DAWN PUTNAM SESSION

Putnam 1972 — Papers A and B

Putnam 120 Archive

Contents

Putnam 1972 — Paper A **5**
1972 A1 5
1972 A2 7
1972 A3 9
1972 A4 12
1972 A5 15
1972 A6 16

Putnam 1972 — Paper B **21**
1972 B1 21
1972 B2 23
1972 B3 25
1972 B4 27
1972 B5 29
1972 B6 32

Putnam 1972 — Paper A

1972 A1

1972 A1. Show that there are no four consecutive binomial coefficients $\binom{n}{r}$, $\binom{n}{r+1}$, $\binom{n}{r+2}$, $\binom{n}{r+3}$ (n, r integers > 0 and $r+3 \leq n$) which are in arithmetic progression.

Fields: combinatorics, algebra of binomial coefficients **Position:** A1 (opener) **Answer:** No such quadruple exists: forcing the first and last triples to each be in AP simultaneously pins $n = 2r+3$, which makes the four terms the four middle entries of an odd row – and unimodality with a repeated maximum at the centre then rules an AP out.

Layer 1 THE WHAT — surface analysis

The natural attempt. Write the arithmetic-progression conditions directly: $2\binom{n}{r+1} = \binom{n}{r} + \binom{n}{r+2}$ and $2\binom{n}{r+2} = \binom{n}{r+1} + \binom{n}{r+3}$. Clear every factorial using $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, divide through by the largest common factor you can spot, and you are left with two polynomial equations in n and r. Nothing stops you from grinding through this: multiply out, collect powers of n and r, and try to eliminate n between the two equations to get a single equation in r alone. This is possible in principle – both equations are genuinely polynomial – but each one is already a relation between n and r of mixed degree once cleared of denominators, and eliminating a variable between two such relations means resultants or brute substitution, either of which produces an unpleasant quartic. You can push through it, but there is no natural stopping point telling you the algebra is nearly finished, and a single slip in expanding $(n-r-1)(n-r)$ or $(r+1)(r+2)$ is invisible until the final answer refuses to simplify.

The model chain. Take the first AP condition alone: $2\binom{n}{r+1} = \binom{n}{r} + \binom{n}{r+2}$. Divide every term by $\binom{n}{r+1}$, using the standard ratios $\binom{n}{r}/\binom{n}{r+1} = \frac{r+1}{n-r}$ and $\binom{n}{r+2}/\binom{n}{r+1} = \frac{n-r-1}{r+2}$. This gives a single equation purely in n and r, with no factorials left:

$$2 = \frac{r+1}{n-r} + \frac{n-r-1}{r+2}. \tag{$*$}$$

Now look at the second AP condition, on $\binom{n}{r+1}, \binom{n}{r+2}, \binom{n}{r+3}$: dividing through the same way produces exactly the same equation $(*)$, with r replaced by $r+1$ throughout and nothing else changed. So both r and $r+1$ satisfy $(*)$, viewed as an equation in one variable (call it x) for the fixed value of n. Clearing denominators, $(*)$ is a quadratic in x, so it has at most two roots. Moreover $(*)$ is invariant under the substitution $x \mapsto n-x-2$ – swap the two fractions and this is immediate, since $n-(n-x-2) = x+2$ and $(n-x-2)+1 = n-x-1$. So if $x = r$ is a root, so is $x = n-r-2$; the quadratic's two roots are $\{r, n-r-2\}$. Since $r+1$ is also a root and $r+1 \neq r$, the only possibility is $r+1 = n-r-2$, i.e. $n = 2r+3$. Substituting back, the four coefficients are $\binom{2r+3}{r}, \binom{2r+3}{r+1}, \binom{2r+3}{r+2}, \binom{2r+3}{r+3}$ – the four middle entries of an odd-length row. Because $2r+3$ is odd, the two central entries coincide: $\binom{2r+3}{r+1} = \binom{2r+3}{r+2}$ (symmetry $\binom{n}{k} = \binom{n}{n-k}$ with $k = r+1$). Binomial coefficients strictly increase up to the centre (the ratio $\binom{n}{k+1}/\binom{n}{k} = \frac{n-k}{k+1} > 1$ exactly while $k < \frac{n-1}{2}$), so $\binom{2r+3}{r} < \binom{2r+3}{r+1}$ strictly, while the

middle difference $\binom{2r+3}{r+2} - \binom{2r+3}{r+1} = 0$. A genuine AP needs a single constant difference across all three gaps; here one gap is 0 and an adjacent one is strictly positive, so no AP is possible. No quadruple exists.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths start by writing down the two AP equations on the four coefficients. They split at the moment of deciding what to do with the two resulting relations. The natural attempt treats them as two independent polynomial constraints on n, r to be solved simultaneously by elimination. The model treats the second relation not as new information, but as the *same* relation as the first, evaluated at a shifted argument.

> *These two equations are not independent – the second one is exactly the first one with r replaced by $r+1$, because "three consecutive binomial coefficients are in AP" is one fixed condition on a starting index, and I am imposing it twice at two adjacent starting points. So instead of eliminating between two different equations, I should ask what it means for one fixed equation to hold at two different inputs.*

Layer 3 THE WHY

The reason the second equation is free – costing no new algebra – is that dividing a three-term AP condition on $\binom{n}{k}, \binom{n}{k+1}, \binom{n}{k+2}$ by the middle term turns it into a relation between two consecutive *ratios* of binomial coefficients, and that ratio $\binom{n}{j+1}/\binom{n}{j} = \frac{n-j}{j+1}$ is a fixed rational function of the index j, independent of which window of the row you are looking at. So "three consecutive entries in AP, starting at index k" is literally one equation $(*)$ in the single unknown k (with n a fixed parameter) – and asking for it to hold at $k = r$ and again at $k = r+1$ is asking one bounded-degree equation to have two prescribed roots, which is a far stronger and more rigid demand than two unrelated polynomial constraints in two unknowns. A quadratic cannot have three roots, so once you know it needs r and $r+1$ as roots, its *other* structural feature – the reflection symmetry $k \mapsto n-k-2$ inherited from $\binom{n}{k} = \binom{n}{n-k}$ – immediately hands you the second root for free, without solving anything: it must be $n-r-2$. Setting $r+1$ equal to that forced second root is not solving a quadratic, it is reading off which of the two roots $r+1$ has to be. This is the same instinct as STYLE.md's circle example: a circle is the set of points equidistant from its centre, so the centre is obviously the object of interest; here, a ratio-normalised AP condition on binomial coefficients is obviously symmetric about $n/2$, because the coefficients themselves are, so the two roots of $(*)$ are obviously a mirror pair before you compute anything. The final contradiction is equally structural, not computational: forcing $n = 2r+3$ does not produce some exotic new case to check by hand, it lands you exactly back on the defining shape of a binomial row – unimodal, symmetric, strictly increasing until the centre – and a four-term AP straddling a symmetric peak is asking for a strictly positive difference and a zero difference to be equal, which is a contradiction built into what "middle of a row" already means.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that the problem gives *two overlapping* AP conditions (on $r, r+1, r+2$ and on $r+1, r+2, r+3$) rather than one. On seeing two conditions that share three of four terms, write down: these are probably the same relation evaluated twice, not two different relations.
2. That cue suggests normalising the AP condition by the middle term to turn it into a relation purely in the ratio $\binom{n}{k+1}/\binom{n}{k} = \frac{n-k}{k+1}$, since that isolates the part of the condition that depends only on the starting index k, exposing the repetition.
3. Hesitation sets in right after clearing denominators, when the natural instinct is to treat the resulting quadratic and its "$r+1$" shifted version as two separate equations to eliminate between. It is reinforced away by noticing the equation is a quadratic in one variable: a

quadratic has only two roots, so demanding it hold at both r and $r+1$ is a much sharper constraint than two generic equations in two unknowns, and is worth exploiting before reaching for elimination.

4. The trap is committing to full factorial expansion and cross-elimination before checking whether the two conditions are secretly the same equation. Once inside a two-variable elimination, there is no natural signal that you are close to finished, and a routine algebra slip (a wrong sign expanding $(n-r-1)(n-r)$) produces an unrecoverable mess with no way to sanity-check it.

5. To recognise this faster: whenever a problem imposes the same-shaped condition at two overlapping windows (three-term AP at index r, then again at index $r+1$), treat it as one equation sampled twice, and look immediately for a symmetry of that equation (here, $k \mapsto n-k-2$, inherited from $\binom{n}{k} = \binom{n}{n-k}$) that hands you the second root without solving anything.

6. Decision rule: when a family of objects indexed by k has both (a) a known reflection symmetry in k and (b) a problem condition imposed at two consecutive values of k, normalise the condition into a single equation in k first – the symmetry will usually supply one root as the mirror of the other, collapsing what looks like a two-variable elimination into a one-line consequence.

1972 A2

1972 A2. Let S be a set and let $*$ be a binary operation on S satisfying the laws $x*(x*y) = y$ for all x, y in S, and $(y*x)*x = y$ for all x, y in S. Show that $*$ is commutative but not necessarily associative.

Fields: abstract algebra, functional equations **Position:** A2 **Answer:** $*$ is commutative (proof below); it need not be associative – $x*y = -x-y$ on $\mathbb{Z}$ satisfies both laws and is not associative.

Caveat. The MAA 1965–1984 OCR renders the internal derivation of commutativity as near-unreadable line noise ("(x*y)*x = (x*y)[(x*y)*y] = y", with no visible equals signs or substitution markers in several places). The two axioms and the final claims (commutative, not associative, with the counterexample $x*y = -x-y$) are legible and unambiguous, and I have independently re-derived the intermediate identity and the commutativity argument from the two given axioms alone; the derivation given in Layer 1 below is my own reconstruction, verified step by step, not a transcription of the garbled OCR.

Layer 1 THE WHAT — surface analysis

The natural attempt. Read the two laws as $x*(x*y) = y$ and $(y*x)*x = y$, and try to reach $x*y = y*x$ by direct substitution games: plug $x*y$ in for y somewhere, plug $y*x$ in for x somewhere, and see what falls out. This does produce true statements – for instance, substituting $y \to x*y$ into the first law gives $x*(x*(x*y)) = x*y$, which after using the first law again on the inner pair collapses right back to $x*y = x*y$, telling you nothing new. You can spend a long time generating true-but-useless identities this way, because each law individually only ever cancels its own operation against itself: the first law says "operating by x on the left twice is the identity," the second says "operating by x on the right twice is the identity." Neither law alone ever brings a left-application and a right-application into the same equation, and $x*y = y*x$ is precisely a claim about a left-application $(x*y)$ equalling a right-application $(y*x)$. Without deliberately manufacturing an equation that mixes both, the natural substitutions stay trapped inside one law at a time and never connect the two sides you are trying to equate.

