the sequence is purely periodic mod 2006, so that zero block reappears among the forward, positive-index terms.

Layer 1 THE WHAT — surface analysis

The natural attempt. The recursion is linear, so the standard machinery is to look for a closed form via its characteristic polynomial: rearranging $x_{k+1} = x_k + x_{k-2005}$ gives $x_{k+1} - x_k - x_{k-2005} = 0$, with characteristic equation $t^{2006} - t^{2005} - 1 = 0$. A strong student who has seen linear-recurrence-mod-N arguments before will reach for this immediately: study the roots of this degree-2006 polynomial modulo the prime factors of $2006 = 2 \cdot 17 \cdot 59$, determine the multiplicative order of those roots, and use that to bound the period of the sequence modulo each prime, then combine via CRT to get the period modulo 2006 itself. This is a real and eventually workable strategy, but it is heavy: it requires factoring a degree-2006 polynomial's behaviour mod three different primes, handling the (likely) case that the polynomial is not separable or that Frobenius-style order arguments need care in each residue field, and even after all of that, knowing the period does not by itself locate *where* a run of 2005 consecutive zeros occurs – it would still need a separate argument for why such a run exists at all, since periodicity alone says values repeat, not that a long run of zeros exists anywhere in the cycle. Pushed forward from $k = 1$ by direct computation instead, the first few terms $1, 2, \ldots, 2006, 2007, 2009, 2012, 2016, \ldots$ show no zeros at all and grow indefinitely, so brute-force forward search for a zero run has no visible foothold within reach of hand computation.

The model chain. The recursion is linear with the coefficient of x_{k-2005} equal to 1, so it can be solved for the *earliest* term in its window without any division:

$$x_{k+1} = x_k + x_{k-2005} \iff x_{k-2005} = x_{k+1} - x_k, \qquad k \geq 2006.$$

This lets the sequence be extended uniquely to negative indices by running the recursion backward, one step at a time, starting from the known block $x_1 = 1, x_2 = 2, \ldots, x_{2006} = 2006$. Taking $k = 2005, 2004, \ldots, 1$ in the rearranged recursion (each valid because x_{k+1}, x_k are both already known, consecutive integers):

$$x_{k-2005} = x_{k+1} - x_k = (k+1) - k = 1, \qquad k = 1, \ldots, 2005,$$

which gives $x_{-2004} = x_{-2003} = \cdots = x_0 = 1$. Combined with the given $x_1 = 1$, this is a run of 2006 consecutive terms all equal to 1: $x_{-2004}, \ldots, x_1$. Extending one further step, take $k = 0, -1, \ldots, -2004$ in the same rearranged recursion; each x_{k+1} and x_k now lies inside that run of 1's, so their difference is 0:

$$x_{k-2005} = x_{k+1} - x_k = 1 - 1 = 0, \qquad k = 0, -1, \ldots, -2004,$$

giving $x_{-2005} = x_{-2006} = \cdots = x_{-4009} = 0$ – exactly 2005 consecutive terms, all divisible by 2006 (indeed all equal to 0). The only issue is that these indices are negative, not the positive indices the problem's sequence actually starts from.

To transplant this block into positive indices, note that the state $(x_{k-2005}, \ldots, x_k)$ (a window of 2006 consecutive terms) determines x_{k+1} by the forward recursion, and – because the coefficient of x_{k-2005} is 1 – the *same* window together with x_{k+1} determines x_{k-2005} by the backward

recursion just used. So the map "shift the window forward by one" is a bijection on the (finite, since we work modulo 2006) set of possible states $(\mathbb{Z}/2006\mathbb{Z})^{2006}$: it is injective (the backward recursion inverts it) on a finite set, hence a bijection. An injective self-map of a finite set is a permutation, and every orbit of a permutation is a finite cycle with no pre-period – so the doubly infinite sequence, reduced modulo 2006, is *purely periodic*, with some period T. The zero block found at indices $-4009, \ldots, -2005$ therefore reappears, modulo 2006, at indices $-4009 + cT, \ldots, -2005 + cT$ for every integer c; choosing c large enough that $-4009 + cT \geq 1$ produces 2005 consecutive positive-index terms of the original sequence, each divisible by 2006.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains recognise the recursion is linear and start reasoning about its behaviour modulo 2006. They part on what "linear" is used for. The natural attempt uses linearity to justify studying the recursion's characteristic polynomial and the algebraic order of its roots – machinery aimed at controlling the sequence's long-run behaviour forward. The model chain instead asks a narrower, more immediate question of the same linearity: can this specific recursion be solved for its *earliest* term as easily as for its latest? Bifurcating point: the line $x_{k+1} = x_k + x_{k-2005}$ rearranged to $x_{k-2005} = x_{k+1} - x_k$ – everyone can write this rearrangement, but only one path asks what it buys immediately, applied to the known initial block $x_1, \ldots, x_{2006} = 1, \ldots, 2006$, rather than treating it as an algebraic curiosity en route to a characteristic-polynomial argument.

> *The recursion is linear, and the coefficient of the earliest term x_{k-2005} is just 1 – so I can solve for it by pure subtraction, no division mod 2006 required, and nothing stops me running the recursion backward from my known starting block instead of forward into the unknown. The moment I do that, differences of consecutive known values appear: $x_{k+1} - x_k = 1$ for every consecutive pair inside $1, 2, \ldots, 2006$, so the terms just before the sequence starts are forced to be $1, 1, 1, \ldots$ – and one more step back, differences of 1's are 0. I get 2005 zeros for free, I just don't have them at a positive index yet.*

> *I have a block of zeros, but at negative indices where the problem's sequence isn't even defined – so the remaining question isn't "does a zero block exist," it's "does whatever happens at negative indices have to happen again later at positive indices." Since the step-forward map on windows of 2006 consecutive residues mod 2006 is invertible (I just inverted it to go backward), it's a bijection on a finite set, and a bijection on a finite set can only produce cycles – so the whole doubly infinite sequence mod 2006 must repeat with some period, and my zero block is somewhere on that repeating cycle, which means it comes back around to positive indices too.*

Layer 3 THE WHY

Why does the recursion being solvable for its earliest term justify running it backward at all, rather than that just being a formal rearrangement with no content? Because reversibility of a recursion is precisely the statement that the map from one window of consecutive terms to the next is injective, and injectivity is what is actually needed to extend a sequence uniquely in *both* directions – forward, injectivity is automatic (a function always has a well-defined image), but backward extension requires that each window have at most one possible predecessor, which is exactly what "solve for x_{k-2005} uniquely in terms of later terms" supplies. That this solving requires no division here (the coefficient of x_{k-2005} is 1, a unit in every ring, not merely in fields where 2006 is invertible) is what makes the reversal available modulo a composite number like 2006 with no extra case-work about invertibility.

The second, larger "of course" is why reversibility upgrades to full periodicity rather than merely "the sequence can be continued backward." Working modulo 2006, a window of 2006 consecutive terms is an element of the finite set $(\mathbb{Z}/2006\mathbb{Z})^{2006}$. The forward-step map on

this finite set is injective exactly because the recursion is reversible (two different windows can't produce the same successor window, since the successor together with the recursion's coefficient-1 term determines the predecessor uniquely). An injective map from a finite set to itself is automatically bijective, by counting – this is not a special fact about this recursion, it is the general fact that finite sets admit no injective self-map that isn't onto. A bijection of a finite set decomposes it entirely into cycles, with no element left as a "tail" outside a cycle; equivalently, iterating a bijection of a finite set can never enter a state it did not also come from, so every orbit, run forward or backward, is periodic with the same period. That is precisely what licenses moving a phenomenon (the run of 2005 zeros) discovered by running the state *backward* from the known initial block into a true statement about the *forward*, positive-index sequence: read forward instead of backward, the same periodic orbit that contains the zero block must pass through it again. The reusable habit: whenever a linear recurrence has a leading (or trailing) coefficient equal to ± 1, check immediately whether it can be run backward from known initial data — a reversible recursion modulo a finite N is automatically purely periodic, and any structural feature (a long run of equal values, a run of zeros, a specific residue pattern) found anywhere on its orbit is guaranteed to recur, including at indices the problem actually asks about.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is a linear recursion whose defining relation involves a fixed window of consecutive terms with the earliest one carrying coefficient 1 (here, $x_{k+1} = x_k + x_{k-2005}$). On seeing it, write down: "this can be solved for x_{k-2005} by pure subtraction, so it can be run backward from the known initial data with no loss of information."

2. That cue suggests extending the sequence to negative indices explicitly, computing what the recursion forces just below the given starting block, and separately establishing that the recursion is a bijection on windows modulo N (hence purely periodic), so anything found in the backward extension recurs forward too.

3. Hesitation sets in right after finding the block of 2005 zeros at negative indices, when it's tempting to treat the problem as solved – but the problem's sequence is only defined for $k \geq 1$. What reinforces pushing on is asking explicitly whether the step-forward map on residue windows is invertible: since it manifestly is (the same subtraction that built the backward extension inverts it), the finite-set-bijection-implies-purely-periodic argument is available immediately, without needing to compute the actual period T.

4. The trap is assuming that finding *a* zero block anywhere (even at negative indices) is not enough and reaching for a direct forward search or a characteristic-polynomial period computation to locate one explicitly among positive indices – unnecessary, since purely periodicity (from the bijection argument) guarantees recurrence without ever computing the period's numerical value.

5. Recognise it faster by treating "linear recurrence with a ± 1 coefficient on the earliest or latest term, working modulo a fixed N" as a standing invitation to check reversibility first, before reaching for characteristic polynomials, generating functions, or CRT-based period bounds.

6. Reusable rule: if a linear recurrence over $\mathbb{Z}$ can be solved for its earliest (or latest) term with coefficient ± 1, then modulo any fixed N the step map on windows of consecutive terms is a bijection of a finite set, hence the sequence mod N is purely periodic; any structural pattern (equal runs, zero runs, a target residue configuration) obtained by extending the sequence backward from known initial data to a convenient nearby index is therefore guaranteed to recur at some positive index too, without needing to compute the period explicitly.

▷ **Cue:** a linear recurrence's earliest (or latest) term in its defining window has coefficient ± 1 $\longrightarrow$ **Move:** solve for that term explicitly (pure addition/subtraction, no inversion needed mod any N) and run the recursion backward from known initial data

▷ **Cue:** a recursion is reversible (injective step map) and is being studied modulo a fixed finite N $\longrightarrow$ **Move:** conclude the step map on windows is a bijection of a finite set, hence the

sequence mod N is purely periodic – any pattern found anywhere on the orbit (including via backward extension) recurs at every later, positive index too

2006 A4

2006 A4. Let $S = \{1, 2, \ldots, n\}$ for some integer $n > 1$. Say a permutation π of S has a *local maximum* at $k \in S$ if

(i) $\pi(k) > \pi(k+1)$ for $k = 1$;

(ii) $\pi(k-1) < \pi(k)$ and $\pi(k) > \pi(k+1)$ for $1 < k < n$;

(iii) $\pi(k-1) < \pi(k)$ for $k = n$.

(For example, if $n = 5$ and π takes values at $1, 2, 3, 4, 5$ of $2, 1, 4, 5, 3$, then π has a local maximum of 2 at $k = 1$, and a local maximum of 5 at $k = 4$.) What is the average number of local maxima of a permutation of S, averaging over all permutations of S?

Fields: combinatorics; probability (linearity of expectation) **Position:** A4 — mid-paper, but one of the cheapest problems on the 2006 paper once the right tool is named. The difficulty is entirely in refusing to count permutations. **Answer:** The average is $\dfrac{n+1}{3}$.

Layer 1 THE WHAT — surface analysis

The natural attempt. You compute small cases. For $n = 2$: 12 has one local maximum (at $k = 2$), 21 has one (at $k = 1$); average 1. For $n = 3$, list the six permutations: 123 (one, at $k = 3$), 132 (one), 213 (two: $2 > 1$ at $k = 1$ and $1 < 3$ at $k = 3$), 231 (one), 312 (two), 321 (one). Total 8, average $\frac{4}{3}$. For $n = 4$ you grind through 24 permutations and get 40, average $\frac{5}{3}$. The pattern $1, \frac{4}{3}, \frac{5}{3}$ suggests $\frac{n+1}{3}$.