The model chain. First isolate an intermediate identity that does not itself look like the goal: $(x * y) * x = y$ for all x, y. To get it, start from the fact that the second law, applied to the pair (x, y) in place of (y, x), reads $(x * y) * y = x$ – call this fact (2′). Now take $(x * y) * x$ and, inside it, replace the trailing x using (2′) read backwards: $x = (x * y) * y$, so

$$(x * y) * x = (x * y) * \big[(x * y) * y\big].$$

The right-hand side has the exact shape of the first law, $a * (a * b)$, with $a = x * y$ and $b = y$; the first law collapses it straight to $b = y$. Hence

$$(x * y) * x = y. \tag{3}$$

Now bring in the second law itself, but apply it to the specific element $x * y$ rather than to a generic pair: the second law says $(Y * X) * X = Y$ for all Y, X; set $Y = x * y$, $X = x$:

$$\big((x * y) * x\big) * x = x * y.$$

But the left-hand inner block $(x * y) * x$ is exactly y, by (3). Substituting,

$$y * x = x * y,$$

which is commutativity. For non-associativity, take $S = \mathbb{Z}$ and define $x * y = -x - y$. Check the laws: $x * (x * y) = x * (-x - y) = -x - (-x - y) = y$, and $(y * x) * x = (-y - x) * x = -(-y - x) - x = y$; both hold. But $x * (y * z) = -x - (-y - z) = -x + y + z$ while $(x * y) * z = -(-x - y) - z = x + y - z$; taking $x = 1, y = 0, z = 2$ gives $-1 + 0 + 2 = 1$ against $1 + 0 - 2 = -1$, so the two are unequal and $*$ is not associative.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths begin by trying substitutions into the two given laws. They split the moment the natural attempt runs out of new information from single-law substitution and the model deliberately builds a *third*, unstated identity – $(x * y) * x = y$ – that is not literally either axiom, by feeding one axiom's output into the other axiom's input.

> *Each law by itself only ever cancels an operation against a repetition of the same operation on the same side, so no amount of substituting within one law will ever put a left-multiplication and a right-multiplication into the same equation. I need to manufacture a hybrid identity myself – take what the second law tells me equals x, and feed that expression, not the bare letter x, into the first law.*

Layer 3 THE WHY

The two axioms are statements about two different operators, $L_x(z) = x * z$ and $R_x(z) = z * x$, and each axiom says its own operator is an involution: $L_x \circ L_x = \mathrm{id}$ from the first law, $R_x \circ R_x = \mathrm{id}$ from the second. Commutativity, $x * y = y * x$, is the statement $L_x(y) = R_x(y)$ – it is inherently a claim that *relates* the two operators, so it cannot be a consequence of either involution property in isolation; an equation purely about $L_x \circ L_x$ can never, by substitution alone, produce a statement equating L_x and R_x, because there is nothing in that equation for R_x to attach to. The only way to force a relation between L_x and R_x is to compose them – apply one, then feed the result into the other – which is exactly what building $(x * y) * x = y$ does: it takes the output of L_x (namely $x * y$) and asks what R_x does to it, using the second axiom (rewritten as $x = (x * y) * y$, i.e. R_y undoing L_x's partner) to smuggle a fresh copy of x in through the first axiom's slot. Once that hybrid identity (3) exists, applying the second axiom once more to the single element $x * y$ finishes the job for free, because (3) has already told you what $(x * y) * x$ equals, and the second axiom tells you independently what $\big((x * y) * x\big) * x$

equals – two descriptions of the same quantity, one of which is literally $y * x$ once you strip the outer layer. This is the general lesson behind any "show two differently-built expressions are equal" problem defined by closure axioms: the axioms as given never mix left and right, so the solver's job is to construct, by hand, the one composite expression that both operators can act on simultaneously, and let each axiom peel off one of its own layers.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that the two given laws are structurally symmetric to each other under swapping the roles of left- and right-multiplication ($x * (x * y) = y$ versus $(y * x) * x = y$), yet the goal, $x * y = y * x$, is a statement that crosses between them. On seeing that, write down: neither law alone can produce the goal; I need an identity that uses both.
2. That cue suggests deliberately constructing a hybrid expression – take an output guaranteed by one law and substitute it into the other law's argument slot, rather than substituting variables (like $y \to x * y$) into a single law repeatedly.
3. The hesitation is exactly the moment after a few substitutions collapse back to trivial identities ($x * y = x * y$): it feels like the axioms are too weak, when in fact the substitutions were simply staying inside one axiom. It is reinforced away by naming explicitly what kind of statement the goal is (a cross-operator equation) and asking which construction could possibly produce that shape.
4. The trap is treating $x * (x * y) = y$ and $(y * x) * x = y$ as interchangeable "cancellation" facts and hoping enough substitution will eventually equate $x * y$ and $y * x$ by accident. Purely single-law substitution can only ever regenerate statements of the form $a = a$; it cannot manufacture a genuinely new equation relating two differently-parenthesised expressions.
5. To recognise this faster: whenever two axioms each describe a separate operator as self-cancelling (an involution), and the goal is to relate the two operators, expect the proof to require exactly one composite step – feed the output of one operator into the other – and look for it immediately rather than iterating substitutions inside a single axiom.
6. Decision rule: given closure axioms of the shape "operator A undoes itself" and "operator B undoes itself" separately, and a goal equating A and B, construct the hybrid expression $B(A(\cdot))$ (or $A(B(\cdot))$) explicitly and expand it using both axioms in turn, rather than searching for the goal inside iterated applications of a single axiom.

1972 A3

1972 A3. If for a sequence $x_1, x_2, x_3, \ldots$, $\lim_{n\to\infty}(x_1 + x_2 + \cdots + x_n)/n$ exists, call this limit the C-limit of the sequence. A function $f(x)$ from $[0,1]$ to the reals is called a supercontinuous function on the interval $[0,1]$ if the C-limit exists for the sequence $f(x_1), f(x_2), f(x_3), \ldots$ whenever the C-limit exists for the sequence $x_1, x_2, x_3, \ldots$. Find all supercontinuous functions on $[0,1]$.

Fields: real analysis, Cesàro summability, functional equations **Position:** A3 **Answer:** The supercontinuous functions on $[0,1]$ are exactly the affine functions $f(x) = Ax + B$.

Caveat. The official solution proves sufficiency (affine $\Rightarrow$ supercontinuous) trivially, and for necessity states the two required lemmas – (1) f is continuous, (2) $f\left(\frac{a+b}{2}\right) = \frac{f(a)+f(b)}{2}$ for all a, b – but proves only (2) in the source, describing (1) as "similar" without giving it. The proof of (1) below is my own construction, using the same block-domination device the book uses for (2); I built and checked it independently rather than transcribing anything, and flag it as reconstructed rather than sourced. The final step, that continuity plus the midpoint identity forces f affine, is likewise supplied by me (a standard dyadic-density

argument) since the book states only that the two facts "imply that f is affine" without spelling out why.

Layer 1 THE WHAT — surface analysis

The natural attempt. Guess that affine functions work (easy to check: for $f(x) = Ax + B$, the average of $f(x_1), \ldots, f(x_n)$ is exactly $A \cdot \frac{x_1+\cdots+x_n}{n} + B$, so its limit exists whenever the C-limit of x_n does). Then try to prove these are the *only* such functions by attacking convergent sequences directly: take $x_n \to x$ in the ordinary sense, note this forces the C-limit of x_n to equal x (a standard fact – Cesàro averages of a convergent sequence converge to the same limit), so supercontinuity gives that the C-limit of $f(x_n)$ exists. The natural next move is to try to conclude $f(x_n) \to f(x)$ directly from "the C-limit of $f(x_n)$ exists" – but this does not follow: a sequence can have a perfectly good Cesàro average without converging at all (e.g. $1, 0, 1, 0, \ldots$ has C-limit $\frac{1}{2}$ but never converges). So the existence of the C-limit of $f(x_n)$, applied to a single ordinarily-convergent input sequence, is compatible with $f(x_n)$ oscillating forever – it gives no usable information about f from one sequence alone. You are stuck holding a true but useless deduction, the same trap as feeding a single substitution into a single axiom: one input sequence, however cleverly chosen, cannot pin down local behaviour of f, because supercontinuity is a condition quantified over *all* sequences, and you have only used one.

The model chain. To use the "for all sequences" strength of the hypothesis, build one deliberately composite sequence out of two candidate limits and let supercontinuity referee between them. For the midpoint identity, fix $a, b \in [0, 1]$, let $c = \frac{a+b}{2}$, and suppose toward contradiction $f(c) \neq \frac{f(a)+f(b)}{2}$. Choose integers $N_1 < N_2 < N_3 < \cdots$ growing so fast that $N_{i+1} > 2^i N_i$, and build $\{x_n\}$ in blocks: for $N_{2i-1} \leq n < N_{2i}$, set $x_n = c$ (constant); for $N_{2i} \leq n < N_{2i+1}$, set x_n to alternate $a, b, a, b, \ldots$. Because each new block's length dwarfs the sum of every block before it (the ratio condition $N_{i+1} > 2^i N_i$ makes earlier blocks' total length a vanishing fraction of the current one), the running Cesàro average of x_n at the end of block i is determined almost entirely by that block's own local average. Both block types have local average c: a constant-c block trivially averages to c, and an alternating a, b block of even length averages to $\frac{a+b}{2} = c$ as well. So the overall running average is squeezed toward c at the end of every block, and $\{x_n\}$ has C-limit exactly c. Now look at $\{f(x_n)\}$: on a c-block it is constant $f(c)$, with local average $f(c)$; on an a, b-block it alternates $f(a), f(b), f(a), f(b), \ldots$, with local average $\frac{f(a)+f(b)}{2}$. By the same domination argument, the running average of $f(x_n)$ is pulled to (a value near) $f(c)$ at the end of every c-block and to (a value near) $\frac{f(a)+f(b)}{2}$ at the end of every ab-block. If these two numbers differ, the partial averages of $f(x_n)$ have two subsequences converging to different limits, so the C-limit of $\{f(x_n)\}$ does not exist – contradicting supercontinuity. Hence $f(c) = \frac{f(a)+f(b)}{2}$ for all $a, b \in [0, 1]$.

The same composite-sequence device gives continuity. Suppose f is discontinuous at some $x \in [0, 1]$: there is $\varepsilon > 0$ and points $y_k \to x$ (ordinary convergence) with $|f(y_k) - f(x)| \geq \varepsilon$ for every k. Build a sequence $\{w_n\}$ in rapidly-growing blocks exactly as above, alternating long runs of the constant x with runs of a single fixed nearby point y_k (taking k large enough, block by block, that y_k is within $1/i$ of x on the i-th such block, so that the terms w_n do converge to x in the ordinary sense overall). An ordinarily-convergent sequence automatically has C-limit equal to its ordinary limit, so $\{w_n\}$ has C-limit x. But by the same block-domination argument, the running average of $f(w_n)$ is pulled to $f(x)$ at the end of every x-block and to (approximately) $f(y_k)$, which stays at least $\varepsilon/2$ away from $f(x)$ for k large, at the end of every y_k-block. So the partial averages of $f(w_n)$ again split into two subsequences with different limits, and the C-limit of $f(w_n)$ fails to exist – contradicting supercontinuity. Hence f is continuous at every $x \in [0, 1]$.