Now you need a proof, and the instinct is induction: build a permutation of $\{1, \ldots, n\}$ from one of $\{1, \ldots, n-1\}$ by inserting the symbol n into one of the n gaps (including the two ends). Track what happens. The new symbol n is larger than everything, so it is always a local maximum: $+1$. Any old local maximum sitting next to the gap now has n as a neighbour and stops being one: -1 for each such peak. Every other peak is untouched, and no new non-n peak can appear, because inserting a larger neighbour never helps anyone be a maximum. Each old peak is adjacent to exactly two gaps (an end peak has the end gap on its outer side), and two peaks are never adjacent, so no gap is next to two peaks. Summing over the n insertion positions of a permutation with P peaks gives $nP + n - 2P$ peaks in total. Averaging over all $(n-1)!$ starting permutations,

$$E_n = \frac{n-2}{n} E_{n-1} + 1, \qquad E_2 = 1,$$

and the guess $E_{n-1} = \frac{n}{3}$ gives $E_n = \frac{n-2}{3} + 1 = \frac{n+1}{3}$. It works. It cost you three enumerations, a careful insertion argument with two edge cases, and a recursion to solve. None of that was necessary.

The model chain. Choose π uniformly at random from the $n!$ permutations. The average asked for is then the expectation $E[M]$, where M is the number of local maxima. Write

$$M = \sum_{k=1}^{n} X_k, \qquad X_k = \begin{cases} 1 & \text{if } \pi \text{ has a local maximum at } k, \\ 0 & \text{otherwise.} \end{cases}$$

Expectation is linear whether or not the X_k are independent, so $E[M] = \sum_k P(X_k = 1)$.

Ends. $X_1 = 1$ exactly when $\pi(1) > \pi(2)$. Swapping the values in positions 1 and 2 is a bijection from permutations with $\pi(1) > \pi(2)$ to those with $\pi(1) < \pi(2)$, so $P(X_1 = 1) = \frac{1}{2}$. The same holds for $k = n$.

Interior. For $1 < k < n$, $X_k = 1$ exactly when $\pi(k)$ is the largest of the three values $\pi(k-1), \pi(k), \pi(k+1)$. Cyclically permuting the values in these three positions (leaving all other positions fixed) is a bijection of the set of permutations that moves the position of the largest of the three through all three places; so each of the three positions is equally likely to hold the largest, and $P(X_k = 1) = \frac{1}{3}$.

Therefore

$$E[M] = 2 \cdot \frac{1}{2} + (n-2) \cdot \frac{1}{3} = \frac{n+1}{3}.$$

(For $n = 2$ there are no interior positions and the formula gives 1, which agrees.)

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains start from the same object: a count M of local maxima, averaged over all permutations. They split at the first line. The natural attempt reads "average over all permutations" as "add up $M(\pi)$ over $n!$ permutations and divide", and therefore organises the sum by *permutation*. The model reorganises the same double sum by *position*: $\sum_\pi \sum_k X_k(\pi) = \sum_k \sum_\pi X_k(\pi)$.

> *I'm asked for the average of a count. A count is a sum of yes/no indicators, one per place where the thing could happen, and the average of a sum is the sum of the averages. So I only need, for each position k, the chance that k is a local maximum — and that only depends on two or three entries.*

> *In a random permutation, any fixed set of positions has its values in a uniformly random relative order. So "$\pi(k)$ is the biggest of its three neighbours" has probability $\frac{1}{3}$ without my knowing anything else.*

Layer 3 THE WHY

Why is linearity the natural move rather than a trick? Look at how the problem *defines* a local maximum. It does so in three clauses, (i), (ii), (iii), and each clause is a condition on a window of two or three consecutive positions. That is the whole cue. The property "has a local maximum at k" is local by definition: it ignores everything outside the window. The count M is a sum of these local tests. Whenever a quantity is a sum of tests each of which only looks at a small piece of the object, the average of the quantity is the sum of the probabilities of the tests, and each probability is a question about a small piece. That is the definition of expectation applied to a sum, $E[\sum X_k] = \sum E[X_k]$, which holds with no independence assumption — and the X_k here are certainly dependent (two adjacent positions can never both be peaks), which is precisely what scares students away from it. Dependence does not matter for linearity. It matters only for variance.

Why is each probability so easy? Because of what a uniformly random permutation is. Its law is unchanged if you rearrange the values sitting in any fixed set of positions, since rearranging is a bijection of S_n. So the relative order of the three values in a window is uniform over the $3! = 6$ orders, and in exactly 2 of them the middle one is the largest. Symmetry does all the work; no counting is needed.

Why did the natural attempt go the long way? "Average over all permutations" is phrased as an average over permutations, and it pulls the eye towards enumerating permutations. The sum $\sum_\pi M(\pi)$ is a double sum over pairs (permutation, position). Summing over permutations first is the order the question is written in; summing over positions first is the order in which each inner sum is trivial to evaluate by symmetry. Swapping the order of a finite double sum is

always legal, so the decision is purely about which inner sum you can do in your head. That is the rule to carry away.

Kedlaya and Ng's second solution runs linearity the other way: instead of positions, it sums over *values* i and asks for the probability that the value i is a peak, using a circular arrangement of the symbols $1, \dots, n, *$ to make the ends behave like interior positions. It gets $\sum_i \frac{i(i-1)}{n(n-1)} = \frac{n+1}{3}$. It is correct, but notice that it is the more expensive choice of indicator: whether value i is a peak depends on where i sits, while whether position k is a peak depends only on a window of fixed shape. Choose indicators whose events depend on as little as possible.

The insertion argument in the natural attempt is not wasted knowledge: removing the maximum element and recursing is the standard route to the full *distribution* of the number of peaks (the reference in the official remark), and for that problem linearity alone is not enough. For the mean, it is overkill.

Caveat. Verified independently by exhaustive enumeration for $n = 2, \dots, 8$: the averages are $1, \frac{4}{3}, \frac{5}{3}, 2, \frac{7}{3}, \frac{8}{3}, 3$, matching $\frac{n+1}{3}$.

Layer $+\alpha$ BACKTESTING

1. The cue is the phrase "average number of" applied to a count of local events, together with the clause-by-clause definition that makes each event depend on a small window. On seeing it, write down $M = \sum_k X_k$ and $E[M] = \sum_k P(X_k = 1)$ before touching any example.

2. Linearity of expectation over indicator variables, followed by a symmetry argument for each indicator.

3. Hesitation comes at "but the X_k are not independent." Reinforce it away by recalling the one-line proof of linearity: $\sum_\pi \sum_k X_k(\pi) = \sum_k \sum_\pi X_k(\pi)$ is a rearrangement of a finite sum, and independence never enters.

4. The trap is enumeration: small cases are easy enough that you are rewarded for listing, and the listing leads to a guess that then needs an induction. A second, smaller trap is treating the end positions like interior ones and writing $\frac{n}{3}$.

5. Slogan: *the average of a count is the sum of the chances; if each chance depends on a small window of a symmetric random object, each is a symmetry fraction.*

6. Decision rule: **when asked for the mean of a quantity that counts occurrences, write it as a sum of indicators, choose the indicators whose events depend on the fewest coordinates, and compute each probability by symmetry — never enumerate the configurations.**

▷ **Cue:** "average number of …" over all permutations, subsets, colourings or other configurations $\longrightarrow$ **Move:** write the count as $\sum$ of indicators and use $E[\sum X_k] = \sum P(X_k = 1)$, independence not required

▷ **Cue:** an event about a uniformly random permutation that involves only the relative order of the values in a fixed set of m positions $\longrightarrow$ **Move:** each of the $m!$ relative orders is equally likely; the probability is (number of good orders)$/m!$

▷ **Cue:** a definition split into cases for the first position, the last position, and the interior $\longrightarrow$ **Move:** the end cases have a different window size; compute their contribution separately (here $\frac{1}{2}$ versus $\frac{1}{3}$)

▷ **Cue:** a double sum over (configuration, location) that is hard when summed over configurations first $\longrightarrow$ **Move:** swap the order and sum over locations first

2006 A5

2006 A5. Let n be a positive odd integer and let θ be a real number such that θ/π is

irrational. Set $a_k = \tan(\theta + k\pi/n)$, $k = 1, 2, \ldots, n$. Prove that

$$\frac{a_1 + a_2 + \cdots + a_n}{a_1 a_2 \cdots a_n}$$

is an integer, and determine its value.

Fields: algebra (symmetric functions, Vieta); trigonometry via complex exponentials **Position:** A5 — late in the paper, but it yields completely to one standard construction. The difficulty is in seeing that the construction is standard. **Answer:** The ratio equals $(-1)^{(n-1)/2}\, n$: that is, n if $n \equiv 1 \pmod 4$ and $-n$ if $n \equiv 3 \pmod 4$.

Layer 1 THE WHAT — surface analysis

The natural attempt. You try $n = 1$: $a_1 = \tan(\theta + \pi) = \tan\theta$, ratio 1. Then $n = 3$, with $t = \tan\theta$ and $s = \tan\frac{\pi}{3} = \sqrt{3}$:

$$a_1 = \frac{t + \sqrt{3}}{1 - \sqrt{3}t}, \quad a_2 = \tan(\theta + \tfrac{2\pi}{3}) = \frac{t - \sqrt{3}}{1 + \sqrt{3}t}, \quad a_3 = t.$$

Then $a_1 a_2 = \dfrac{t^2 - 3}{1 - 3t^2}$, so $a_1 a_2 a_3 = \dfrac{t^3 - 3t}{1 - 3t^2} = -\tan 3\theta$, and after putting $a_1 + a_2$ over the common denominator $1 - 3t^2$ you find $a_1 + a_2 + a_3 = \dfrac{9t - 3t^3}{1 - 3t^2} = 3\tan 3\theta$. Ratio -3. Encouraging: an integer, and the θ-dependence cancelled because both sum and product are multiples of $\tan 3\theta$.

For general n this route stalls. The addition formula writes each a_k as $\dfrac{t + \tan(k\pi/n)}{1 - t\tan(k\pi/n)}$, and the sum and product then involve elementary symmetric functions of $\tan(\pi/n), \ldots, \tan((n-1)\pi/n)$, which you do not know. You can guess from $n = 1, 3$ that the answer is $\pm n$ and that sum and product are both multiples of $\tan n\theta$, but you have no mechanism to prove either. This is the wall: the angles $k\pi/n$ are equally spaced, and trigonometric addition formulae do not see equal spacing.

The model chain. *Step 1: the a_k are the roots of an explicit polynomial.* For any real x with $\cos x \neq 0$,

$$\frac{1 + i\tan x}{1 - i\tan x} = \frac{\cos x + i\sin x}{\cos x - i\sin x} = \frac{e^{ix}}{e^{-ix}} = e^{2ix}.$$

Put $x_k = \theta + k\pi/n$. Then $e^{2inx_k} = e^{2in\theta}e^{2\pi ik} = e^{2in\theta}$. Write $\omega = e^{2in\theta}$. Raising the display to the nth power,

$$(1 + ia_k)^n = \omega\,(1 - ia_k)^n,$$

so each a_k is a root of

$$Q(x) = (1 + ix)^n - \omega(1 - ix)^n = c_n x^n + c_{n-1}x^{n-1} + \cdots + c_0.$$

Step 2: the hypotheses make this legitimate. Since θ/π is irrational: no x_k is an odd multiple of $\frac{\pi}{2}$, so every a_k is defined; no x_k is a multiple of π, so every $a_k \neq 0$; and $2n\theta$ is not a multiple of π, so $\omega \neq \pm 1$. The x_k, $k = 1, \ldots, n$, are distinct modulo π (they differ by $j\pi/n$ with $0 < |j| < n$) and tan is injective modulo π, so the a_k are n distinct numbers.

Step 3: read off the coefficients. Because n is odd, $(-i)^n = -i^n$ and $(-i)^{n-1} = i^{n-1}$. Hence

$$c_n = i^n - \omega(-i)^n = i^n(1 + \omega), \qquad c_{n-1} = ni^{n-1} - \omega n(-i)^{n-1} = ni^{n-1}(1 - \omega), \qquad c_0 = 1 - \omega.$$

As $\omega \neq -1$, $c_n \neq 0$, so Q has degree exactly n; it has the n distinct roots $a_1, \ldots, a_n$, which are therefore all its roots. By Vieta,

$$a_1 + \cdots + a_n = -\frac{c_{n-1}}{c_n}, \qquad a_1 \cdots a_n = (-1)^n\frac{c_0}{c_n} = -\frac{c_0}{c_n}.$$

Step 4: divide.

$$\frac{a_1 + \cdots + a_n}{a_1 \cdots a_n} = \frac{c_{n-1}}{c_0} = \frac{n i^{n-1}(1-\omega)}{1-\omega} = n\, i^{n-1} = n\,(-1)^{(n-1)/2},$$

using $\omega \neq 1$ to cancel. This is n when $n \equiv 1 \pmod 4$ and $-n$ when $n \equiv 3 \pmod 4$; in particular it is an integer.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains know they need the sum and the product of the a_k. They split at the choice of representation for $\tan(\theta + k\pi/n)$. The natural attempt uses the real addition formula and is left holding unknown symmetric functions of $\tan(k\pi/n)$. The model uses $\dfrac{1 + i\tan x}{1 - i\tan x} = e^{2ix}$, under which the shift by $k\pi/n$ becomes multiplication by an nth root of unity, and the equal spacing collapses into a single polynomial equation.