Finally, continuity plus $f\left(\frac{a+b}{2}\right) = \frac{f(a)+f(b)}{2}$ forces f to be affine. Let $g(x) = f(x) - f(0) - x\big(f(1) - f(0)\big)$; then g is continuous, $g(0) = g(1) = 0$, and g inherits the exact midpoint identity

(the affine correction term satisfies it with equality on the nose, so subtracting it preserves the identity for g). By the midpoint identity, $g(\frac{1}{2}) = \frac{g(0)+g(1)}{2} = 0$; then $g(\frac{1}{4}) = g(\frac{3}{4}) = 0$ from the midpoints of $(0, \frac{1}{2})$ and $(\frac{1}{2}, 1)$; inductively g vanishes at every dyadic rational in $[0, 1]$. The dyadic rationals are dense in $[0, 1]$, and g is continuous, so $g \equiv 0$ on $[0, 1]$. Hence $f(x) = f(0) + x\big(f(1) - f(0)\big)$, which is affine, as claimed.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths begin by testing the hypothesis on an ordinarily-convergent sequence $x_n \to x$, which legitimately gives that the C-limit of $f(x_n)$ exists. They split at what to do with that one deduction: the natural attempt tries to extract continuity from it directly and finds the deduction too weak (a Cesàro-summable sequence need not converge), while the model abandons single, "clean" input sequences and deliberately manufactures a two-valued composite sequence whose blocks pit two competing candidate limits against each other.

> *Supercontinuity is a condition over* all *sequences with a C-limit, so if one clean convergent sequence gives me too little information, the fix is not to analyse it harder – it is to build a genuinely different input, one made of two clashing pieces, so that if f does not behave consistently between them, the hypothesis itself gets violated and hands me the contradiction I want.*

Layer 3 THE WHY

The single deduction from an ordinary convergent sequence is weak because "the C-limit of $f(x_n)$ exists" is a statement about long-run averages, and averages are blind to oscillation as long as the oscillation itself averages out consistently – exactly the $1, 0, 1, 0, \ldots$ phenomenon. To force f to reveal itself, you need an input sequence whose Cesàro average has a definite, provable value (so the hypothesis actually applies) while its structure still lets you read off two separate local behaviours of f at the end of two different kinds of block. The rapidly-growing block-length condition $N_{i+1} > 2^i N_i$ is the engine that makes both of these possible simultaneously: it guarantees that at the precise moment you sample the running average (the end of a block), that average is dominated by the block just completed, no matter how earlier blocks behaved, because their combined length is forced to be a vanishing fraction of the current block. This is what turns "the whole sequence has C-limit c" and "the running average is repeatedly pulled toward two different candidate values under f" from contradictory-sounding claims into simultaneously true ones: the C-limit of x_n genuinely exists (both block types individually average to c, so there is no real ambiguity about where $\{x_n\}$'s average is heading), while $\{f(x_n)\}$'s running average is repeatedly re-anchored near two different numbers, $f(c)$ and $\frac{f(a)+f(b)}{2}$, forever – which is precisely what "no limit exists" looks like for a sequence of partial averages. The midpoint identity is then not an arbitrary lemma to memorise: it is the shape forced on any function for which averaging never lies about it, and continuity plus the midpoint identity pins a function down completely on a dense set (the dyadics) by the same halving structure the block construction already exploited, after which continuity alone extends equality from a dense set to the whole interval. The construction and the conclusion use the same idea twice: dyadic self-similarity is what both proves the lemma and then promotes the lemma to the full answer.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that the hypothesis is universally quantified over sequences ("whenever the C-limit exists"), while a single convergent test sequence $x_n \to x$ only yields one true-but-weak consequence. On seeing a hypothesis quantified over all objects of a type, and finding one instance too weak, write down: build a composite instance instead.

2. That cue suggests constructing a sequence deliberately assembled from two competing pieces (here, a constant block and an alternating block, both engineered to have the same target

average), so that supercontinuity's single conclusion has to reconcile two different local behaviours of f at once.

3. The hesitation is believing that "the C-limit of $f(x_n)$ exists" for one convergent $x_n \to x$ should already force $f(x_n) \to f(x)$. It is reinforced away by the concrete counterexample $1, 0, 1, 0, \ldots$: Cesàro-summable does not imply convergent, so that single deduction genuinely carries no information about pointwise behaviour of f and must be abandoned in favour of a richer input.

4. The trap is spending effort trying to squeeze continuity or the midpoint identity out of one clean sequence, when the hypothesis's real power only appears when the input sequence itself is built to expose a potential inconsistency – the growth condition on block lengths is not decoration, it is what makes the running average well-defined in one interpretation (the x_n sequence) while remaining genuinely torn in another (the $f(x_n)$ sequence).

5. To recognise this faster: whenever a hypothesis is stated as "property P holds for every object satisfying condition Q" and a single natural test object under-determines the conclusion, look for a way to interleave two test objects, in blocks whose lengths grow fast enough that the interleaved object still satisfies Q unambiguously.

6. Decision rule: to disprove an equality (or force one) using an averaging-type hypothesis quantified over all sequences, construct a two-phase sequence with rapidly growing block lengths so that (a) the overall Cesàro average is provably a single value, forcing the hypothesis to apply, while (b) the two block types, sampled at block ends, pull the transformed sequence's running average toward two different values – turning "the hypothesis must hold" into the equality you wanted.

1972 A4

1972 A4. Of all ellipses inscribed in a square, show that the circle has the maximum perimeter.

Fields: geometry, calculus (integral inequalities), analytic geometry **Position:** A4 **Answer:** Every ellipse inscribed in a square of side $2R$ has semi-axes a, b with $a^2 + b^2 = 2R^2$, and then its perimeter is at most $2\pi R$, with equality only for $a = b = R$, the inscribed circle.

Layer 1 THE WHAT — surface analysis

The natural attempt. Set the square as $|x| \le R$, $|y| \le R$. The first job is to describe the ellipses that fit. A strong student writes the general conic $Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$ with $B^2 - 4AC < 0$ and imposes tangency to each of the four lines $x = \pm R$, $y = \pm R$: substitute $x = R$, get a quadratic in y, demand zero discriminant, and repeat four times. That is exactly what the book's solution does, and it works: subtracting the conditions in pairs forces $D = E = 0$ and then $A = C$, so the ellipse is centred at the centre of the square with its axes along the diagonals. Then, with semi-axes $a \ge b$ along the diagonals, tangency to $x + y = R\sqrt{2}$ (in rotated coordinates) is another discriminant computation, yielding $a^2 + b^2 = 2R^2$.

That is already a page of algebra, and the actual question has not started. Now you must compare perimeters. The perimeter of $x = a\cos t$, $y = b\sin t$ is

$$P(a, b) = \int_0^{2\pi} \sqrt{a^2 \sin^2 t + b^2 \cos^2 t}\, dt,$$

an elliptic integral with no closed form. The reflex is Lagrange multipliers, or substituting $b = \sqrt{2R^2 - a^2}$ and differentiating under the integral sign in a. You get

$$\frac{dP}{da} = \int_0^{2\pi} \frac{a\sin^2 t - a\cos^2 t}{\sqrt{a^2 \sin^2 t + (2R^2 - a^2)\cos^2 t}}\, dt,$$

and deciding the sign of that for $a > R$ is its own small problem. It can be done (the book does it by folding $[0, \pi/2]$ in half at $\pi/4$ and showing that $u \mapsto \sqrt{p-u} + \sqrt{p+u}$ decreases), but the student who has not seen the fold stares at an integral with a sign-changing numerator and no obvious way to compare the two halves. That is the wall.

The model chain. **Step 1: what "inscribed" forces.** Take any ellipse with centre O' and semi-axes a, b, and work in its own axes, so it is $x = a\cos t$, $y = b\sin t$. For a unit vector $\mathbf{n} = (\cos\alpha, \sin\alpha)$, the largest value of $\mathbf{n}\cdot(x,y)$ on the ellipse is

$$\max_t \left(a\cos\alpha\cos t + b\sin\alpha\sin t\right) = \sqrt{a^2\cos^2\alpha + b^2\sin^2\alpha} =: h(\alpha),$$

by the $A\cos t + B\sin t \le \sqrt{A^2+B^2}$ bound, attained. The smallest value is $-h(\alpha)$. So the two tangent lines with normal $\mathbf{n}$ are at distance $h(\alpha)$ on either side of the centre.

An ellipse inscribed in the square lies inside it and touches all four sides, so each side-line is a tangent line. The two sides with normal $\mathbf{n}$ are then the two tangents with that normal; they are $2R$ apart and symmetric about O', hence $h(\alpha) = R$. The other pair of sides has normal $\mathbf{n}^\perp$, angle $\alpha + \pi/2$, so also $h(\alpha + \pi/2) = R$. Written out:

$$a^2\cos^2\alpha + b^2\sin^2\alpha = R^2, \qquad a^2\sin^2\alpha + b^2\cos^2\alpha = R^2.$$

Add them:

$$a^2 + b^2 = 2R^2. \tag{$*$}$$

I never needed to know the orientation α. (Subtracting instead gives $(a^2-b^2)\cos 2\alpha = 0$, so a non-circular inscribed ellipse does have its axes on the diagonals, but the argument below does not use this.)

Step 2: the perimeter bound. Let $g(t) = a^2\sin^2 t + b^2\cos^2 t \ge 0$, so $P = \int_0^{2\pi}\sqrt{g(t)}\,dt$. Since $\int_0^{2\pi}\sin^2 t\,dt = \int_0^{2\pi}\cos^2 t\,dt = \pi$,

$$\int_0^{2\pi} g(t)\,dt = \pi(a^2+b^2) = 2\pi R^2$$

by $(*)$. The Cauchy–Schwarz inequality for integrals, applied to $\sqrt{g}$ and 1, gives

$$P^2 = \Big(\int_0^{2\pi}\sqrt{g}\cdot 1\,dt\Big)^2 \le \int_0^{2\pi} g\,dt \cdot \int_0^{2\pi} 1\,dt = 2\pi R^2\cdot 2\pi = (2\pi R)^2.$$

So $P \le 2\pi R$, the circumference of the inscribed circle (which is an inscribed ellipse with $a = b = R$).

Step 3: equality. Equality in Cauchy–Schwarz for continuous functions holds only when $\sqrt{g}$ is a constant multiple of 1, i.e. g is constant. But $g(0) = b^2$ and $g(\pi/2) = a^2$, so $a = b$, and then $(*)$ gives $a = b = R$. The circle is the unique maximiser.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains agree that the perimeter is $\int_0^{2\pi}\sqrt{a^2\sin^2 t + b^2\cos^2 t}\,dt$ and that "inscribed" imposes a condition on a, b. They split at the moment the constraint $(*)$ is written down. The natural attempt reads $(*)$ as a relation to *substitute*: $b = \sqrt{2R^2 - a^2}$, then differentiate. The model reads $(*)$ as a statement about the *integrand*: the thing under the square root has a fixed total, $\int g = \pi(a^2+b^2)$.