> *Sum over product — that's e_1/e_n, two elementary symmetric functions. Symmetric functions of n numbers come from the polynomial whose roots they are. So I want a degree-n polynomial with roots $a_1, \ldots, a_n$, and I'll read the answer off its coefficients.*

> *The angles are θ plus multiples of π/n: equally spaced, n of them, going once round modulo π. Equally spaced angles mean roots of unity. So turn $\tan x$ into e^{2ix} — the map $a \mapsto \frac{1+ia}{1-ia}$ does it — and then n times each angle is the same thing, so all the a_k satisfy one equation.*

Layer 3 THE WHY

Why is "find the polynomial" the right first thought? Because of the shape of the target. The expression $\dfrac{a_1 + \cdots + a_n}{a_1 \cdots a_n}$ is symmetric in the a_k, and it is a ratio of two elementary symmetric functions, e_1/e_n. The only general-purpose machine that computes elementary symmetric functions without computing the individual numbers is Vieta: if you know the monic polynomial $\prod(x - a_k)$, you know every e_j. So the problem is really "produce the polynomial whose roots are the a_k". Kedlaya and Ng's solution opens with exactly that sentence; here is the reason behind it.

Why the complex exponential? Ask what is special about the numbers $\theta + k\pi/n$. They are an arithmetic progression of n terms with common difference π/n, and tan has period π, so they fill out one full period evenly. Arithmetic progressions of angles filling a period are, after exponentiating, the nth roots of a single complex number. That is the definition of the nth roots of unity: the solutions of $z^n = 1$ are $e^{2\pi ik/n}$, equally spaced round the circle. So you need a map that sends $\tan x$ to something like $e^{i\cdot(\text{multiple of } x)}$. The map is forced by Euler's formula: $\tan x = \dfrac{\sin x}{\cos x}$ and $\cos x \pm i \sin x = e^{\pm ix}$, so $1 \pm i\tan x = \sec x\, e^{\pm ix}$, and the ratio kills the $\sec x$. Once $\dfrac{1 + ia_k}{1 - ia_k} = e^{2ix_k}$, raising to the nth power sends every x_k to the same value, because $2n \cdot k\pi/n$ is a multiple of 2π. That single equation, cleared of denominators, is Q.

Why is n odd in the hypothesis? Step 3 is where it is used, and it is worth seeing that it is essential rather than decorative. For odd n, $(-i)^{n-1} = i^{n-1}$, which makes c_{n-1} a multiple of $1 - \omega$, the same factor as c_0; that common factor is what cancels the θ-dependence. For even n the signs swap: $c_{n-1} = n i^{n-1}(1 + \omega)$ while $c_0 = 1 - \omega$, and the ratio $n i^{n-1}\frac{1+\omega}{1-\omega}$ genuinely depends on θ. (Numerically, for $n = 2$ the ratio is about 2.19 at $\theta = 0.37$ and about -0.058 at $\theta = 0.8$.) The irrationality of θ/π is doing a humbler job: it guarantees every tangent is defined and nonzero and that $\omega \neq \pm 1$, so that Q has degree n and the cancellation is legal. It is more than is needed, and you should say which consequences you use.

A cross-check worth knowing. The same polynomial gives $a_1 \cdots a_n = -\frac{c_0}{c_n} = -\frac{1-\omega}{i^n(1+\omega)}$, and since $\frac{1-\omega}{1+\omega} = -i\tan n\theta$, this simplifies to $(-1)^{(n-1)/2}\tan n\theta$, while the sum is $n\tan n\theta$. That is exactly the pattern the natural attempt saw at $n=3$: $-\tan 3\theta$ and $3\tan 3\theta$. (The sum can also be found as $-\frac{d}{d\theta}\log\prod_k \cos x_k$, using the product formula $\prod_k \cos x_k = \pm 2^{1-n}\cos n\theta$ for odd n; but that product formula is itself proved with roots of unity, so it is the same idea wearing a different coat.)

The official remark says that the ratio of any two *odd* elementary symmetric functions of the a_k is independent of θ. You can see why from Q: for odd n, the coefficient c_j of Q equals $\binom{n}{j} i^j\big(1-(-1)^j\omega\big)$, so every c_j with j even carries the factor $1-\omega$ and every c_j with j odd carries $1+\omega$. The odd-index e_j are, up to sign, c_{n-j}/c_n with $n-j$ even; they all share $1-\omega$, and their ratios are free of θ.

Caveat. Verified independently: numerically for $n = 1,3,5,7,9,11$ at $\theta = 0.37$ the ratio is $1,-3,5,-7,9,-11$ to floating-point accuracy; for $n = 2,4$ it varies with θ, confirming that oddness is essential. The identities $\prod a_k = (-1)^{(n-1)/2}\tan n\theta$ and $\sum a_k = n\tan n\theta$ were also checked numerically for $n = 3,5,7,9$.

Layer $+\alpha$ BACKTESTING

1. The first cue is the target itself: a symmetric expression in n numbers, specifically (sum)/(product). Write down at once: *find the polynomial with roots $a_1,\ldots,a_n$; answer = ratio of its coefficients.* The second cue is $\theta + k\pi/n$ for $k = 1,\ldots,n$: an equally spaced set filling one period of tan. Write down $\dfrac{1+i\tan x}{1-i\tan x} = e^{2ix}$.

2. Vieta's formulas applied to a polynomial built from complex exponentials — the roots-of-unity method for equally spaced angles.

3. Hesitation occurs at "I need a polynomial for tangents, but tangents are not roots of unity." It is reinforced away by remembering that a Möbius map $a \mapsto \frac{1+ia}{1-ia}$ turns $\tan x$ into e^{2ix}; once the numbers are exponentials, powering is the natural operation.

4. The trap is the real addition formula, which works for $n = 3$ and rewards you enough to keep going, but leaves unknown symmetric functions of $\tan(k\pi/n)$ for general n. A second trap is sign bookkeeping: $(-i)^{n-1}$ versus $(-i)^n$, and the sign $(-1)^n$ in Vieta for the product. Getting one sign wrong gives $\pm n$ with the wrong case split.

5. Slogan: *equally spaced angles are roots of unity in disguise; convert the trigonometric function to an exponential and the whole family satisfies one polynomial.*

6. Decision rule: **when a symmetric function of $f(\theta + 2\pi k/m)$, $k = 1,\ldots,m$, is requested, express f through e^{ix}, find the single polynomial equation all the values satisfy, and read the answer from its coefficients by Vieta.**

▷ **Cue:** a symmetric expression (sum, product, their ratio, sum of squares) in n given numbers $\longrightarrow$ **Move:** build the polynomial with those numbers as roots and use Vieta

▷ **Cue:** angles $\theta + k\pi/n$ or $\theta + 2k\pi/n$, $k = 1,\ldots,n$ $\longrightarrow$ **Move:** exponentiate; n times the angle is constant, so the values are roots of one degree-n equation

▷ **Cue:** $\tan x$ to be related to e^{ix} $\longrightarrow$ **Move:** use $\dfrac{1+i\tan x}{1-i\tan x} = e^{2ix}$, equivalently $1 \pm i\tan x = \sec x\, e^{\pm ix}$

▷ **Cue:** a hypothesis such as "n odd" or "θ/π irrational" in a trigonometric identity $\longrightarrow$ **Move:** locate exactly which sign or nonvanishing it secures (here $(-i)^{n-1} = i^{n-1}$ and $\omega \neq \pm 1$), and check the even case fails

2006 A6

2006 A6. Four points are chosen uniformly and independently at random in the interior of a given circle. Find the probability that they are the vertices of a convex quadrilateral.

Fields: geometric probability; multivariable integration (polar coordinates, scaling) **Position:** A6 — the hardest slot on the paper, and it earns it: this is Sylvester's four-point problem for the disc. The reduction is quick; the integral is where exams are lost, and the choice of coordinates decides whether it is one page or four. **Answer:** The probability is $1 - \dfrac{35}{12\pi^2} \approx 0.7045$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The first step is one most strong students find. Take the circle to have radius 1, so the disc D has area π. With probability 1 no three of the four points are collinear. Four points in general position either form a convex quadrilateral or one of them lies inside the triangle formed by the other three, and at most one can do so (if S is inside triangle PQR, then triangle QRS lies inside PQR and cannot contain the vertex P in its interior). Conditioned on P, Q, R, the point S is uniform on D, so $\Pr(S \in PQR) = E([PQR])/\pi$, where $[PQR]$ is the area. With four choices of the interior point,

$$\Pr(\text{convex}) = 1 - \frac{4\,E([PQR])}{\pi}.$$

So everything rests on the expected area of a random triangle in the disc.

Now the natural move is coordinates centred at the centre O, because that is where the disc is simple. Write P, Q, R in polar form (r_j, φ_j). Then $[OPQ] = \frac{1}{2} r_1 r_2 |\sin(\varphi_2 - \varphi_1)|$ is pleasant, and

$$E([OPQ]) = \tfrac{1}{2}\Big(\int_0^1 2r^2\,dr\Big)^2 \cdot \frac{1}{\pi}\int_0^\pi \sin\psi\,d\psi = \frac{4}{9\pi}.$$

But $[PQR]$ is not $[OPQ] + [OQR] + [ORP]$. That identity holds only when O is inside triangle PQR. Otherwise $[PQR] = \pm[OPQ] \pm [OQR] \pm [ORP]$ with signs depending on whether the directions of P, Q, R from O lie in a semicircle and, if they do, whether the middle point is inside or outside the triangle formed by O and the other two. You now have three regimes, each with its own integral, one of which carries the indicator of a point lying inside a random triangle. This is Kane's first solution, and it does go through — but it is a page of case analysis with sign errors waiting at every step, and it is the wall most candidates hit at about the forty-minute mark. (Cartesian coordinates are worse: $\frac{1}{2}|\det|$ integrated over a six-dimensional region, with the absolute value unresolved.)

The model chain. *Step 1 (reduction).* As above, $\Pr(\text{convex}) = 1 - 4E([ABC])/\pi$ for A, B, C independent and uniform in the unit disc D.

Step 2 (scale out the outermost point). Let $c = \max(OA, OB, OC)$. For one uniform point, $\Pr(OA \le r) = r^2$; by independence $\Pr(c \le r) = r^6$, so c has density $6c^5$ on $[0,1]$. Condition on c and on which point attains the maximum (each of the three with probability $\frac{1}{3}$, and the area does not care which). Given that A is the outermost point at distance c, A is uniform on the circle of radius c (rotational symmetry), and B, C are independent and uniform in the disc of radius c (a uniform point conditioned to lie in a subdisc is uniform there). Dilating by $1/c$ about O maps this configuration to one with A on the unit circle and B, C uniform in D, and multiplies areas by $1/c^2$. Hence

$$E([ABC] \mid c) = c^2 X, \qquad X := E([RBC]) \text{ with } R \text{ a fixed point of the unit circle, } B, C \text{ uniform in } D,$$

and $E([ABC]) = \displaystyle\int_0^1 6c^5 \cdot c^2 X\,dc = \tfrac{3}{4}X.$

Step 3 (polar coordinates at R). Place R at the origin with the disc above the tangent line at R: $D = \{x^2 + (y-1)^2 < 1\}$. In polar coordinates (r, θ) about R this is $r^2 < 2r\sin\theta$, i.e.

$$D = \{(r,\theta) : 0 < \theta < \pi,\ 0 < r < 2\sin\theta\},$$

and the uniform density is $\frac{1}{\pi}\, r\, dr\, d\theta$. For $B = (r, \theta)$ and $C = (r', \theta')$, the triangle has a vertex at the pole, so

$$[RBC] = \tfrac{1}{2}\, rr' |\sin(\theta - \theta')|.$$

Step 4 (integrate the radii). $\int_0^{2\sin\theta} r \cdot r\, dr = \frac{8}{3}\sin^3\theta$. Therefore

$$X = \frac{1}{\pi^2} \cdot \frac{1}{2} \cdot \left(\frac{8}{3}\right)^2 I = \frac{32}{9\pi^2}\, I, \qquad I = \int_0^{\pi}\!\int_0^{\pi} \sin^3\theta\, \sin^3\theta'\, |\sin(\theta - \theta')|\, d\theta'\, d\theta.$$

Step 5 (the angular integral). The integrand is symmetric in θ, θ', and for $0 \le \theta' \le \theta \le \pi$ we have $\sin(\theta - \theta') \ge 0$. So

$$I = 2\int_0^{\pi} \sin^3\theta \int_0^{\theta} \sin^3\theta' \,(\sin\theta\cos\theta' - \cos\theta\sin\theta')\, d\theta'\, d\theta.$$

The first inner integral is $\int_0^{\theta} \sin^3\theta' \cos\theta'\, d\theta' = \frac{1}{4}\sin^4\theta$. Write $G(\theta) = \int_0^{\theta} \sin^4\theta'\, d\theta'$ for the second. Then

$$I = 2\Big[\frac{1}{4}\int_0^{\pi} \sin^8\theta\, d\theta - \int_0^{\pi} \sin^3\theta\cos\theta\, G(\theta)\, d\theta\Big].$$

Integrate the second term by parts with $\frac{d}{d\theta}\big(\frac{1}{4}\sin^4\theta\big) = \sin^3\theta\cos\theta$: the boundary term $\big[\frac{1}{4}\sin^4\theta\, G(\theta)\big]_0^{\pi}$ vanishes, leaving $-\frac{1}{4}\int_0^{\pi}\sin^8\theta\, d\theta$. Hence

$$I = 2\Big[\frac{1}{4} + \frac{1}{4}\Big]\int_0^{\pi} \sin^8\theta\, d\theta = \int_0^{\pi} \sin^8\theta\, d\theta = \pi \cdot \frac{7\cdot 5\cdot 3\cdot 1}{8\cdot 6\cdot 4\cdot 2} = \frac{35\pi}{128}$$

by the Wallis reduction $\int_0^{\pi}\sin^m = \frac{m-1}{m}\int_0^{\pi}\sin^{m-2}$.

Step 6 (assemble).

$$X = \frac{32}{9\pi^2}\cdot\frac{35\pi}{128} = \frac{35}{36\pi}, \qquad E([ABC]) = \frac{3}{4}X = \frac{35}{48\pi}, \qquad \Pr(\text{convex}) = 1 - \frac{4}{\pi}\cdot\frac{35}{48\pi} = 1 - \frac{35}{12\pi^2}.$$

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains agree through the reduction to $E([PQR])$. They split at the choice of origin. The natural attempt puts the origin at the centre O, where the *disc* is simple but the *area* is not: $[PQR]$ becomes a signed sum of three triangles through O with configuration-dependent signs. The model puts the origin at a vertex of the triangle, where the area is a single product $\frac{1}{2}rr'\sin\angle$, and uses scaling to make sure that vertex lies on the boundary, where the disc is still simple ($r < 2\sin\theta$).

> *Non-convex means one point is inside the triangle of the other three, and the chance of that, given the triangle, is its area over π. So the whole problem is the average area of a random triangle in the disc. Everything else is bookkeeping.*

> *Area is $\frac{1}{2}ab\sin C$ only from a vertex. From the centre it's a sum of three pieces with signs I'll get wrong. So my polar origin has to be one of the three points.*

> *But a random interior point makes the disc boundary ugly in polar coordinates. A point on the circle makes it $r = 2\sin\theta$. Area scales like length squared, so shrink the picture until the farthest of the three points from the centre sits on the rim: that costs me a factor c^2 and one easy integral in c, and pins a vertex to the boundary.*

Layer 3 THE WHY

This is a hard problem and there is no way to make the final integral vanish. What can be made natural is the *choice of coordinates*, which decides whether the integral is tractable. Three facts drive it, and each is at the level of a definition.

First: why reduce to expected area at all? Because of what "uniform" means. A uniform point lands in a region with probability (area of region)/(area of disc). The event "not convex" is a union of four events of the form "this point is inside the triangle of those three", and each is, conditionally, an area ratio. So the probability is an expected area, by the definition of the uniform distribution plus conditioning. The mutual exclusivity is the one rigour point here: a vertex of a triangle is never interior to a smaller triangle inside it.

Second: why a vertex as origin? Because the only formula for the area of a triangle that is a product — and therefore separates into independent integrals — is $\frac{1}{2}\,|RB|\,|RC|\sin\angle BRC$, and it is written from a vertex. From any other point the area is a signed combination of three sub-triangles, which is where Kane's case analysis comes from. The moment you notice that "area" wants a vertex, the origin is chosen. Kedlaya and Ng present Elkies' solution with the words "we introduce polar coordinates centered at R"; here is why R.

Third: why can we insist that the vertex be on the circle? Two facts. Area is homogeneous of degree 2: dilate by λ about O and areas multiply by λ^2. And the uniform distribution on a disc is self-similar: conditioned on lying within radius c, a uniform point is uniform on the smaller disc. So if you condition on the largest of the three distances from O, the configuration is a scaled copy of "one point on the rim, two inside", and the scale factor contributes only the one-variable integral $\int 6c^7\,dc$. This is the extremal principle in probabilistic dress: *condition on the extreme point*, because the extreme point is the one whose conditional law is simplest (it lives on a curve, not a region). Then a boundary point is exactly where polar coordinates for a circle are clean: by the inscribed-angle theorem, the chord from R making angle θ with the tangent has length $2\sin\theta$, which is the equation $r = 2\sin\theta$. A random interior pole would give $r(\theta) = -p\cos(\theta-\alpha) + \sqrt{1-p^2\sin^2(\theta-\alpha)}$, which no one wants to integrate.

Why the absolute value stops being a problem. With the pole on the boundary, every direction into the disc has $\theta \in (0,\pi)$, so the two angles θ, θ' live in one interval and $|\sin(\theta-\theta')|$ is resolved by ordering them — a symmetry, not a case analysis. From the centre, angles range over a full circle and the sign of sin of a difference depends on wrap-around, which is precisely what produced the semicircle cases.

Why the integral collapses to $\int \sin^8$. After expanding $\sin(\theta-\theta')$, one inner integral is an exact derivative, $\frac{1}{4}\sin^4$, and the other, G, is paired with that same derivative $\sin^3\theta\cos\theta$; integrating by parts moves the derivative onto G and returns $\sin^4\theta$. Both halves become $\frac{1}{4}\int\sin^8$. The official text calls the integral "unpleasant but straightforward". It is only unpleasant if you expand $\sin^4$ into multiple angles first. Look for the exact derivative before expanding; that is the cheap route, and zombie expansion here costs a quarter of an hour.

For context: the answer $\frac{35}{12\pi^2}$ for the non-convex probability is classical (Sylvester's problem for the disc). Crofton's formula gives another derivation; Savitt's second official solution computes the distribution of the distance from O to a random chord. Neither is cheaper under exam conditions than the scaling-plus-vertex-pole route.

Caveat. Verified independently. The angular integral $I = 35\pi/128$, $X = 35/(36\pi)$ and $E([ABC]) = 35/(48\pi)$ were confirmed symbolically (sympy). A Monte Carlo simulation with 4×10^6 samples gave $E([ABC]) \approx 0.23211$ (exact 0.23210), $X \approx 0.30944$ (exact 0.30947), and convex-position probability ≈ 0.70485 (exact $1 - 35/(12\pi^2) \approx 0.70448$), all within sampling error.

Layer $+\alpha$ BACKTESTING

1. The first cue is "uniformly … at random in the interior" combined with a condition on the

configuration (convexity). Write down: *non-convex* $\iff$ *one point in the triangle of the other three; probability* $= 4E[\text{area}]/(\text{area of disc})$. The second cue is that the target is now an expected *area*: write down $\frac{1}{2}ab\sin C$ and ask where its origin must be.

2. Reduce to an expectation by conditioning; then choose coordinates adapted to the quantity (polar at a vertex) and use scaling to pin that vertex to the boundary.

3. Hesitation occurs at the choice of origin. Everyone's reflex for a disc is the centre. Reinforce it away with the question: *which is harder to describe, the region or the integrand?* Here the integrand (area) is the harder one, so it gets the good coordinates, and scaling repairs the region.

4. Two traps. First, writing $[PQR] = [OPQ] + [OQR] + [ORP]$ without checking that O lies inside the triangle — it does so only with probability $\frac{1}{4}$. Second, expanding $\sin^3\theta \sin^3\theta' \sin(\theta - \theta')$ into multiple angles before looking for the exact derivative, which turns a three-line integral into a page.

5. Slogan: *put the origin where the integrand is simplest, and use homogeneity to move the region's awkwardness into a one-variable scale integral.*

6. Decision rule: **in a geometric-probability expectation of a homogeneous quantity (length, area, volume) over a self-similar region, condition on the extreme point, scale it to the boundary, and centre coordinates at that point; the scale contributes** $\int(\textbf{density of the maximum}) \cdot c^{\deg}\,dc$ **and the rest is one fixed-boundary integral.**

▷ **Cue:** "points chosen uniformly at random" and a question about whether one lies inside a figure formed by the others $\longrightarrow$ **Move:** condition on the others; the probability is expected area of the figure divided by the total area

▷ **Cue:** expected area of a triangle with random vertices $\longrightarrow$ **Move:** use polar coordinates centred at a vertex, so the area is $\frac{1}{2}rr'\sin(\theta - \theta')$, a product

▷ **Cue:** a homogeneous quantity of degree d averaged over i.i.d. uniform points in a disc or ball $\longrightarrow$ **Move:** condition on the largest radius c (density from $\Pr(c \le r) = r^{2m}$ for m points in a disc), scale by $1/c$, and the answer is $X \cdot E[c^d]$ with one point on the boundary

▷ **Cue:** polar coordinates about a point on a circle of radius 1 $\longrightarrow$ **Move:** the circle is $r = 2\sin\theta$ for $\theta \in [0, \pi]$ (inscribed-angle theorem)

▷ **Cue:** a double integral $\int\int f(\theta)f(\theta')|\sin(\theta - \theta')|$ over a square with both angles in $[0, \pi]$ $\longrightarrow$ **Move:** use symmetry to restrict to $\theta' < \theta$, expand $\sin(\theta - \theta')$, and look for an exact derivative before expanding powers

Putnam 2006 — Paper B

2006 B1

2006 B1. Show that the curve $x^3 + 3xy + y^3 = 1$ contains only one set of three distinct points, A, B, and C, which are vertices of an equilateral triangle, and find its area.

Fields: algebra (factoring cubics) and coordinate geometry **Position:** B1 – opener, deceptively geometric for what is really an algebra problem in disguise **Answer:** Unique; the area is $3\sqrt{3}/2$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The instinct on meeting an equation like $x^3 + 3xy + y^3 = 1$ is to treat it as a genuine curve and go looking for its geometric symmetry directly. The exponents $3, 1, 1, 3$ and the coefficient 3 on the mixed term look like a disguised cube – $(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$ – so the natural first move is to try to complete the cube, writing $x^3 + y^3 + 3xy = (x+y)^3 - 3x^2y - 3xy^2 + 3xy = (x+y)^3 - 3xy(x+y-1)$. Setting $s = x + y$ gives $s^3 - 3xy(s-1) = 1$, which is progress – one variable (s) and one product (xy) instead of two free coordinates – but it is progress into a dead end for the purpose of the problem, because it expresses a relation between s and xy, not a parametrisation of the curve, and the equilateral-triangle condition needs actual coordinates, not a symmetric-function identity. A student pursuing this line either tries to solve for y as a function of x (a genuine cubic in y for each fixed x, unpleasant and not obviously enlightening), or tries assuming the equilateral triangle is centred at the origin or has some rotational symmetry compatible with the equation's shape – a guess that has no justification from the algebra actually in hand, and which turns out to be wrong, since the true configuration is not centred at the origin at all.

The model chain. Move every term to one side: $x^3 + y^3 + 3xy - 1 = 0$, and rewrite $3xy$ as $-3xy(-1)$ and -1 as $(-1)^3$, so that the left side is exactly $x^3 + y^3 + z^3 - 3xyz$ with $z = -1$. This is the standard factorisation

$$a^3 + b^3 + c^3 - 3abc = (a+b+c)(a^2 + b^2 + c^2 - ab - bc - ca),$$

applied with $a = x,\ b = y,\ c = -1$. So

$$x^3 + y^3 + 3xy - 1 = (x+y-1)\big(x^2 + y^2 + 1 - xy + x + y\big).$$

The second factor is a positive-definite quadratic form up to the constant, since

$$x^2 + y^2 + 1 - xy + x + y = \tfrac{1}{2}\big[(x-y)^2 + (x+1)^2 + (y+1)^2\big] \geq 0,$$

with equality only at $x = y = -1$. So the "curve" is the union of the line $x + y = 1$ and the single point $(-1, -1)$. Three distinct points on this set cannot all lie on the line (collinear, no triangle) and cannot include two copies of $(-1, -1)$ (only one such point exists), so *exactly one* vertex of any triangle inscribed in the curve is $(-1, -1)$, and the other two lie on $x + y = 1$. For the triangle to be equilateral, the altitude from $(-1, -1)$ must be perpendicular to the base, i.e.

it must be the perpendicular dropped from $(-1,-1)$ to the line $x+y=1$, and its foot must be the midpoint of the base – both automatic for an equilateral triangle's altitude, and both requirements a single line satisfies uniquely (the perpendicular from a point to a line is unique). This pins the whole triangle down: the height is

$$h = \frac{|(-1)+(-1)-1|}{\sqrt{1^2+1^2}} = \frac{3}{\sqrt{2}},$$

and since an equilateral triangle of side s has height $h = \frac{\sqrt{3}}{2}s$, we get $s = 2h/\sqrt{3} = \sqrt{6}$, hence area $\frac{\sqrt{3}}{4}s^2 = \frac{\sqrt{3}}{4}\cdot 6 = \dfrac{3\sqrt{3}}{2}$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths notice the same surface pattern – exponents $3,1,1,3$ screaming "cube of a sum" – but they complete the cube differently. The natural attempt completes $(x+y)^3$, which is the cube that uses only the two variables already present and is therefore the first one that comes to hand. The model completes $a^3+b^3+c^3-3abc$ by manufacturing a third variable, $c=-1$, out of the bare constant on the right-hand side.