> *The constraint is $a^2 + b^2$ fixed, and a^2 and b^2 are exactly what sits under the square root, linearly — so the average of the thing under the root is fixed, and the square root is concave. I don't need to vary anything; the integral of a square root is largest when what's under it doesn't vary.*

> *Before I solve for which ellipses are inscribed, what does inscribed give me in a form I can use? Two perpendicular pairs of tangents at distance R — and the tangent distance in direction α squared is $a^2\cos^2\alpha + b^2\sin^2\alpha$. Two perpendicular directions: add, and the angle disappears.*

Layer 3 THE WHY

Why is Cauchy–Schwarz (or Jensen for the concave square root) the natural tool rather than a trick? Because of what a perimeter integral *is*. The perimeter of a closed curve is $\int |\mathbf{r}'(t)|\,dt$, and $|\mathbf{r}'(t)|$ is the square root of a sum of squares. Constraints on shape — area, moment of inertia, tangent distances — almost always control *squared* quantities, because distances enter geometry through Pythagoras. So you are routinely asked to bound $\int \sqrt{\text{quantity}}$ given control of $\int$ quantity. The inequality that converts one into the other is $\left(\int \sqrt{g}\right)^2 \le \left(\int 1\right)\left(\int g\right)$, and its equality case is "g constant", which for a curve means "constant speed at every direction", which for an ellipse means circle. The answer to the problem is the equality case of the inequality you should reach for; that is the signal that you have chosen the right inequality.

Now the constraint. Why is adding two perpendicular tangent distances the right move? Because of the definition of h: $h(\alpha)^2 = a^2\cos^2\alpha + b^2\sin^2\alpha$ is a weighted average of a^2 and b^2 with weights $\cos^2\alpha, \sin^2\alpha$, and the perpendicular direction swaps the weights. Adding a direction and its perpendicular is using $\cos^2 + \sin^2 = 1$ — Pythagoras again. A square is the shape that tests an ellipse in two perpendicular directions with the same distance; so a square sees only $a^2 + b^2$, never a and b separately. That is why the family of inscribed ellipses is one-parametric, and why the orientation is irrelevant to the perimeter question.

The book's solution spends most of its length proving the axes lie along the diagonals via a general conic and four discriminants. That fact is true, and the subtraction $(a^2 - b^2)\cos 2\alpha = 0$ gives it in one line, but it is not needed at all. The book then computes $L = \int_0^{\pi/2}\left[R^2 - \frac{1}{2}c^2\cos 2t\right]^{1/2} dt$ with $c^2 = a^2 - b^2$, folds at $\pi/4$, and uses that $\sqrt{p-u} + \sqrt{p+u}$ decreases. Here is where that comes from: it is Jensen's inequality for the concave square root, applied pointwise to the pair of values $R^2 \mp \frac{1}{2}c^2\cos 2t$ whose average is R^2. The fold pairs each t with a partner so that the averages are all equal to R^2; Cauchy–Schwarz does the same pairing globally, all at once.

Layer $+\alpha$ BACKTESTING

1. The cue is "maximum perimeter" for a family of ellipses. Write down the perimeter integral $\int_0^{2\pi}\sqrt{a^2\sin^2 t + b^2\cos^2 t}\,dt$ immediately and admit it has no closed form — so the answer must come from an inequality, not a computation.

2. An integral of a square root with a constraint on squares suggests Cauchy–Schwarz / Jensen; "inscribed in a square" suggests computing tangent (support) distances in two perpendicular directions.

3. Hesitation comes at "which ellipses are inscribed?" — the pull to classify them fully (centre, orientation) before doing anything else. Reinforce with the question: what is the least I need from the constraint? Only a relation between a and b.

4. The trap is substituting $b = \sqrt{2R^2 - a^2}$ and differentiating an elliptic integral. It can be pushed through, but you will be deciding the sign of an integral with a sign-changing numerator — zombie calculus on a problem whose answer is an equality case.

5. When an extremal answer is the "most symmetric" object (circle, regular polygon, equal variables), look for an inequality whose equality case is that symmetry, before you look for a derivative.

6. Decision rule: to bound $\int\sqrt{g}$ when $\int g$ is controlled, use $\left(\int\sqrt{g}\right)^2 \le$ (length of interval) $\int g$; the extremiser is g constant. To extract a constraint from tangency in perpendicular directions, square the support distances and add.

- ▷ **Cue:** a perimeter or arc-length integral $\int \sqrt{\cdots}$ must be maximised under a constraint on squared quantities $\longrightarrow$ **Move:** bound it by Cauchy–Schwarz against 1; the extremal shape is the one with constant integrand
- ▷ **Cue:** a convex curve is tangent to two perpendicular pairs of parallel lines $\longrightarrow$ **Move:** write the squared support distance $h(\alpha)^2$ and add $h(\alpha)^2 + h(\alpha + \pi/2)^2$ to kill the orientation
- ▷ **Cue:** the problem's answer is the most symmetric member of a family $\longrightarrow$ **Move:** look for an inequality whose equality case is that symmetry rather than differentiating

1972 A5

1972 A5. Show that if n is an integer greater than 1, then n does not divide $2^n - 1$.

Fields: number theory (orders modulo a prime, Fermat's little theorem) **Position:** A5 **Answer:** No $n > 1$ divides $2^n - 1$: the smallest prime factor p of such an n would force the order of 2 modulo p to be a divisor of n lying strictly between 1 and p, contradicting the minimality of p.

Layer 1 THE WHAT — surface analysis

The natural attempt. Try small cases: $2^2 - 1 = 3$, $2^3 - 1 = 7$, $2^4 - 1 = 15$, $2^5 - 1 = 31$, $2^6 - 1 = 63$, $2^9 - 1 = 511 = 7 \cdot 73$, $2^{15} - 1 = 32767 = 7 \cdot 31 \cdot 151$. Nothing divides. Then the prime case: for a prime p, Fermat gives $2^p \equiv 2 \pmod{p}$, so $2^p - 1 \equiv 1 \pmod{p}$ and $p \nmid 2^p - 1$. Good — primes are done in one line.

Now the natural extension: take composite n and pick *some* prime $p \mid n$. You want $p \nmid 2^n - 1$. Fermat tells you $2^{p-1} \equiv 1$, so $2^n \equiv 2^{n \bmod (p-1)} \pmod{p}$. For this to be $\neq 1$ you need $n \bmod (p-1)$ to avoid multiples of the order of 2 — but n is arbitrary, so $n \bmod (p-1)$ can be anything. Take $n = 21$, $p = 7$: $21 \bmod 6 = 3$ and $2^3 = 8 \equiv 1 \pmod{7}$. So $7 \mid 2^{21} - 1$. An arbitrary prime factor is not enough; the argument with "a prime factor of n" stalls here. The student often tries induction on n next, or on the number of prime factors, and finds that $2^{ab} - 1$ factoring as $(2^a - 1)(\cdots)$ does not interact with $ab \mid \cdot$ in any usable way. That is the wall.

The model chain. Suppose, for contradiction, that $n > 1$ and $n \mid 2^n - 1$.

n **is odd.** $2^n - 1$ is odd, and n divides it.

Choose the smallest prime. Let p be the smallest prime factor of n; p is odd since n is. Then $p \mid 2^n - 1$, i.e. $2^n \equiv 1 \pmod{p}$.

The order. Let d be the least positive integer with $2^d \equiv 1 \pmod{p}$ (it exists: p is odd, so $\gcd(2, p) = 1$ and Fermat gives $2^{p-1} \equiv 1$). Standard lemma: if $2^m \equiv 1 \pmod{p}$ then $d \mid m$. Proof: write $m = qd + r$ with $0 \leq r < d$; then $1 \equiv 2^m = (2^d)^q 2^r \equiv 2^r$, and minimality of d forces $r = 0$.

Apply the lemma twice: $d \mid n$ (from $2^n \equiv 1$) and $d \mid p - 1$ (from Fermat). Also $d \neq 1$, because $2^1 - 1 = 1$ is not divisible by p. So $d \geq 2$ and d has a prime factor q.

Contradiction. $q \mid d \mid n$, so q is a prime factor of n; and $q \leq d \leq p - 1 < p$. That contradicts the choice of p as the smallest prime factor of n. Hence no $n > 1$ divides $2^n - 1$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains reach "let p be a prime factor of n; then $2^n \equiv 1 \pmod{p}$ and $2^{p-1} \equiv 1 \pmod{p}$". The natural attempt picks an *arbitrary* prime factor and tries to show $2^n \not\equiv 1$ directly, which the example $n = 21$, $p = 7$ shows is false in general. The model picks the *smallest* prime factor and never tries to show $2^n \not\equiv 1$; it instead combines the two congruences into "the order d divides both n and $p - 1$".

> *I have two exponents that both give* 1*:* n *and* $p - 1$*. Whatever the order of* 2 *is, it divides both. Now*

> *$p-1$ has only prime factors smaller than p — so if I choose p to be the* smallest *prime of n, then n and $p-1$ share no prime at all, and the order is forced to be* 1.

> *An arbitrary prime factor failed on $n=21$ because $7-1=6$ shares the factor 3 with 21. What choice of prime makes $p-1$ coprime to n? The smallest one.*

Layer 3 THE WHY

Why is "smallest prime factor" not a rabbit from a hat? Look at what the two congruences give you. $2^n \equiv 1$ and $2^{p-1} \equiv 1 \pmod p$ together say $2^{\gcd(n,\,p-1)} \equiv 1 \pmod p$ — the order divides the gcd, which is the lemma above in one line. Since $2^1 \not\equiv 1$, the whole problem reduces to showing $\gcd(n, p-1) = 1$ for some prime $p \mid n$. Now go to the definitions: $\gcd(n, p-1) > 1$ means some prime divides both n and $p-1$, and every prime dividing $p-1$ is less than p. So the gcd is 1 precisely when n has no prime factor below p — that is, when p is the least prime factor of n. The choice is not inspired; it is the solution of the equation "which p makes $\gcd(n, p-1) = 1$ automatic?". The failed attempt with $n = 21$, $p = 7$ was the evidence: $\gcd(21, 6) = 3$, and 3 is a smaller prime of n. Switching to $p = 3$: $\gcd(21, 2) = 1$, and indeed $2^{21} \equiv 2 \pmod 3$.

This is the extremal principle in number theory: when a property must hold for *some* prime factor, and the obstruction always comes from a smaller prime, choose the smallest. The book's solution does exactly this but presents the choice with no motivation. The motivation is the gcd.

It also shows what the problem is really testing, because the same argument run with a general base $a \geq 2$ proves a sharper statement: *if $n > 1$ and $n \mid a^n - 1$, then the smallest prime factor p of n divides $a-1$.* (If $p \mid a$ then $a^n \equiv 0 \not\equiv 1$; otherwise the order of a modulo p divides $\gcd(n, p-1) = 1$, so $a \equiv 1 \pmod p$. No parity step is needed: for $p = 2$, $p - 1 = 1$.) The statement is sharp — $3 \mid 4^3 - 1 = 63$ with $3 \mid 4 - 1$, and $2 \mid 3^2 - 1 = 8$ with $2 \mid 3 - 1$. For $a = 2$ we have $a - 1 = 1$, which no prime divides; that single fact is where the problem lives.

Layer $+\alpha$ BACKTESTING

1. The first cue is "n divides $2^n - 1$": an exponent tied to the modulus. Write down immediately "let p be a prime factor of n; then $2^n \equiv 1 \pmod p$, and also $2^{p-1} \equiv 1 \pmod p$."
2. Two exponents both giving 1 modulo p suggests the order of 2 modulo p, which divides each of them, hence their gcd.
3. Hesitation occurs at "which prime p?". An arbitrary prime fails (test $n = 21$). Reinforce with: the gcd with $p - 1$ can only involve primes smaller than p, so take the smallest p.
4. The trap is proving the prime case by Fermat and then trying to extend to composites by factorising $2^{ab} - 1$ or inducting on n — neither sees the interaction between n and $p - 1$.
5. Whenever a divisibility statement mixes n with $p - 1$ for primes $p \mid n$ (orders, Fermat, Euler), expect the least prime factor to be the right choice, because it makes $\gcd(n, p-1) = 1$.
6. Decision rule: if $a^m \equiv 1$ and $a^k \equiv 1 \pmod p$, then $a^{\gcd(m,k)} \equiv 1$; engineer gcd $= 1$ by an extremal choice (smallest prime factor) and conclude $a \equiv 1 \pmod p$.

▷ **Cue:** $n \mid a^n - 1$ or $a^n \equiv 1 \pmod n$ appears in a statement $\longrightarrow$ **Move:** take p the smallest prime factor of n; the order of a mod p divides $\gcd(n, p-1) = 1$
▷ **Cue:** two different exponents both give $a^{(\cdot)} \equiv 1 \pmod p$ $\longrightarrow$ **Move:** pass to the order: it divides both exponents, hence their gcd
▷ **Cue:** an argument "for some prime factor" fails because of a smaller prime factor $\longrightarrow$ **Move:** choose the prime extremally (smallest) so the obstruction cannot occur

1972 A6

1972 A6. Let $f(x)$ be an integrable function in $0 \le x \le 1$ and suppose $\int_0^1 f(x)\,dx = 0$, $\int_0^1 xf(x)\,dx = 0$, …, $\int_0^1 x^{n-1}f(x)\,dx = 0$ and $\int_0^1 x^n f(x)\,dx = 1$. Show that $|f(x)| \ge 2^n(n+1)$ in a set of positive measure.

Fields: real analysis (integral inequalities, orthogonality to polynomials) **Position:** A6 **Answer:** Since f is orthogonal to all polynomials of degree $< n$, $\int_0^1 (x-\frac{1}{2})^n f(x)\,dx = 1$; if $|f| < 2^n(n+1)$ almost everywhere, this integral would be strictly less than $2^n(n+1)\int_0^1 |x-\frac{1}{2}|^n\,dx = 1$.

Caveat. The OCR of the statement in the MAA volume is garbled at "$\int_0^1 x^{n-1}f(x)\,dx = 0$ and $\int_0^1 x^n f(x)\,dx = 1$"; the reading above is the standard one and is confirmed by the book's own solution, which uses exactly these hypotheses. "Integrable" is read as Lebesgue integrable, as the phrase "set of positive measure" requires.

Layer 1 THE WHAT — surface analysis

The natural attempt. The conditions look like a moment problem, and the only condition that is not zero is $\int_0^1 x^n f = 1$. So bound it. Suppose $|f| < M$ almost everywhere. Then

$$1 = \int_0^1 x^n f(x)\,dx \le \int_0^1 x^n |f(x)|\,dx < M\int_0^1 x^n\,dx = \frac{M}{n+1}.$$

So $M > n+1$, i.e. $|f| \ge n+1$ on a set of positive measure. That is correct, and it is off from the target by a factor of exactly 2^n. The zero conditions have not been used at all, which is the tell-tale sign.

The next attempt is usually to use them one at a time: $\int (x^n - x^{n-1})f = 1$, so $1 \le M\int_0^1 x^{n-1}(1-x)\,dx = M/(n(n+1))$; better for $n \ge 2$ but not the right shape. Trying $x^n - \frac{1}{2}x^{n-1}$, or $x^{n-1}(x-c)$ and optimising c, gives improvements that never reach 2^n and the student starts hunting coefficients blindly. That is the wall: a family of polynomials, no principle for choosing one.

The model chain. **Step 1: f integrates to zero against every polynomial of degree $< n$.** If $r(x) = \sum_{k=0}^{n-1} c_k x^k$ then $\int_0^1 rf = \sum c_k \int_0^1 x^k f = 0$ by linearity. Hence for every *monic* polynomial q of degree n, writing $q = x^n + r$,

$$\int_0^1 q(x)f(x)\,dx = \int_0^1 x^n f(x)\,dx + \int_0^1 r(x)f(x)\,dx = 1.$$

Step 2: choose $q = (x-\frac{1}{2})^n$. This is monic of degree n, so $\int_0^1 (x-\frac{1}{2})^n f(x)\,dx = 1$. Compute

$$\int_0^1 |x-\tfrac{1}{2}|^n\,dx = 2\int_0^{1/2} u^n\,du = \frac{2}{n+1}\Big(\frac{1}{2}\Big)^{n+1} = \frac{1}{2^n(n+1)}.$$

Step 3: contradiction. Let $M = 2^n(n+1)$ and suppose the conclusion fails, i.e. the set $\{x : |f(x)| \ge M\}$ has measure zero, so $|f(x)| < M$ for almost every x. The function $\phi(x) = |x-\frac{1}{2}|^n\,(M - |f(x)|)$ is then non-negative almost everywhere and in fact strictly positive for almost every x (it can vanish only at $x = \frac{1}{2}$ or where $|f| \ge M$, a null set). A non-negative function with integral zero vanishes almost everywhere, so $\int_0^1 \phi > 0$, that is,

$$\int_0^1 |x-\tfrac{1}{2}|^n |f(x)|\,dx < M\int_0^1 |x-\tfrac{1}{2}|^n\,dx = 1.$$

But

$$1 = \int_0^1 (x-\tfrac{1}{2})^n f(x)\,dx \le \int_0^1 |x-\tfrac{1}{2}|^n |f(x)|\,dx < 1,$$

a contradiction. So $|f| \geq 2^n(n+1)$ on a set of positive measure.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains start from "if $|f| < M$ then $1 = \int (\text{something})\, f \leq M \int |\text{something}|$". They split at the choice of the something. The natural attempt uses x^n, because that is the integral the problem hands you with value 1. The model notices that the zero conditions make *every* monic degree-n polynomial give the value 1, and so picks the one with the smallest $\int_0^1 |q|$.

> *The hypotheses say f can't see any polynomial of degree below n. So x^n isn't special: I can add any lower-degree polynomial to it for free and the integral stays 1. Which monic degree-n polynomial is smallest on $[0,1]$? One centred in the middle of the interval — $(x - \frac{1}{2})^n$.*

> *My bound is off by exactly 2^n. Where does 2^n come from on $[0,1]$? From $(\frac{1}{2})^n$ — the distance to the midpoint. So the polynomial should be centred at $\frac{1}{2}$.*

Layer 3 THE WHY

Why is shifting the centre the natural move? Start from the definition of what the hypotheses say. "$\int x^k f = 0$ for $k < n$" is the statement that f is orthogonal to the space $\mathcal{P}_{n-1}$ of polynomials of degree less than n. Orthogonality to a subspace means that the pairing $q \mapsto \int qf$ depends only on q *modulo* that subspace. Modulo $\mathcal{P}_{n-1}$, a polynomial of degree n is determined by its leading coefficient. So the one number the problem gives you, $\int x^n f = 1$, is really a statement about the whole coset $x^n + \mathcal{P}_{n-1}$, i.e. about all monic polynomials of degree n at once. Having freedom in the choice of q, the bound $1 \leq M \int_0^1 |q|$ is best when $\int_0^1 |q|$ is smallest. That is a minimisation over a family, and you do it the ordinary way: take a one-parameter family you can compute, $q_c = (x - c)^n$, where

$$\int_0^1 |x - c|^n \, dx = \frac{c^{n+1} + (1-c)^{n+1}}{n+1} \qquad (0 \leq c \leq 1),$$

which is convex and symmetric about $c = \frac{1}{2}$, so minimised there, with value $1/(2^n(n+1))$. The target constant $2^n(n+1)$ is the reciprocal of this minimum. The problem is telling you which polynomial to use, if you ask where its constant comes from.

The book's solution is two lines: "the conditions imply $\int (x - \frac{1}{2})^n f = 1$", then the estimate. Here is where the $\frac{1}{2}$ comes from: it is the centre that minimises $\int_0^1 |x - c|^n$, and it is recoverable from the 2^n in the statement.

A remark on sharpness, for orientation. Over *all* monic q of degree n (not only $(x-c)^n$), the minimum of $\int_{-1}^1 |q|$ is 2^{1-n}, attained by $2^{-n}U_n$ with U_n the Chebyshev polynomial of the second kind (a classical theorem of Korkine and Zolotarev, cited not proved here); rescaled to $[0,1]$ the minimum of $\int_0^1 |q|$ is 4^{-n}. So the same argument gives $|f| \geq 4^n$ on a set of positive measure, and $4^n \geq 2^n(n+1)$ because $2^n \geq n+1$, with equality at $n = 1$ (where $(x - \frac{1}{2})$ is already optimal). The problem's constant is the one the simplest family reaches; the method is the same.

Difficulty note: this is an A6, but it is not a hard A6 once the orthogonality is read correctly. The difficulty is entirely in seeing that x^n may be replaced.

Layer $+\alpha$ BACKTESTING

1. The first cue is a list of vanishing moments $\int x^k f = 0$, $k < n$. Write down: "so $\int pf = 0$ for every polynomial p of degree $< n$, and $\int qf = 1$ for every monic q of degree n."

2. A lower bound on $\sup |f|$ from $\int qf = 1$ is Hölder/triangle inequality: $1 \leq \sup |f| \cdot \int |q|$. The method is: minimise $\int |q|$ over the allowed q.

3. Hesitation comes after the first estimate gives $n+1$ instead of $2^n(n+1)$. Reinforce with: an unused hypothesis is a hypothesis that should change your choice of test function.

4. The trap is treating $\int x^n f = 1$ as the only usable equation because it is the only non-zero one, and then hunting coefficients for $x^n + \cdots$ blindly instead of recognising the free coset.

5. When a constant in the statement looks like 2^n or $(\frac{1}{2})^n$ on $[0,1]$, suspect the midpoint; when the constant is a reciprocal of an integral, identify which integral and reverse-engineer the test function.

6. Decision rule: if f is orthogonal to a subspace V, any pairing $\int gf$ may be computed with g replaced by any $g+v$, $v \in V$; to bound f from below, pick the representative of $g+V$ of smallest norm.