> *The constant* 1 *on the right is not just "the thing to subtract off" – it is exactly* $(-1)^3$*, and if I treat it as a third cube alongside* x^3 *and* y^3*, the coefficient* 3 *on* xy *is crying out to become* $-3xyz$ *with* $z=-1$*, because* $-3xy(-1)=3xy$ *is precisely the cross term the three-variable identity produces.*

> *A curve defined by one cubic equation in two variables is not obliged to be irreducible or connected – before hunting for its symmetry group or trying to parametrise it, I should ask whether it factors, because a factored cubic in two variables is either a union of a line and a conic, or three lines, and either case is far easier to reason about geometrically than the original curve.*

Layer 3 THE WHY

The reason $(x+y)^3$ is the wrong cube to complete, even though it is the more obvious one, is that it only ever uses the resources the problem hands you on the left-hand side; it throws away the fact that the right-hand side is also a perfect cube. The identity $a^3+b^3+c^3-3abc = (a+b+c)(a^2+b^2+c^2-ab-bc-ca)$ is not a trick pulled from nowhere – it is the factorisation of a circulant determinant, or equivalently the statement that $a+b+c$, $a+\omega b+\omega^2 c$, $a+\omega^2 b+\omega c$ (with ω a primitive cube root of unity) are the eigenvalue-style linear combinations that diagonalise the cyclic symmetry of $(a,b,c)\mapsto(b,c,a)$ acting on $a^3+b^3+c^3-3abc$. That structural fact is why the identity exists at all, and it is symmetric in three slots, not two – so the moment the constant term is recognised as a cube in its own right, using all three slots is simply respecting the symmetry the identity actually has, not adding a clever extra ingredient. Kedlaya's solution states the factorisation in one line without motivating why $c=-1$ suggests itself; the motivation is that $1=(-1)^3$ is the unique way to write the constant as a cube with the correct sign to make $-3abc$ match the given $+3xy$, and once that matching is seen, the algebra is forced, not guessed. The payoff – that the curve degenerates into a line and a point – is itself a consequence of the second factor being (half) a sum of three squares: a positive-definite quadratic form vanishes at only one point, so factoring a cubic into a linear times a positive-definite quadratic is exactly the mechanism by which a plane cubic curve can fail to be a curve at all in the usual sense, collapsing to a line plus an isolated point. Seeing that the second factor was going to be positive-definite is itself predictable in hindsight: it is $\frac{1}{2}$ of a sum of three squares $(x-y)^2+(x+1)^2+(y+1)^2$, which is the real part of $|a+\omega b+\omega^2 c|^2$-type expressions that the same cyclic identity always produces when a,b,c are forced real – so once

the factorisation is written down, checking positive-definiteness is routine, not an additional stroke of luck.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the exponent pattern $x^3, 3xy, y^3$ against a constant right-hand side of 1: three cubed terms and one degree-two cross term with coefficient 3. On seeing it, write down immediately that this matches $a^3+b^3+c^3-3abc$ and go looking for what plays the role of c.

2. That cue suggests treating the constant term as a third cube, $c^3=-1 \Rightarrow c=-1$ (matching the sign needed so $-3abc=+3xy$), and applying the standard factorisation rather than expanding $(x+y)^3$.

3. Hesitation occurs right after obtaining the factorisation, when a student is tempted to treat the second (quadratic) factor as just another branch of the curve to be explored geometrically. Reinforce it away by completing the square on that factor immediately – three cross terms and a constant almost always complete into a sum of squares, and checking for positive-definiteness costs one line and tells you at once whether that branch is a genuine curve or a single point.

4. The trap is treating a cubic-looking equation in two variables as necessarily an irreducible, one-piece curve, and searching for a symmetry (rotation, centring at the origin) that the equation's raw shape does not actually possess.

5. Recognisable faster next time by treating any equation of the shape $x^3+y^3+(\text{coefficient})\,xy=$ constant as a candidate for the three-cube identity the instant the coefficient on xy is ± 3 times something and the constant is a perfect cube (including negative cubes and $1=1^3=(-1)^3$ with sign chosen to match).

6. Reusable rule: before treating a two-variable cubic equation as a genuine curve to be parametrised or symmetrised directly, test whether it factors – specifically, whether the constant term can be absorbed as a third variable into $a^3+b^3+c^3-3abc$ – because a reducible cubic collapses the geometry (line plus conic, or line plus point) far more than any amount of direct manipulation of the original two variables will reveal.

▷ **Cue:** a two-variable cubic $x^3+y^3+kxy=c$ with $k=\pm 3$ and c a perfect cube $\longrightarrow$ **Move:** rewrite as $a^3+b^3+c^3-3abc$ with $a=x, b=y, c=\mp\sqrt[3]{c}$ chosen to match the sign of k, then factor via $(a+b+c)(a^2+b^2+c^2-ab-bc-ca)$

▷ **Cue:** a quadratic factor left over after factoring a cubic $\longrightarrow$ **Move:** complete the square to check whether it is positive-definite (collapsing to a single point) before treating it as a genuine second branch of the curve

2006 B2

2006 B2. Prove that, for every set $X=\{x_1, x_2, \ldots, x_n\}$ of n real numbers, there exists a non-empty subset S of X and an integer m such that

$$\left| m+\sum_{s\in S} s \right| \leq \frac{1}{n+1}.$$

Fields: combinatorics (pigeonhole on fractional parts) **Position:** B2 – second problem, a clean pigeonhole exercise that punishes the wrong pigeonhole **Answer:** Always true; the bound $1/(n+1)$ is achieved by a sharper argument than the obvious pigeonhole.

Layer 1 THE WHAT — surface analysis

The natural attempt. The phrase "there exists a subset and an integer m with $|m+\sum_S s|$ small" is a strong cue for partial sums and fractional parts: build $s_0=0$, $s_1=x_1$, $s_2=x_1+x_2$, …, $s_n=$

$x_1 + \cdots + x_n$, look at the $n+1$ fractional parts $\{s_0\}, \ldots, \{s_n\} \in [0,1)$, and note that if $\{s_i\}$ and $\{s_j\}$ ($i < j$) are close, then $S = \{x_{i+1}, \ldots, x_j\}$ and $m = \lfloor s_i \rfloor - \lfloor s_j \rfloor$ give $|m + \sum_S s| = |\{s_j\} - \{s_i\}|$ small, exactly the target shape. This is the right skeleton. The natural next move is the standard pigeonhole: split $[0,1)$ into n equal boxes of width $1/n$ and place the $n+1$ fractional parts into them; two must share a box, giving $|\{s_j\} - \{s_i\}| < 1/n$. This is a complete, correct proof – of a weaker statement. It delivers $1/n$, and the problem demands $1/(n+1)$, a strictly smaller bound, so the natural attempt is not wrong, it is simply insufficient, and a student who has proven $1/n$ and checks it against the target is left looking for where the extra room comes from, with the pigeonhole machinery already spent.

The model chain. Use the same $n+1$ partial sums $s_0, \ldots, s_n$ and the same fractional parts, but do not box them: sort $\{s_0\}, \ldots, \{s_n\}$ into ascending order $t_0 \le t_1 \le \cdots \le t_n$ in $[0,1)$. Consider the $n+1$ nonnegative gaps

$$t_1 - t_0,\ t_2 - t_1,\ \ldots,\ t_n - t_{n-1},\ 1 - t_n,$$

which sum to exactly 1 (they tile $[0,1)$ from 0 up to 1, including the final gap back to 1). An average of $n+1$ nonnegative numbers summing to 1 is $1/(n+1)$, so at least one of them is $\le 1/(n+1)$. If that minimal gap is one of the first n (a difference $t_k - t_{k-1} = |\{s_i\} - \{s_j\}|$ for some $i < j$), take $S = \{x_{i+1}, \ldots, x_j\}$ and $m = \lfloor s_i \rfloor - \lfloor s_j \rfloor$ exactly as before. If instead the minimal gap is the last one, $1 - t_n \le 1/(n+1)$, take $S = X$ (all of it) and $m = -\lceil s_n \rceil$, giving $|m + s_n| = \lceil s_n \rceil - s_n = 1 - \{s_n\} = 1 - t_n \le 1/(n+1)$ (or 0 if s_n is already an integer). Either way, the bound $1/(n+1)$ is achieved.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths build the same $n+1$ partial sums and reduce to bounding the smallest gap between fractional parts. They split at the moment of choosing the counting device: the natural attempt partitions $[0,1)$ into boxes and counts collisions (a pigeonhole on *position*), while the model sorts the values and sums the gaps between consecutive ones (an averaging argument on *spacing*).

> *I do not need two points to literally land in the same box to get a small gap – I only need* some *gap between consecutively-sorted points to be small, and $n+1$ nonnegative numbers that sum to a fixed total of 1 must have a minimum at or below their average, $1/(n+1)$, whether or not any of them coincide.*

> *The extra "+1" the problem wants is exactly the gap I was discarding: I have $n+1$ values in $[0,1)$, which give n internal gaps between them, but there is also the boundary gap back up to 1 – counting that gap as a genuine $(n+1)$th piece of the tiling, with its own subset $S = X$, is what turns $1/n$ into $1/(n+1)$.*

Layer 3 THE WHY

The classic pigeonhole (points into boxes) proves something strictly stronger than what is needed and, in doing so, wastes information: it guarantees two points *coincide in a box of a chosen size*, which is a binary yes/no fact, and the box size has to be chosen in advance without knowing where the points actually fall. Averaging is the correct tool here because the real object of interest – the $n+1$ gaps between consecutively sorted points, including the wraparound gap to 1 – is not a set of independent random landings but a partition of a segment of fixed total length 1 into $n+1$ nonnegative pieces. A partition into $n+1$ nonnegative numbers summing to a constant always has a piece at or below the average; this is not a probabilistic heuristic, it is the trivial fact that if every piece exceeded the average, the sum would exceed $n+1$ times the average, i.e. exceed the total, a contradiction. Kedlaya's solution states the sort-and-sum-gaps

step in a single line ("since $0 = t_0 \leq \cdots \leq t_n < 1$, the differences ... are nonnegative and add up to 1, hence one of these differences is no more than $1/(n+1)$") without pausing on why this beats the box argument; the reason is that boxing forces a commitment to n bins *before* looking at the data, discarding exactly the freedom that the sorted-gap method keeps: the bins are chosen *by* the data, as the gaps between the actual sorted values, so there are $n+1$ of them (one per point, wrapping around) rather than n (a fixed grid). The deeper point is that pigeonhole and averaging are the same principle at different resolutions – pigeonhole is averaging applied to a 0/1-valued (occupancy) function, while here the finer, continuous quantity (gap size) carries more information than mere occupancy, and extracting a bound from an average rather than from a collision is strictly more efficient whenever, as here, the quantity being bounded is itself the thing being summed to a known total.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the target expression $m + \sum_{s \in S} s$ together with an absolute-value bound: this is the shape of "distance to the nearest integer of a partial sum," so on seeing it, write down immediately the partial sums $s_0 = 0, s_1, \ldots, s_n$ and their fractional parts.

2. That cue suggests a pigeonhole-type argument on the $n+1$ fractional parts in $[0, 1)$, since a small difference of two fractional parts translates directly (via $m = \lfloor s_i \rfloor - \lfloor s_j \rfloor$) into the target quantity being small.

3. Hesitation occurs on checking the bound: the box-pigeonhole naturally produces $1/n$ (using n boxes to force a collision among $n+1$ points), one factor short of $1/(n+1)$. Reinforce it away by asking whether the argument needs a genuine collision at all, or only a small gap between sorted values – the latter is exactly what an averaging argument over $n+1$ pieces summing to 1 supplies, with no collision required.

4. The trap is assuming the strongest-sounding tool (force a literal coincidence via pigeonhole) is the tool the problem is asking for, when the problem only asks for smallness, and a softer averaging argument gets a sharper constant for free by using every one of the $n+1$ data points as a boundary of its own gap, including the wraparound gap to 1.