▷ **Cue:** $\int x^k f = 0$ for all $k < n$ and one non-zero moment is given $\longrightarrow$ **Move:** replace x^n by any monic degree-n polynomial; choose the one minimising $\int |q|$

▷ **Cue:** a target constant 2^n appears for a problem on $[0,1]$ $\longrightarrow$ **Move:** centre the test function at $\frac{1}{2}$, since $|x - \frac{1}{2}| \leq \frac{1}{2}$

▷ **Cue:** a first estimate is off from the target by a clean factor and some hypotheses were unused $\longrightarrow$ **Move:** use the unused hypotheses to change the test function, not to add a separate estimate

Putnam 1972 — Paper B

1972 B1

1972 B1. Show that the power series representation for the series $\sum_{n=0}^{\infty} \frac{x^n(x-1)^{2n}}{n!}$ cannot have three consecutive zero coefficients.

Caveat. The source OCR renders the exponent on $(x-1)$ as the unreadable fragment "$(x-1Jln)$". Reconstructed as $(x-1)^{2n}$, not $(x-1)^n$: the book's own solution proves the general statement "the series expansion about any point for $\exp(P(x))$, if $P(x)$ is a cubic polynomial, will not have three consecutive zero coefficients," and its alternate solution uses the coefficient bound $0 \le k-n \le 2n$, i.e. the product $x^n(1-x)^{2n}$. Both only make sense if the summed polynomial is $x(x-1)^2$, a cubic, matching $f(x) = \sum x^n(x-1)^{2n}/n! = e^{x(x-1)^2}$. Cross-checked against both of the book's independently stated solutions; treated as settled.

Fields: analysis, power series, exponential generating functions **Position:** B1 (opener) **Answer:** No: at most two consecutive Maclaurin coefficients of $e^{x(x-1)^2}$ can vanish, because three consecutive vanishing derivatives at 0 propagate forward forever and force the function to be a polynomial, which $e^{x(x-1)^2}$ is not.

Layer 1 THE WHAT — surface analysis

The natural attempt. The first move is forced and correct: $\sum_{n=0}^{\infty} y^n/n! = e^y$ with $y = x(x-1)^2$, so $f(x) = e^{P(x)}$ where $P(x) = x(x-1)^2 = x^3 - 2x^2 + x$ is a cubic. Having named f, the natural next step is to go get the coefficients: expand $(x-1)^{2n}$ by the binomial theorem, $(x-1)^{2n} = \sum_{j=0}^{2n} \binom{2n}{j}(-1)^{2n-j}x^j$, so $x^n(x-1)^{2n} = \sum_j (-1)^{2n-j}\binom{2n}{j}x^{n+j}$. Dividing by $n!$ and summing over n, the coefficient of x^k collects every n for which $n \le k \le 3n$ (since $j = k-n$ must lie in $[0, 2n]$), i.e. $\lceil k/3 \rceil \le n \le k$:

$$c_k = \sum_{n=\lceil k/3 \rceil}^{k} \frac{(-1)^{k-n}\binom{2n}{k-n}}{n!}.$$

This is a real formula and it is exactly how the book's own alternate solution proceeds: multiply through by $(k-1)!$, observe every term except $n = k$ becomes an integer, and the $n = k$ term becomes $1/k$, so $(k-1)!\,c_k \notin \mathbb{Z}$ for $k > 1$ and hence $c_k \neq 0$ for every $k > 1$ (with $c_0 = c_1 = 1$ checked directly) — actually proving *no* coefficient vanishes, a stronger conclusion than asked. It works. But arriving at it costs an explicit binomial sum, a summation-index inequality to track, and a divisibility argument about $\binom{2n}{k-n}/n! \cdot (k-1)!$ being an integer for every n in range except the top one — real bookkeeping, and a wrong index inequality anywhere silently breaks the whole argument.

The model chain. Do not expand the series. Use what "$f = e^{P(x)}$" actually buys you: a differential equation. Since $f = e^P$, $f' = P'f$. Differentiate this $k-1$ more times with Leibniz's rule:

$$f^{(k+1)} = (fP')^{(k)} = \sum_{i=0}^{k} \binom{k}{i} f^{(k-i)} (P')^{(i)}.$$

Because P is cubic, P' is quadratic, so $(P')^{(i)} = 0$ for $i \geq 3$: only $i = 0, 1, 2$ survive, and $(P')^{(2)} = P'''$ is a nonzero constant ($P''' = 6$). So for every $k \geq 2$,

$$f^{(k+1)} = f^{(k)}P' + kf^{(k-1)}P'' + \binom{k}{2} f^{(k-2)}P'''. \qquad (*)$$

This is a fixed-order-3 linear recursion on the derivative sequence, evaluated at any point x_0 — in particular at $x_0 = 0$, where the coefficients are $c_j = f^{(j)}(0)/j!$, so $f^{(j)}(0) = 0 \iff c_j = 0$. Suppose, for contradiction, three consecutive Maclaurin coefficients vanish: $c_{k-2} = c_{k-1} = c_k = 0$ for some $k \geq 2$, i.e. $f^{(k-2)}(0) = f^{(k-1)}(0) = f^{(k)}(0) = 0$. Then $(*)$ gives $f^{(k+1)}(0) = 0$ too. But now $f^{(k-1)}(0) = f^{(k)}(0) = f^{(k+1)}(0) = 0$ — the same three-in-a-row pattern shifted up by one — so applying $(*)$ again gives $f^{(k+2)}(0) = 0$, and by induction $f^{(j)}(0) = 0$ for every $j \geq k-2$. So the Taylor series of f at 0 terminates: f agrees, on an open interval around 0, with the degree-$(k-3)$ polynomial given by its own truncated Taylor series (since f is entire, being $e^{P(x)}$ with P a polynomial). Two entire functions agreeing on an open set are identical, so f would equal a polynomial everywhere. That is impossible: $e^{P(z)}$ is never zero for any complex z, while every nonconstant polynomial has a root, and f is not constant (its derivative $f' = P'f$ is not identically 0, as P' is a nonzero quadratic and f is nowhere 0). Contradiction. So no three consecutive Maclaurin coefficients of f can vanish.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains agree instantly that $f(x) = e^{P(x)}$ with $P(x) = x(x-1)^2$ cubic — that identification is forced by the summation formula for e^y. They split immediately after: the natural attempt treats "control the coefficients" as "compute the coefficients," reaching for the binomial expansion of $(x-1)^{2n}$; the model treats "control the coefficients" as "control the derivatives at 0 via the defining relation $f' = P'f$," never expanding $(x-1)^{2n}$ at all.

> *I don't need a formula for each coefficient — I need to know they can't align three zeros in a row, and f satisfies its own differential equation $f' = P'f$; differentiating that equation is a statement entirely about f's derivatives feeding back into each other, which is precisely a claim about consecutive coefficients, with no expansion required.*

Layer 3 THE WHY

The reason to differentiate $f' = P'f$ rather than expand the series is that $f = e^{P(x)}$ is defined by an ODE, and an ODE is a machine for producing relations *between consecutive derivatives* — which is exactly the shape of the thing being asked about ("three consecutive coefficients"). Multiplying out $(x-1)^{2n}$ throws that structure away and replaces it with raw combinatorics that has to rediscover, term by term, a fact the differential equation hands over for free.

The reason the recursion $(*)$ has exactly three terms, no more, is the cubic degree of P: P cubic $\Rightarrow$ P' quadratic $\Rightarrow$ $(P')^{(i)} = 0$ for $i \geq 3$. This is not a coincidence to be verified after the fact; it is the entire reason the problem specifies a cubic and not some other polynomial. A recursion of bounded order d behaves exactly like a linear recurrence relation: d consecutive zero "seed" values force every subsequent value to be zero, by the same one-line induction used for a linear recurrence such as the Fibonacci relation. The bound "d" here is $\deg P - 1$ read off from how many derivatives of P' survive — so the book's own remark that this generalises to "any cubic polynomial" is not decoration, it identifies the actual mechanism: for $\exp(Q(x))$

with $\deg Q = d$, no d consecutive Maclaurin coefficients can vanish, by the identical argument with $(*)$ having d terms.

The final contradiction — "f becomes a polynomial, but f is not a polynomial" — is where the problem's specific choice of $f = \exp(\text{something})$ does its last piece of work: an exponential of a nonconstant polynomial is transcendental in the cheapest possible sense available here, it is never zero, and every nonconstant polynomial is zero somewhere. No growth-rate or degree argument is needed; non-vanishing versus having-a-root is already enough to separate the two function classes completely.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the shape $\sum_n y^n/n!$ with y a specific polynomial in x: on seeing it, write down "$f = e^{P(x)}$, $\deg P = 3$" before touching a single binomial expansion.

2. That cue suggests differentiating the defining relation $f' = P'f$ repeatedly via Leibniz's rule, rather than expanding the series to find an explicit coefficient formula.

3. Hesitation sets in right after naming $f = e^{P(x)}$, where the instinct is "now find the coefficients" — the default meaning of "control a coefficient" is "compute it." Reinforce against it by asking whether the claim needed is about individual coefficient values or only about which *patterns* of zero/nonzero can occur; the latter is a job for a recursion, not a formula.

4. The trap is that the binomial-expansion route is not a dead end — it genuinely finishes the problem (and proves more) — so there is no wall to hit that would force a reroute; the cost is invisible until you are already committed to tracking a summation index inequality and an integrality argument, at which point the shorter differential route no longer looks obviously better in hindsight, only in advance.

5. Recognise this faster next time by treating any "exp(polynomial)" object as automatically carrying a first-order ODE ($f' = P'f$), and by treating "no d consecutive zero coefficients" claims as recursion-order statements: differentiate the ODE $d-1$ times and count how many terms survive.

6. Decision rule: for $f = \exp(Q(x))$ with Q a polynomial of degree d, questions about how many consecutive Maclaurin coefficients can vanish are answered by the order-d linear recursion obtained from repeatedly differentiating $f' = Q'f$ — d consecutive zero derivatives at the expansion point propagate to all later derivatives being zero, contradicting that $\exp(Q)$ is nowhere zero while agreeing with a polynomial would force it to have a root.

1972 B2

1972 B2. A particle moving on a straight line starts from rest and attains a velocity v_0 after traversing a distance s_0. If the motion is such that the acceleration was never increasing, find the maximum time for the traverse.

Fields: calculus, concavity, optimisation via inequalities **Position:** B2 **Answer:** $t_{0,\max} = \dfrac{2s_0}{v_0}$, attained exactly by constant acceleration.

Layer 1 THE WHAT — surface analysis

The natural attempt. "Acceleration never increasing" reads like a constraint to search over, so the natural move is to guess candidate acceleration profiles and see which gives the longest time. Constant acceleration $a(t) = a_0$ is the obvious first try, since it is trivially non-increasing (it doesn't change at all): then $v(t) = a_0 t$, $v(t_0) = v_0 \Rightarrow a_0 = v_0/t_0$, and $s_0 = \int_0^{t_0} v\,dt = \frac{1}{2}a_0 t_0^2 = \frac{1}{2}v_0 t_0$, giving $t_0 = 2s_0/v_0$ — the right answer, reached correctly. The trouble is knowing this is the *maximum* and not just *a* value. A careful student tries to check other legal profiles: an acceleration that starts high and decreases, e.g. $a(t) = a_0 - kt$ for $k > 0$, still non-increasing,

gives $v(t) = a_0 t - \frac{1}{2}kt^2$, and now t_0, a_0, k are linked by two equations ($v(t_0) = v_0$, $\int_0^{t_0} v = s_0$) with a free parameter k — a two-parameter family to optimise t_0 over, and nothing yet says why every member of this family, let alone every non-increasing-acceleration profile whatsoever, gives $t_0 \leq 2s_0/v_0$. Checking families one at a time never proves a bound over *all* legal motions; it only ever produces more candidates.

The model chain. Translate the physical hypothesis into one calculus fact and stop generating candidates. "Acceleration never increasing" means $a(t) = v'(t)$ is a non-increasing function of t, i.e. $v''(t) = a'(t) \leq 0$: the velocity graph is concave on $[0, t_0]$. A concave function lies on or above the straight line (chord) joining any two of its points — that is literally the definition of concavity applied to the two points $t = 0$ and $t = t_0$. Here $v(0) = 0$ and $v(t_0) = v_0$, so the chord is the line $\ell(t) = \dfrac{v_0}{t_0} t$, and concavity gives

$$v(t) \geq \ell(t) \quad \text{for all } t \in [0, t_0].$$

Since the integral of a pointwise inequality preserves the inequality,

$$s_0 = \int_0^{t_0} v(t)\,dt \;\geq\; \int_0^{t_0} \ell(t)\,dt = \frac{1}{2} \cdot \frac{v_0}{t_0} \cdot t_0^2 = \frac{1}{2} v_0 t_0.$$

Rearranging, $t_0 \leq \dfrac{2s_0}{v_0}$ — a bound that holds for *every* motion with non-increasing acceleration simultaneously, with no case-work and no search over families. Equality in the chord inequality holds exactly when v is linear on $[0, t_0]$, i.e. constant acceleration, which is precisely the natural attempt's first guess — now certified as optimal rather than merely plausible. So the maximum traverse time is $t_{0,\max} = 2s_0/v_0$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains correctly identify constant acceleration as a candidate achieving $t_0 = 2s_0/v_0$ — that much is shared. They split on how to establish it is the *maximum*: the natural attempt tries additional explicit acceleration profiles one by one, hoping none beats the constant case; the model instead translates "acceleration never increasing" directly into "v is concave" and applies the single geometric fact that a concave function's graph lies above its chord, which bounds the integral for every legal motion at once.

> *I don't need to search over acceleration profiles at all — "acceleration never increasing" is exactly the definition of v being concave, and a concave function with known values at both endpoints is automatically bounded below, everywhere in between, by the straight line joining those endpoints; that single inequality, integrated, is the whole problem.*

Layer 3 THE WHY

The reason the chord inequality settles the problem in one line is that it is definitional, not derived: concavity of v on $[0, t_0]$ *means* $v(\lambda \cdot t_0 + (1-\lambda)\cdot 0) \geq \lambda v(t_0) + (1-\lambda)v(0)$ for $\lambda \in [0, 1]$, which is exactly "the graph lies on or above the line through the endpoints." Once the physical hypothesis is correctly translated into this one word — concave — nothing further needs to be invented; the rest of the argument is monotonicity of the integral (a bigger integrand pointwise gives a bigger integral) applied to that one inequality. The real content of the problem is entirely in the translation step: "acceleration is a non-increasing function of time" must be read as a statement about the *second* derivative of position ($a = v' = x''$), and "a non-increasing" as $a' \leq 0$, i.e. $v'' \leq 0$. Once that reading is made, treating the problem as an optimisation over explicit profiles (the natural attempt) is strictly more work for the same conclusion, because it re-derives, profile by profile, a fact the definition of concavity already guarantees for every profile in the constraint class simultaneously. The equality case reinforces this: equality in "graph above

chord" holds exactly at linearity, so the bound is not merely an estimate that happens to be met by the constant-acceleration guess — the constant-acceleration case *is* the chord, definitionally, which is why it was the natural first guess to try and also the exact extremiser.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the phrase "acceleration was never increasing" itself: on seeing it, write down immediately $a'(t) \leq 0$, i.e. $v''(t) \leq 0$ — the physical constraint restated as a concavity condition on v, before considering any specific profile.
2. That cue suggests the chord inequality for concave functions: $v(t) \geq$ the straight line joining $(0, v(0))$ and $(t_0, v(t_0))$, applied on the interval $[0, t_0]$.
3. Hesitation occurs right after guessing constant acceleration works, where it is tempting to stop at "found a candidate achieving the bound" without a device that shows no other legal motion can beat it. Reinforce against it by asking: is there a single inequality, true for the whole constraint class at once, that this candidate merely achieves with equality — rather than one profile among many to be checked?
4. The trap is mistaking "never increasing" for an invitation to enumerate and optimise over acceleration *families* (constant, linearly decreasing, and so on), when it is instead a single global geometric property (concavity of v) that bounds every member of the entire constraint class in one inequality.
5. Recognise this faster next time by treating any phrase of the form "the rate of change of X is never increasing/decreasing" as an instruction to write down the sign of the second derivative of the underlying quantity, and then reach for the chord (or tangent-line) inequality for that concavity/convexity before trying explicit functional forms.
6. Decision rule: to bound an accumulated quantity (an area, a distance, an integral) governed by a function with known concavity and fixed endpoint values, compare the function directly to the straight line joining those endpoints and integrate the resulting pointwise inequality; equality holds exactly at linearity, so this simultaneously identifies both the bound and its extremiser.

1972 B3

1972 B3. Let A and B be two elements in a group such that $ABA = BA^2B$, $A^3 = 1$ and $B^{2n-1} = 1$ for some positive integer n. Prove $B = 1$.

Fields: group theory, element relations, commutativity from order conditions **Position:** B3 **Answer:** $B = 1$, forced once A and B are shown to commute.

Layer 1 THE WHAT — surface analysis

The natural attempt. The relation $ABA = BA^2B$ has two unknowns tangled together, and the natural instinct is to isolate B directly. $A^3 = 1$ lets you replace A^2 by A^{-1}, giving $ABA = BA^{-1}B$. From here it is tempting to multiply by things to peel a B off one side: multiply on the right by B^{-1} to get $ABAB^{-1} = BA^{-1}$, or conjugate by A^{-1} on the left to get $BA = A^{-1}BA^{-1}B$. Each of these is legal algebra, and each produces a new relation that still has A and B mixed in a longer, not shorter, word. Nothing cancels, because A and B are not assumed to commute, so there is no way to gather all the B's on one side and all the A's on the other. Continuing to multiply and substitute produces an ever-growing word in A, B, A^{-1}, B^{-1} with no visible target length, and no guarantee that any particular sequence of moves terminates in something solvable — the order hypotheses $A^3 = 1$ and $B^{2n-1} = 1$ sit unused on the sidelines the whole time, because nothing in this line of attack asks what they are for.

The model chain. Do not try to isolate B from a single relation between two non-commuting elements. Aim instead for an intermediate goal: prove A and B commute, then revisit the

original relation once cancellation is legal. With $A^2 = A^{-1}$, the hypothesis reads

$$ABA = BA^{-1}B. \tag{†}$$

Compute AB^2 two different ways using (†). First, trivially, $AB^2 = (ABA)A^{-1}B$ (insert $A \cdot A^{-1} = 1$). Substituting (†) for ABA:

$$AB^2 = (BA^{-1}B)A^{-1}B = BA^{-1}\big(BA^{-1}B\big).$$

The bracketed piece $BA^{-1}B$ is exactly the right-hand side of (†) again, so substitute it back as ABA:

$$AB^2 = BA^{-1}(ABA) = B(A^{-1}A)BA = B \cdot BA = B^2A.$$

So $AB^2 = B^2A$: A commutes with B^2. Commuting with an element means commuting with all its powers, so by a one-line induction A commutes with B^{2r} for every positive integer r (base case just shown; inductive step: if $AB^{2r} = B^{2r}A$ then $AB^{2(r+1)} = AB^{2r}B^2 = B^{2r}AB^2 = B^{2r}B^2A = B^{2(r+1)}A$). Take $r = n$: $AB^{2n} = B^{2n}A$. But $B^{2n-1} = 1$ gives $B^{2n} = B^{2n-1} \cdot B = B$, so this says exactly $AB = BA$: A and B commute.

Now revisit the original relation with commutativity in hand. $ABA = A \cdot AB = A^2B$ (using $BA = AB$), and $BA^2B = A^2B \cdot B = A^2B^2$ (using $BA^2 = A^2B$, which follows from $BA = AB$). So $ABA = BA^2B$ becomes

$$A^2B = A^2B^2.$$

Group elements are invertible, so multiplying both sides by A^{-2} cancels freely (no zero divisors in a group): $B = B^2$. Multiplying both sides by B^{-1} gives $B = 1$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains agree on the forced first substitution $A^2 \to A^{-1}$, giving $ABA = BA^{-1}B$. They split on what to do with that relation: the natural attempt keeps manipulating it directly, trying to isolate B from a relation between two elements that are not yet known to commute; the model sets a different target entirely — compute AB^2 two ways to test whether A commutes with B^2 — treating commutativity as something to be manufactured first, before attempting to solve for B at all.

> *One relation in two non-commuting unknowns is not enough to isolate either one directly — but I haven't used the order hypotheses yet, and $A^3 = 1$ together with $B^{2n-1} = 1$ (odd order) is exactly the kind of setup that lets me manufacture a commuting relation first, so let me compute the same product two ways and see what falls out, instead of trying to solve for B from (†) itself.*

Layer 3 THE WHY

The reason $AB^2 = B^2A$ is the right target, rather than any other combination, is the numerology of the hypothesis: B has *odd* order $2n - 1$, and the one algebraic fact that hands over for free is $B^{2n} = B^{2n-1} \cdot B = B$ — an even power of B collapsing to B itself. That fact is useless unless you already have control over some even power of B, so the natural target to chase is exactly "does A commute with B^2?", because commuting with B^2 propagates for free (by the one-line induction, itself just the elementary fact that commuting with x means commuting with every power of x) to commuting with B^{2n}, and $B^{2n} = B$ closes the bridge the problem's exponent $2n - 1$ was built to cross. This is the general shape worth keeping: whenever a hypothesis gives $x^m = 1$ for *odd* m, any established fact about "commutes with x^2" upgrades immediately to "commutes with x" via $x^{m+1} = x$ — a bridge that does not exist so cheaply when m is even, since then $x^{m+1} = x$ still requires an even power, not x itself.