5. Recognisable faster next time by counting exactly how many objects and how many "slots" or "gaps" are in play before reaching for pigeonhole: if the count is (number of points) = (number of natural gaps, including a wraparound), an averaging-over-gaps argument is available and will usually beat a fixed-grid pigeonhole by exactly one in the denominator.

6. Reusable rule: when a pigeonhole argument on N points into $N-1$ fixed boxes gives a bound of $1/(N-1)$ but the target bound is $1/N$, look for a way to sort the N points and sum the N gaps between them (including any natural wraparound) to a known constant; the minimum of N nonnegative terms summing to a constant is at most their average, and this frequently recovers exactly the sharper constant the fixed-grid version loses.

▷ **Cue:** a bound of the form $1/(n+1)$ is wanted but a naive pigeonhole on $n+1$ points into n boxes only delivers $1/n$ $\longrightarrow$ **Move:** sort the $n+1$ values and sum the $n+1$ gaps between consecutive sorted values (including the wraparound gap back to the interval's endpoint); their minimum is at most the average $1/(n+1)$

▷ **Cue:** target expression has the shape $m + \sum_{s \in S} s$ for an integer m and subset S $\longrightarrow$ **Move:** build the partial sums $s_0 = 0, s_1, \ldots, s_n$ of the full ordered list and work with their fractional parts; a small difference of two fractional parts $\{s_i\}, \{s_j\}$ converts directly to $S = \{x_{i+1}, \ldots, x_j\}$, $m = \lfloor s_i \rfloor - \lfloor s_j \rfloor$

2006 B3

2006 B3. Let S be a finite set of points in the plane. A linear partition of S is an unordered pair $\{A, B\}$ of subsets of S such that $A \cup B = S$, $A \cap B = \emptyset$, and A and B lie on opposite

sides of some straight line disjoint from S (A or B may be empty). Let L_S be the number of linear partitions of S. For each positive integer n, find the maximum of L_S over all sets S of n points.

Fields: combinatorics (line arrangements) and induction **Position:** B3 – a counting problem whose upper bound is the real content, and where induction on a single removed point does all the work **Answer:** Maximum is $\binom{n}{2}+1$, achieved e.g. by a convex n-gon.

Layer 1 THE WHAT — surface analysis

The natural attempt. The construction end of the problem is genuinely easy and worth doing first: put n points on a circle, in convex position. A line separates the polygon into an arc of consecutive vertices and its complement, so every nontrivial linear partition corresponds to choosing which of the $\binom{n}{2}$ pairs of edges the separating line crosses, plus the one trivial partition (everything on one side). This gives $\binom{n}{2}+1$ and, crucially, a strong guess at the answer before the upper bound is even attempted – a useful habit, since it tells you what number you are trying to prove is unbeatable. The natural attempt at the *upper bound*, though, reaches first for the picture that produced the construction: draw all $\binom{n}{2}$ lines through pairs of points of S, and count the number of regions this arrangement cuts the plane into, since intuitively a linear partition is "a region of directions" for the separating line. For lines in general position, an arrangement of k lines cuts the plane into $1+k+\binom{k}{2}$ regions, which for $k=\binom{n}{2}$ is a much larger number than $\binom{n}{2}+1$ and visibly the wrong quantity – most of those regions correspond to the *position* of the separating line, not to distinct partitions of S, since many parallel-in-effect positions of the line induce the identical partition. Untangling which regions of that arrangement actually correspond to distinct partitions, and handling the special (non-generic) case where three or more points of S are collinear, turns into a delicate argument about directions at infinity rather than a clean count, and a student who starts here is doing real work for a bound that the counting device was not built to deliver directly.

The model chain. Prove the bound $\binom{n}{2}+1$ by induction on n, with no genericity assumption needed. Base cases $n=1,2$ are immediate: one point admits only the trivial partition ($L_S = 1 = \binom{1}{2}+1$); two points admit the trivial partition and the one partition separating them ($L_S = 2 = \binom{2}{2}+1$). For the inductive step, fix any point $P \in S$ and let $S' = S \setminus \{P\}$, with $|S'| = n-1$. Every linear partition $\{A,B\}$ of S restricts to a partition $\{A\setminus\{P\}, B\setminus\{P\}\}$ of S' (still separable by the same line, or a line perturbed slightly if it passed through P), so restriction is a well-defined map from partitions of S to partitions of S'. This map is surjective, since any separating line for S' can be translated slightly to put P on either side (unless the line already can be chosen to run exactly through P separating S', in which case both sides are achievable). A partition of S' has *two* preimages (one with P adjoined to each side) exactly when it can be achieved by some line through P; otherwise it has a unique preimage. Rotating a line through P by 180, the partition of S' it induces changes only when the line passes through one of the $n-1$ points of S', so there are at most $n-1$ distinct partitions of S' achievable by a line through P. Hence

$$L_S \;\le\; L_{S'} + (n-1) \;\le\; \left(\binom{n-1}{2}+1\right) + (n-1) = \binom{n}{2}+1,$$

using $\binom{n-1}{2} + (n-1) = \binom{n}{2}$. Combined with the convex-position construction achieving equality, the maximum is exactly $\binom{n}{2}+1$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths correctly identify that the answer should be $\binom{n}{2}+1$ and both correctly build the convex-position example. They split entirely on how to prove no configuration does *better*: the natural attempt tries to count regions in the static arrangement of all $\binom{n}{2}$ connecting lines at once, a global object built from the whole set S simultaneously; the model removes a single point and relates L_S to $L_{S\setminus\{P\}}$, a local, one-point-at-a-time recursion.

> *I do not need to understand the whole arrangement of $\binom{n}{2}$ lines at once – I only need to know how many* new *partitions appear when I add one point P to an already-understood set S', and that number is controlled just by how the partitions of S' interact with lines through the single new point P, which is a much smaller object (one point, $n-1$ lines to it) than the whole arrangement.*

> *The quantity that is easy to bound is not "how many regions does this arrangement have" but "how many partitions of S' can be split into two by inserting P", and that quantity is exactly the number of distinct partitions of S' realisable by a line through P – which is at most $n-1$, because rotating a line through a fixed point P only changes the partition of S' each time it sweeps past one of the $n-1$ other points.*

Layer 3 THE WHY

The arrangement-of-all-lines approach is not wrong in spirit – it is in fact salvageable (it is essentially the projective-duality argument Kedlaya records as a third solution) – but it is the harder route because it tries to answer a global question (how many partitions does S admit) with a global construction (every one of the $\binom{n}{2}$ lines determined by pairs of points), when the actual combinatorial content of a linear partition is local: a partition is determined by an interval of *directions* of the separating line, and what matters is only where those intervals begin and end, i.e. which single connecting line each interval's boundary crosses – not the full arrangement of all of them at once. The induction sidesteps this by asking a strictly smaller question at each step: having already classified every partition of the $(n-1)$-point set S', by how much can the count possibly grow when one more point P is reinserted? Since P can only ever be assigned to one side or the other of an existing partition of S', the count grows by at most one for each partition of S' that is *ambiguous* with respect to P – realisable with P on either side – and ambiguity is controlled by a one-dimensional rotation argument around the single point P, which is elementary regardless of how the other $n-1$ points are arranged, collinear or not. This is also why the induction needs no genericity hypothesis at all, while the arrangement-counting approach does: a arrangement of $\binom{n}{2}$ lines has degenerate (higher-multiplicity) intersection points whenever three or more points of S are collinear or three connecting lines happen to concur, and the region-count formula $1+k+\binom{k}{2}$ is valid only for lines in general position, forcing a separate perturbation argument to handle collinear configurations – exactly the case that produces the largest L_S (a convex polygon has no three collinear points, but nearby extremal configurations can, and the proof needs to cover them too). Kedlaya's second solution states "choose a point P in the convex hull of S" and proceeds directly to the rotation argument without dwelling on why induction on a single point beats the global count; the reason is precisely that the object being bounded (partitions of S' ambiguous with respect to one new point) is inherently one-dimensional – a rotation through 180 – while the object the natural attempt reaches for (regions of an arrangement of $\binom{n}{2}$ lines) is inherently two-dimensional and carries genericity baggage the problem never asked for.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that the construction (convex position) is easy and gives $\binom{n}{2}+1$ immediately by an "edges crossed" count. On seeing it, write down that the upper bound is the actual content of the problem, and that the construction's clean combinatorial description (choose a pair of edges) is a hint that the right proof technique should also be combinatorially clean, not a raw geometric arrangement count.
2. That cue suggests trying induction on $|S|$, specifically by removing one point and asking how the partition count can grow when it is reinserted – since the target formula $\binom{n}{2}+1$ has the shape of "previous value plus $(n-1)$", exactly matching $\binom{n-1}{2}+1+(n-1)=\binom{n}{2}+1$.
3. Hesitation occurs when trying to bound "how many partitions of S' become ambiguous when P is reinserted." Reinforce it away by physically imagining rotating a separating line

through P by 180: the induced partition of S' is locally constant and can only change when the rotating line sweeps past one of the $n-1$ points of S', so there are at most $n-1$ distinct partitions achievable this way – a one-dimensional sweep argument, not a count of regions.

4. The trap is reaching for the most symmetric-looking global object (the full arrangement of all $\binom{n}{2}$ connecting lines) to prove an upper bound, when the quantity actually being bounded only concerns how a *single* new point interacts with an already-classified smaller configuration – and committing to the global arrangement forces an unnecessary and error-prone genericity assumption.

5. Recognisable faster next time by testing, for any extremal counting problem over configurations of n objects, whether the target formula decomposes as $f(n-1)+g(n)$ for some simple g (here $g(n)=n-1$); if so, that decomposition is a direct invitation to induct by removing one object and bounding only the *increment* caused by reinserting it, rather than re-deriving the whole count from a static global picture.

6. Reusable rule: when an extremal count over n-point (or n-object) configurations has a target formula with a clean recursive shape $f(n)=f(n-1)+g(n)$, prove the bound by removing one point, applying the inductive hypothesis to the rest, and bounding only how many "boundary" cases become ambiguous when that one point is reinserted – usually via a one-parameter sweep or rotation argument around the removed point – rather than analysing the full configuration's arrangement at once.

▷ **Cue:** an extremal count over n-point configurations has a target formula of the recursive shape $f(n)=f(n-1)+g(n)$ (e.g. $\binom{n}{2}+1=\binom{n-1}{2}+1+(n-1)$) $\longrightarrow$ **Move:** induct by removing one point, apply the inductive hypothesis to the remaining $n-1$ points, and bound only the increment caused by reinserting the removed point – typically via a one-parameter rotation/sweep argument around that single point

▷ **Cue:** a static arrangement of all $\binom{n}{2}$ pairwise-connecting lines looks like the natural object to count regions in, but requires a general-position (no-three-collinear) assumption to get a clean formula $\longrightarrow$ **Move:** prefer a one-point-at-a-time induction, which needs no genericity hypothesis, over a global arrangement count that does

2006 B4

2006 B4. Let Z denote the set of points in $\mathbb{R}^n$ whose coordinates are 0 or 1. (Thus Z has 2^n elements, which are the vertices of a unit hypercube in $\mathbb{R}^n$.) Given a vector subspace V of $\mathbb{R}^n$, let $Z(V)$ denote the number of members of Z that lie in V. Let k be given, $0 \le k \le n$. Find the maximum, over all vector subspaces $V \subseteq \mathbb{R}^n$ of dimension k, of the number of points in $V \cap Z$. [Editorial note: the proposers probably intended to write $Z(V)$ instead of "the number of points in $V \cap Z$", but this changes nothing.]

Fields: linear algebra, combinatorics **Position:** B4 – middle of the paper, but one of the lighter B4s once the right object is chosen **Answer:** The maximum is 2^k, attained by the coordinate subspace $\{x_1 = \cdots = x_{n-k} = 0\}$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The answer 2^k is guessed in thirty seconds: a coordinate k-plane, say the span of $e_{n-k+1}, \ldots, e_n$, contains every 0/1 vector supported on those last k coordinates, and there are 2^k of them. The whole problem is the upper bound, and the natural attempt goes after it geometrically and by induction. Slice the cube by the two hyperplanes $V_0 = \{x_n = 0\}$ and $V_1 = \{x_n = 1\}$. Every point of Z lies in exactly one of them. If V is not contained in V_0, then $V \cap V_0$ has dimension $k-1$, and $V \cap V_1$ is either empty or an *affine* $(k-1)$-plane. So the induction hypothesis you want is about affine planes, not subspaces, and you have to go back and strengthen the statement to "any affine k-plane in $\mathbb{R}^n$ meets Z in at most 2^k points". Then

you need a separate case for $V \subseteq V_0$ or $V \subseteq V_1$ (drop the last coordinate and induct on n), the base case $k = n = 0$, and care about which of k and n decreases in each branch. This is Kedlaya's first solution and it is correct. It costs a strengthened hypothesis discovered midway, a double induction on $k + n$, and a case split, all to prove a bound that is one line long in the right language.

The model chain. Let V have dimension k and choose a basis $v_1, \ldots, v_k$. Write them as the rows of a $k \times n$ matrix M. Since the rows are independent, M has rank k, so its column rank is k and some k columns, indexed by a set $T = \{t_1, \ldots, t_k\}$, are linearly independent; equivalently the $k \times k$ submatrix M_T on those columns is invertible. Let $\pi_T : \mathbb{R}^n \to \mathbb{R}^k$ be the projection $x \mapsto (x_{t_1}, \ldots, x_{t_k})$.

Claim: π_T restricted to V is injective. A vector of V is $x = cM$ for a row vector $c \in \mathbb{R}^k$, and $\pi_T(x) = cM_T$. If $cM_T = 0$ then $c = 0$ because M_T is invertible, so $x = 0$. A linear map with trivial kernel is injective.

Now if $z \in V \cap Z$, then $\pi_T(z) \in \{0,1\}^k$, because every coordinate of z is 0 or 1. So π_T maps $V \cap Z$ injectively into $\{0,1\}^k$, a set of size 2^k. Hence $|V \cap Z| \leq 2^k$.

The coordinate subspace $\{x_1 = \cdots = x_{n-k} = 0\}$ has dimension k and contains all 2^k vectors with $0/1$ entries in the last k places and zeros elsewhere. So the maximum is exactly 2^k. (At $k = 0$ this reads $V = \{0\}$, which contains the single point $0 \in Z$, and $2^0 = 1$.)

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains agree on the answer and the example. They split at the first line of the upper bound. The natural attempt asks "how does V sit inside the cube?" and slices the cube. The model asks "what does a k-dimensional space let me *forget*?" and projects V down to k coordinates. Everything after that choice is forced in both chains; the choice itself is the problem.

> *A k-dimensional space is determined by k numbers. So a point of V should be determined by k of its coordinates, as long as I pick the right k. If that is true, a point of $V \cap Z$ is determined by k entries each equal to 0 or 1, and there are only 2^k ways to choose those.*

> *The target 2^k is itself a count of $0/1$ strings of length k. When the answer to a counting bound is "the number of objects of a smaller kind", I should be looking for an injection into those objects, not an induction.*

> *Which k coordinates? The ones where the basis matrix has a nonzero $k \times k$ minor. Rank k guarantees such a minor exists, and an invertible minor is exactly what makes the projection injective.*

Layer 3 THE WHY

Why is "project to k coordinates" natural rather than a trick? Go to the definition of dimension. A subspace has dimension k exactly when it is the image of an injective linear map $\mathbb{R}^k \to \mathbb{R}^n$, i.e. when it can be parametrised by k free real numbers. That is what dimension *means*. The question then is not whether V has k free parameters — it does, by definition — but whether those parameters can be taken to be k of the ambient coordinates themselves. That is the content of the rank statement: row rank equals column rank, so the $k \times n$ basis matrix has k independent columns, so on V the other $n - k$ coordinates are linear functions of those k. This is the same fact you use at A-level when you solve a system with free variables: the pivot columns are the free parameters, and every other variable is expressed in terms of them.

Once the free parameters are ambient coordinates, the cube structure does the rest. Z is defined coordinate by coordinate — each coordinate in $\{0,1\}$ — so the only way to exploit it is through coordinates. The natural attempt does use coordinates (it slices by x_n), but one at a

time, which is why it needs induction. The model uses k coordinates at once, which is why it doesn't.

Zara's solution, as printed by Kedlaya, phrases this as "row rank at most k, hence column rank at most k; choose k coordinates such that each point of $V \cap Z$ is determined by those k". The printed version is terse about *why* the determination holds. Here it is in full: if two points of V agree on the k chosen coordinates, their difference is a vector of V that vanishes on those coordinates, and injectivity of π_T on V (from the invertible minor) forces that difference to be 0. Note that injectivity is needed on all of V, not merely on $V \cap Z$; the linear argument gives it for free.

The second official solution, reducing mod 2 and counting a k-dimensional subspace of $\mathbb{F}_2^n$, is worth knowing as a different circuit (integrality lets you transport linear dependence to a finite field, where counting is exact), but it needs the step from real to rational to integer dependences, which is more to verify in the arena. When two routes exist, take the one whose every step is a definition.

A last point on the bound's sharpness. The injection is into $\{0,1\}^k$, and the coordinate subspace hits every element of it; so the inequality is attained, and there is nothing lost. When an injection argument meets an example that realises the full target set, you are done and need no further casework.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the answer shape: the example gives 2^k, and 2^k is the number of $0/1$ strings of length k. On seeing it, write down "find k coordinates that determine a point of V".

2. That cue suggests an injection argument: a map from $V \cap Z$ to $\{0,1\}^k$ that is visibly one-to-one. The map is the coordinate projection π_T; injectivity comes from rank.

3. Hesitation occurs at "which k coordinates?" — it is not true that *any* k coordinates work (the span of $e_1 + e_2$ is not determined by x_3). It is reinforced away by recalling that a rank-k matrix has an invertible $k \times k$ minor; choose its columns.

4. The trap is slicing the cube and inducting. It works, but it forces you to discover mid-proof that the hypothesis must be strengthened to affine planes, because $V \cap \{x_n = 1\}$ is not a subspace. A second, subtler trap is claiming injectivity only on Z-points without justification; do it on all of V by linearity.

5. Faster next time: whenever the conjectured bound is "the number of configurations of a smaller parameter set" (2^k, $\binom{k}{j}$, q^k), look for the parameters and an injection before any induction.

6. Reusable rule: a k-dimensional linear object is controlled by k well-chosen coordinates (pivot columns); to bound how many points of a coordinate-defined grid it can contain, project onto those coordinates and count the image.

▷ **Cue:** a bound to prove equals the number of strings of length k over a small alphabet, where k is a dimension $\longrightarrow$ **Move:** find k coordinates that determine each object (pivot columns / invertible $k \times k$ minor) and inject into the strings

▷ **Cue:** a subspace meets a set defined coordinate-by-coordinate (lattice, cube, grid) $\longrightarrow$ **Move:** choose the k pivot coordinates of a basis matrix; on the subspace, all other coordinates are linear functions of these

▷ **Cue:** an induction by slicing produces affine rather than linear pieces $\longrightarrow$ **Move:** either strengthen the statement to affine sets at the start, or look for a direct non-inductive argument that never slices

2006 B5

2006 B5. For each continuous function $f : [0,1] \to \mathbb{R}$, let $I(f) = \int_0^1 x^2 f(x)\,dx$ and $J(x) = \int_0^1 x\,(f(x))^2\,dx$. Find the maximum value of $I(f) - J(f)$ over all such functions f.

Fields: analysis (calculus of variations, inequalities) **Position:** B5 – nominally hard slot, but a completing-the-square problem once seen **Answer:** The maximum is $1/16$, attained at $f(x) = x/2$.

Caveat. The statement as printed writes $J(x)$ where $J(f)$ is plainly intended (the definition integrates over x and the question asks about $J(f)$). The analysis reads it as $J(f) = \int_0^1 xf(x)^2\,dx$.

Layer 1 THE WHAT — surface analysis

The natural attempt. A strong student who has met the calculus of variations writes down the functional $F(f) = \int_0^1 (x^2 f - xf^2)\,dx$ and perturbs: replace f by $f + \varepsilon g$ for an arbitrary continuous g and differentiate at $\varepsilon = 0$. This gives

$$\frac{d}{d\varepsilon} F(f + \varepsilon g)\Big|_{\varepsilon=0} = \int_0^1 \big(x^2 - 2xf(x)\big) g(x)\,dx.$$

For this to vanish for every g you need $x^2 - 2xf(x) = 0$, so $f(x) = x/2$ on $(0,1]$ and, by continuity, on $[0,1]$. Substituting, $F(x/2) = \int_0^1 (x^3/2 - x^3/4)\,dx = 1/16$. So the candidate answer is $1/16$.

Here the natural attempt stalls, and a careful student knows it. A stationary point of a functional is not automatically a maximum — it could be a minimum or a saddle, and on an infinite-dimensional space there is no "compact domain" to fall back on. To finish you need a second-variation argument ($\frac{d^2}{d\varepsilon^2} F = -2\int xg^2 \le 0$), then an argument that a concave functional's stationary point is a global maximum, and you have to be sure the fundamental lemma of the calculus of variations applies. All true; all extra; and much of it beyond what a Putnam grader will accept without proof.

The model chain. For each fixed $x \in [0,1]$, look at the integrand as a quadratic in the number $y = f(x)$:

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

(Check: $-x(y^2 - xy + x^2/4) + x^3/4 = -xy^2 + x^2 y$.) Hence for every continuous f,

$$I(f) - J(f) = \int_0^1 \frac{x^3}{4}\,dx - \int_0^1 x\Big(f(x) - \frac{x}{2}\Big)^2 dx = \frac{1}{16} - \int_0^1 x\Big(f(x) - \frac{x}{2}\Big)^2 dx.$$

On $[0,1]$ we have $x \ge 0$, so the subtracted integrand is nonnegative and the subtracted integral is ≥ 0. Therefore $I(f) - J(f) \le 1/16$ for every f. For $f(x) = x/2$, which is continuous, the subtracted integral is 0 and equality holds. The maximum is $1/16$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains start from the same integrand $x^2 f - xf^2$. The natural attempt treats f as a point in a function space and differentiates in that space. The model notices there are no derivatives of f and no coupling between different values of x, and so treats $f(x)$ at each x as an independent real number. The split happens at the line "the integrand is $x^2 f(x) - xf(x)^2$": one reader sees a functional, the other sees a quadratic in $f(x)$.

> *The integrand only involves $f(x)$ at the same point x — no f', no $f(1-x)$, nothing that ties different points together. So I can maximise the integrand pointwise: for each x, which real number y makes $x^2y - xy^2$ largest?*

> *That is a downward quadratic in y with leading coefficient $-x \le 0$. I complete the square, and the maximum value and the maximiser both fall out, with the inequality proof built in.*

> *If the pointwise maximiser $y = x/2$ is itself a continuous function of x, then it is admissible and the pointwise bound is attained. Check that before writing anything else.*

Layer 3 THE WHY

Why is pointwise maximisation the natural first move rather than an inspiration? Go back to what an integral is: a (limit of a) sum of the integrand's values. If you are asked to make a sum large and each term depends on its own private variable, you make each term as large as possible separately. $\int_0^1 G(x, f(x))\,dx$ with no derivatives in G is exactly such a sum — the value $f(x_1)$ does not appear in the term at x_2. So maximising $I(f) - J(f)$ decouples into a continuum of one-variable problems, one for each x. The calculus of variations machinery exists to handle integrands with f' in them, where neighbouring values *are* coupled. Using it here is bringing a crane to lift a chair.

Each decoupled problem is "maximise $x^2y - xy^2$ over real y", a quadratic in y. For a quadratic, completing the square is not one technique among many; it is the proof of where the vertex is and what the extreme value is, simultaneously. And the completed form $-x(y - x/2)^2 + x^3/4$ shows directly why the bound holds (a square times $-x \le 0$) and exactly when equality holds ($y = x/2$). That is why Kedlaya's solution is one displayed chain with no discussion: the square is the whole proof. What his solution does not say is where $x/2$ comes from; the answer is that it is the vertex $-b/(2a)$ of $-xy^2 + x^2y$, namely $x^2/(2x)$.

Two structural facts make the pointwise bound legitimate, and both come from the problem's phrasing. First, the domain is $[0,1]$, so the weight x in J is nonnegative; if the domain had included negative x, the quadratic would open upwards there, $I - J$ would be unbounded above, and the answer would be "no maximum". Second, the class is *all* continuous functions, and the pointwise optimiser $x/2$ is continuous, so nothing is lost by the continuity restriction. Had the constraint been, say, $f(0) = 1$, the pointwise optimiser would violate it, the supremum would still be $1/16$ (approach $x/2$ by continuous functions that fix a tiny neighbourhood of 0) but not attained, and the word "maximum" would be false. The problem is set so that the pointwise answer is admissible; your job is to check that it is, not to assume it.

Note also that the natural attempt found the right function and the right number. It did not fail; it produced a candidate without a proof. Completing the square produces the candidate and the proof in one line. That difference is the difference between a solution and a guess on a Putnam script.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that the integrand contains $f(x)$ and $f(x)^2$ at the same point and nothing else: no f', no shifted arguments. On seeing it, write the integrand as a quadratic in $y = f(x)$ with coefficients depending on x.
2. The method is pointwise optimisation by completing the square: the integrand becomes (a constant in f) minus (a nonnegative weight times a square), and the integral of the constant is the bound.
3. Hesitation occurs after the Euler–Lagrange-style computation gives $f = x/2$: "is this a maximum?" It is reinforced away by never needing to ask — completing the square proves the inequality and identifies the equality case at once.

4. The trap is the calculus-of-variations reflex, which finds the stationary point but leaves the global-maximum claim unproved. A second trap is forgetting to check the sign of the weight x and the admissibility (continuity) of the pointwise maximiser.

5. Recognise faster by asking of any integral functional: does the integrand couple different points of the domain (derivatives, shifts, double integrals)? If not, optimise pointwise.

6. Reusable rule: to extremise $\int G(x, f(x))\,dx$ with no derivative of f, extremise $G(x, \cdot)$ for each x; if G is quadratic in $f(x)$, complete the square, and then check the pointwise extremiser belongs to the admissible class.

▷ **Cue:** an integral functional whose integrand involves $f(x)$ only at the same point x, with no derivatives $\longrightarrow$ **Move:** optimise the integrand pointwise in the real variable $y = f(x)$

▷ **Cue:** a quantity is quadratic in the unknown and you want its maximum or minimum $\longrightarrow$ **Move:** complete the square; the square term gives the inequality and its equality case together

▷ **Cue:** a pointwise or unconstrained optimiser has been found $\longrightarrow$ **Move:** check it lies in the admissible class (continuity, boundary conditions, sign of weights) before claiming "maximum"

2006 B6

2006 B6. Let k be an integer greater than 1. Suppose $a_0 > 0$, and define

$$a_{n+1} = a_n + \frac{1}{\sqrt[k]{a_n}}$$

for $n > 0$. Evaluate

$$\lim_{n\to\infty} \frac{a_n^{k+1}}{n^k}.$$

Fields: analysis (asymptotics of recursions) **Position:** B6 – last problem, but one of the most approachable B6s once the right quantity is tracked **Answer:** The limit is $\left(\frac{k+1}{k}\right)^k$.

(The recursion is meant for $n \geq 0$, so that a_1 is defined from a_0; the printed "$n > 0$" is a slip and changes nothing.)

Layer 1 THE WHAT — surface analysis

The natural attempt. First, a heuristic, and a good one. The increments $a_n^{-1/k}$ shrink as a_n grows, so a_n grows sublinearly; guess $a_n \approx Cn^\alpha$. Then $a_{n+1} - a_n \approx C\alpha n^{\alpha-1}$ and $a_n^{-1/k} \approx C^{-1/k} n^{-\alpha/k}$. Matching exponents, $\alpha - 1 = -\alpha/k$, so $\alpha = k/(k+1)$; matching constants, $C\alpha = C^{-1/k}$, so $C^{(k+1)/k} = (k+1)/k$. Then $a_n^{k+1}/n^k \approx C^{k+1} = ((k+1)/k)^k$. The answer is now known.

The trouble starts when you try to prove it. The natural next step is to squeeze a_n between two sequences of the form $C_\pm n^{k/(k+1)}$ by induction: assume $a_n \leq C_+ n^{k/(k+1)}$ and try to show $a_{n+1} \leq C_+ (n+1)^{k/(k+1)}$. But $a_{n+1} = a_n + a_n^{-1/k}$ is *increasing* in a_n only when a_n is large enough (its derivative is $1 - \frac{1}{k} a_n^{-(k+1)/k}$), the inductive step needs an expansion of $(n+1)^{k/(k+1)}$ to second order, the constants $C_\pm$ must be pushed toward C in a limiting argument, and the starting value a_0 is arbitrary, so the induction must begin at some large N with its own bounds. Kedlaya's third solution does a version of this, tracking $b_n = a_n - Ln^{k/(k+1)}$, and it runs to a page of convexity estimates. It is correct; under exam conditions it is where a B6 gets lost.

The model chain. *Step 1: $a_n \to \infty$.* Every increment $a_n^{-1/k}$ is positive, so (a_n) is increasing. If it were bounded above by some M, every increment would be at least $M^{-1/k} > 0$, and then $a_n \geq a_0 + nM^{-1/k} \to \infty$, a contradiction. So $a_n \to \infty$.

Step 2: the right quantity. Put $m = (k+1)/k$, so $1 < m \le 3/2$ since $k \ge 2$, and set $b_n = a_n^m$. Note the target is

$$\frac{a_n^{k+1}}{n^k} = \left(\frac{a_n^{(k+1)/k}}{n}\right)^k = \left(\frac{b_n}{n}\right)^k,$$

so it suffices to find $\lim b_n/n$.

Step 3: the increments of b_n tend to m. Write $a_{n+1} = a_n\big(1 + a_n^{-1-1/k}\big) = a_n(1+t_n)$ with $t_n = a_n^{-m} > 0$. Then

$$b_{n+1} - b_n = a_n^m\big((1+t_n)^m - 1\big) = \frac{(1+t_n)^m - 1}{t_n}.$$

For $t > 0$ and $1 < m < 2$, Taylor's theorem gives $(1+t)^m = 1 + mt + \frac{1}{2}m(m-1)(1+\xi)^{m-2}t^2$ for some $\xi \in (0,t)$, and $0 < (1+\xi)^{m-2} \le 1$ because $m - 2 < 0$. Hence

$$m \;\le\; b_{n+1} - b_n \;\le\; m + \tfrac{1}{2}m(m-1)\,t_n.$$

Since $a_n \to \infty$, $t_n \to 0$, so $b_{n+1} - b_n \to m$.

Step 4: Cesàro. Let $d_i = b_{i+1} - b_i$, so $b_n = b_0 + \sum_{i=0}^{n-1} d_i$ and

$$\frac{b_n}{n} = \frac{b_0}{n} + \frac{1}{n}\sum_{i=0}^{n-1} d_i.$$

The first term tends to 0. For the second, given $\varepsilon > 0$ choose N with $|d_i - m| < \varepsilon$ for $i \ge N$; then

$$\left|\frac{1}{n}\sum_{i=0}^{n-1} d_i - m\right| \le \frac{1}{n}\sum_{i=0}^{N-1} |d_i - m| + \frac{n-N}{n}\,\varepsilon,$$

and the first term on the right is a fixed number divided by n, so for n large the whole is below 2ε. Thus $b_n/n \to m$.

Step 5. By continuity of $u \mapsto u^k$, $a_n^{k+1}/n^k = (b_n/n)^k \to m^k = \left(\frac{k+1}{k}\right)^k$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains agree on the heuristic: a_n grows like $n^{k/(k+1)}$. They split at the moment of making it rigorous. The natural attempt keeps working with a_n itself and tries to trap it between $C_\pm n^{k/(k+1)}$. The model changes variable to $b_n = a_n^{(k+1)/k}$, the power of a_n that grows *linearly*. After that, the whole problem is "increments tend to a constant, so the average does too".

> *The limit I am asked for is $(a_n^{(k+1)/k}/n)^k$. The problem is telling me that $a_n^{(k+1)/k}$ is the thing that grows like n. A sequence that grows like n is best understood through its differences: I should compute $a_{n+1}^{(k+1)/k} - a_n^{(k+1)/k}$ and hope it tends to a constant.*

> *Replace the recursion by the differential equation $y' = y^{-1/k}$. Separating, $y^{1/k}\,dy = dt$, so $\frac{k}{k+1}y^{(k+1)/k} = t + c$. The quantity with constant derivative is $y^{(k+1)/k}$. So in the discrete problem, track $a_n^{(k+1)/k}$.*

> *Once the increments of b_n tend to a constant, I do not need any rate — Cesàro averaging turns "differences converge" into "b_n/n converges" with no further estimates. So I only need $t_n \to 0$, which is $a_n \to \infty$.*

Layer 3 THE WHY

Why is $a_n^{(k+1)/k}$ the natural quantity and not a guess? There are two independent reasons, and either suffices.

The first is in the problem's own phrasing. You are asked for $\lim a_n^{k+1}/n^k$. The only way that ratio has a finite nonzero limit is if a_n^{k+1} and n^k grow at the same rate, i.e. if $a_n^{(k+1)/k}$ grows like n. The examiners have written the conjugating power into the question. Reading the target expression as $(\,\cdot\,/n)^k$ is the first move, and it costs nothing.

The second reason goes to the axioms of what a recursion like $a_{n+1} - a_n = g(a_n)$ *is*: a discrete version of $y' = g(y)$ (Stanley's heuristic, quoted in Kedlaya's remarks). An autonomous first-order ODE is solved by separation, which means finding G with $G'(y) = 1/g(y)$; then $\frac{d}{dt}G(y(t)) = G'(y)y' = 1$, so $G(y)$ grows exactly linearly. For $g(y) = y^{-1/k}$, $G(y) = \frac{k}{k+1}y^{(k+1)/k}$. So the quantity b_n is not pulled from a hat; it is the discrete shadow of the separation-of-variables substitution, the variable in which the dynamics become "add a constant each step". The discrete version is not exact — that is what the Taylor remainder in Step 3 measures — but the error is $O(t_n)$ and $t_n \to 0$.

Why then does the natural attempt struggle? Because a_n is the *wrong coordinate*: in it, the step size depends on the current state, and every inductive bound has to control how errors in a_n propagate through the nonlinear map $a \mapsto a + a^{-1/k}$. In the coordinate b_n, the step is essentially constant, and errors do not propagate at all — they add. Cesàro averaging is exactly the tool for "a sum of terms each close to m".

Kedlaya's first solution does this in three lines, beginning "Taking $m = (k+1)/k$ and $x = a_{n+1}/a_n$". It does not say why that m. Here is where it comes from: the target expression, or equivalently the separated ODE. (His display writes $x = a_{n+1}/a_n = 1 + a_n^{-(k+1)/k}$ and then applies the bound to $(1+x)^m$; the variable in the Taylor bound should be $a_n^{-(k+1)/k}$, i.e. $x - 1$, which is what Step 3 above uses.) Note also that the argument never used $k > 1$ in an essential way: for $k = 1$, $m = 2$ and $(1+t)^2 - 1 = 2t + t^2$ exactly, so the same proof gives the limit 2. The hypothesis $k > 1$ only makes $m < 2$, where the Taylor remainder bound in the form used is valid.

This is hard for a B6 in only one respect: choosing the coordinate. Everything after is A-level Taylor expansion and a standard averaging lemma. That is typical of recursion asymptotics, and it is why the choice deserves a rule rather than admiration.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the target a_n^{k+1}/n^k. On seeing it, write $a_n^{k+1}/n^k = (a_n^{(k+1)/k}/n)^k$ and define $b_n = a_n^{(k+1)/k}$.

2. That suggests the method: show $b_{n+1} - b_n$ converges to a constant m, then conclude $b_n/n \to m$ by Cesàro (Stolz–Cesàro with denominator n).

3. Hesitation occurs when computing $b_{n+1} - b_n$: the expression $(a_n + a_n^{-1/k})^{(k+1)/k} - a_n^{(k+1)/k}$ looks messy. It is reinforced away by factoring out a_n first, $a_{n+1} = a_n(1+t_n)$, so the difference becomes $((1+t_n)^m - 1)/t_n$, a function of one small quantity.

4. The trap is working with a_n directly and trying to sandwich it by $C_\pm n^{k/(k+1)}$ via induction, which requires monotonicity of the step map, second-order expansions of $(n+1)^{k/(k+1)}$, and a limiting argument in the constants. A second trap is forgetting to prove $a_n \to \infty$, on which $t_n \to 0$ depends.

5. Recognise faster: any recursion $a_{n+1} = a_n + g(a_n)$ with g slowly varying is a discretised ODE $y' = g(y)$; separate variables to find the G with $G(a_n)$ approximately linear, and track $G(a_n)$.

6. Reusable rule: to find the growth of a recursively defined sequence, change variable to the quantity whose increments converge (read it off the target expression or from separating the associated ODE), prove the increments converge, and average.

- ▷ **Cue:** a limit of the form a_n^p/n^q is requested $\longrightarrow$ **Move:** rewrite as $(a_n^{p/q}/n)^q$ and study the increments of $a_n^{p/q}$
- ▷ **Cue:** a recursion $a_{n+1} = a_n + g(a_n)$ with g small for large a_n $\longrightarrow$ **Move:** compare with $y' = g(y)$; separate variables to get G with $G' = 1/g$; the sequence $G(a_n)$ has increments tending to 1
- ▷ **Cue:** the increments of a sequence tend to a limit L $\longrightarrow$ **Move:** conclude $b_n/n \to L$ by Cesàro averaging, needing no rate of convergence
- ▷ **Cue:** a difference like $(a+h)^m - a^m$ with $h/a \to 0$ $\longrightarrow$ **Move:** factor out a^m and expand $(1+t)^m$ with Taylor remainder in the single small variable $t = h/a$