The reason $AB^2 = B^2A$ is *derivable at all* from (†) is a self-referential trick worth naming explicitly: substituting (†) into itself. The word $BA^{-1}B$ that appears after the first substitution

is not a dead end to expand further — it is syntactically identical to the right-hand side of (†), so it can be read backwards and replaced by ABA again. This is the move the natural attempt's manipulations never surface, because they are chasing "isolate B" rather than "recognise the relation's own shape reappearing inside a longer word." Once the relation's right-hand side is spotted a second time, the $A^{-1}A$ pair that remains collapses the whole expression to B^2A — cancellation that was never available while A, B were still tangled, and is only available because the self-substitution was aimed, not blind.

Finally, once commutativity is established, the last step — cancelling A^2 from $A^2B = A^2B^2$ — needs no cleverness at all: it is legal precisely because a group has no zero divisors (every element is invertible), so $A^2X = A^2Y \Rightarrow X = Y$ unconditionally. The entire difficulty of the problem is manufacturing the right to do that cancellation; once granted, $B = B^2 \Rightarrow B = 1$ is definition-chasing.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is $A^3 = 1$ appearing alongside a relation that mixes A^2 into a word with B: on seeing it, write down $A^2 = A^{-1}$ immediately and rewrite the given relation as $ABA = BA^{-1}B$ before attempting any further manipulation.
2. That cue, combined with a second element of stated (odd) order, suggests the sub-goal "prove A, B commute" rather than "solve for B directly" — specifically by computing some product such as AB^2 two different ways using the given relation.
3. Hesitation appears right after the first substitution, when the word $BA^{-1}\big(BA^{-1}B\big)$ (length 5) looks like it is getting longer, not shorter. Reinforce against it by noticing the bracketed piece $BA^{-1}B$ is exactly the right-hand side of the original relation, so it can be substituted back rather than expanded further — the relation is being used a second time, in reverse.
4. The trap is treating the single given relation as directly solvable for B, as if in an abelian group; any manipulation attempted before commutativity is established either fails to cancel or produces an ever-longer word, and the odd-order hypothesis on B sits unused the entire time such an attempt is made.
5. Recognise this faster next time by checking, as soon as two group elements satisfy one mixed relation plus separate order conditions, whether those order conditions (especially an *odd* order) are enough to manufacture commutativity first — that reframes the whole problem from "solve one equation in two unknowns" to "prove a lemma, then solve one equation in one unknown."
6. Decision rule: given $x^m = 1$ with m odd, any proven fact of the form "y commutes with x^2" upgrades for free to "y commutes with x" via $x^{m+1} = x$; when a relation entangles two non-commuting elements with one of odd order, search for a way to produce the even-power commuting fact before attempting to isolate either variable directly.

1972 B4

1972 B4. Let n be an integer greater than 1. Show that there exists a polynomial $P(x, y, z)$ with integral coefficients such that $x = P(x^n, x^{n+1}, x + x^{n+2})$.

Fields: algebra (polynomial identities) / number theory (numerical semigroups) **Position:** B4 – fourth problem of the second session **Answer:** Take $P(x, y, z) = z\sum_{r=0}^{n-2}(-y)^r + (-1)^{n-1}x^n$; substituting gives $x - (-1)^{n-1}x^{n^2} + (-1)^{n-1}x^{n^2} = x$.

Layer 1 THE WHAT — surface analysis

The natural attempt. Name the three inputs: $a = x^n$, $b = x^{n+1}$, $c = x + x^{n+2}$. You want x, and the only input containing a bare x is c. So write $x = c - x^{n+2}$ and try to manufacture x^{n+2} out of a and b. For $n = 2$ that works at once: $x^4 = a^2$, so $P = z - x^2$ in the problem's variables,

and $P(x^2, x^3, x + x^4) = x + x^4 - x^4 = x$. Encouraged, you try $n = 3$: you need x^5 from x^3 and x^4. Products of a and b give exponents $3, 4, 6, 7, 8, \ldots$ – every sum of 3s and 4s – and 5 is not among them. Same wall for every $n \geq 3$: the exponents you can reach with monomials in a, b are the numbers $\alpha n + \beta(n+1)$ with $\alpha, \beta \geq 0$, and between $n+1$ and $2n$ there is nothing, so $n+2$ (which is less than $2n$ once $n > 2$) is unreachable. At this point most people try to cancel x^{n+2} with some other combination of a and b, fail, and start writing out tables of exponents for $n = 3, 4, 5$ hoping a pattern appears. That is the zombie route: it is brute force on the wrong object.

The model chain. Keep the error term, and ask what it is made of. The error x^{n+2} is $x \cdot x^{n+1} = x \cdot b$, and x is the very thing you are building. So feed the equation back into itself:

$$x = c - xb = c - (c - xb)b = c - cb + xb^2 = \cdots$$

After m rounds,

$$x = c \sum_{r=0}^{m-1} (-b)^r + (-1)^m x\, b^m,$$

which you can verify in one line: $c \sum_{r=0}^{m-1}(-b)^r = x(1+b) \cdot \frac{1-(-b)^m}{1+b} = x - (-1)^m x b^m$ (the factor $1 + b$ cancels as a polynomial identity, since $(1+b)\sum_{r=0}^{m-1}(-b)^r = 1 - (-b)^m$). Each round pushes the error's exponent up by $n+1$: after m rounds it is $x\, b^m = x^{m(n+1)+1}$. The only question left is when that exponent becomes reachable by a and b. Take $m = n - 1$: the exponent is $(n-1)(n+1) + 1 = n^2$, and $x^{n^2} = (x^n)^n = a^n$. The sign of the correction is where people slip, so read it off the identity rather than guessing: from $c \sum_{r=0}^{m-1}(-b)^r = x - (-1)^m x^{m(n+1)+1}$ with $m = n - 1$,

$$c \sum_{r=0}^{n-2} (-b)^r = x - (-1)^{n-1} x^{n^2} = x - (-1)^{n-1} a^n.$$

Hence

$$P(x, y, z) = z \sum_{r=0}^{n-2} (-1)^r y^r + (-1)^{n-1} x^n$$

has integer coefficients and satisfies $P(x^n, x^{n+1}, x + x^{n+2}) = x$ identically. Check $n = 2$: $P = z - x^2$, as found by hand. Check $n = 3$: $P = z - zy + x^3$, and $(x + x^5) - (x + x^5)x^4 + x^9 = x + x^5 - x^5 - x^9 + x^9 = x$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains write $x = c - x^{n+2}$. At that line the natural attempt treats x^{n+2} as a target to be built from a, b alone; the model reads x^{n+2} as $x \cdot b$ – an error term that *contains the unknown* – and substitutes the equation into itself.

> *I can't build x^{n+2} from x^n and x^{n+1} – but x^{n+2} is x times x^{n+1}, and x is what I'm solving for. So I substitute $x = c - xb$ into its own error term, and every round the leftover power climbs by $n+1$. I keep going until the leftover power is one I* can *build.*

> *Which powers can I build from x^n and x^{n+1}? Sums $\alpha n + \beta(n+1)$. Once the exponent is at least $(n-1)n$ everything is reachable, so I only need the error exponent $m(n+1)+1$ to be large – and $m = n - 1$ gives exactly $n^2 = n \cdot n$.*

Layer 3 THE WHY

Why is self-substitution natural rather than a trick? Look at what c is: $c = x + x^{n+2} = x(1 + x^{n+1}) = x(1+b)$. So the relation between the inputs and the unknown is $x = c/(1+b)$.

If you were allowed rational functions, the problem would be over in one line. You are not; you are allowed polynomials. And the standard way to turn $1/(1+b)$ into something polynomial is the geometric series $1 - b + b^2 - \cdots$. The iteration in the model chain *is* that geometric series, truncated after m terms; the leftover $\pm xb^m$ is the tail you cut off. So the real first thought is "c factors as $x(1+b)$, so $x = c/(1+b)$, so expand $1/(1+b)$." The iteration is what that looks like before you have recognised it.

Why does truncation work here when it usually does not? Because of the second ingredient: the tail is a single monomial $x^{m(n+1)+1}$, and monomials in x of high enough degree are free. The exponents reachable from n and $n+1$ form a numerical semigroup; since $\gcd(n, n+1) = 1$, it contains every integer beyond its Frobenius number $n(n+1) - n - (n+1) = n^2 - n - 1$. (Directly: any $N \geq n(n-1)$ can be written $N = qn + s$ with $0 \leq s \leq n-1$ and $q \geq n-1 \geq s$, and then $N = (q-s)n + s(n+1)$.) So the obstruction that killed the natural attempt – the gap between $n+1$ and $2n$ – exists only at low degree. The model's move is to push the problem out of the region with gaps into the region without. The natural attempt fought the gap at its worst point, degree $n+2$.

It is also worth seeing that $m = n-1$ is forced, not merely convenient. You need $m(n+1)+1 = \alpha n + \beta(n+1)$ with $\alpha, \beta \geq 0$. Put $k = \alpha + \beta$; then $kn \leq m(n+1)+1 \leq k(n+1)$, and the right-hand inequality gives $k \geq m+1$ (since $k = m$ would need $m(n+1)+1 \leq m(n+1)$). With $k \geq m+1$, the left-hand inequality gives $(m+1)n \leq m(n+1)+1$, i.e. $m \geq n-1$. So the book's degree is the least that this route can achieve.

The official solution writes out $z, zy, zy^2, \ldots$, and says "multiply alternately by $+1$ and -1 and add." That is the geometric series written as a telescoping sum. Here is where it comes from: $c = x(1+b)$, and you invert $1+b$ the only way polynomials allow.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that the third input, $x + x^{n+2}$, is the only one with a bare x, and that its second term differs from the second input by one factor of x. On seeing it, write $c = x(1+b)$.

2. $x = c/(1+b)$ suggests inverting $1+b$ by a truncated geometric series, then disposing of the truncation remainder separately.

3. Hesitation comes after $x = c - x^{n+2}$, when x^{n+2} turns out to be unreachable and it feels as if the method has failed. Reinforce it away by asking whether the obstructing term contains the unknown. If it does, substitute, don't build.

4. The trap is testing $n = 2$ first. There $x^4 = (x^2)^2$ and one subtraction finishes, which suggests a one-step answer that does not exist for $n \geq 3$.

5. Whenever the target T appears as $T \cdot$ (something small) inside an expression you are allowed to use, write $T = E/(1+u)$ and expand; whenever the leftover is a monomial, ask which monomials are free (here: all degrees $\geq n(n-1)$, by the two-coin Frobenius bound).

6. Reusable rule: if the error term of an approximation is the unknown times a known quantity, iterate the approximation into itself (equivalently, expand a geometric series) until the error lands in a region you control outright.

▷ **Cue:** an allowed expression factors as (target)$\times(1+u)$ with u also allowed $\longrightarrow$ **Move:** write target $= E/(1+u)$ and expand $1/(1+u) = \sum(-u)^r$ to finitely many terms

▷ **Cue:** the error term after one step is the unknown times a known quantity $\longrightarrow$ **Move:** substitute the equation into its own error term; each round multiplies the error by that quantity

▷ **Cue:** need a power x^N from x^p and x^q with $\gcd(p,q) = 1$ $\longrightarrow$ **Move:** every $N \geq (p-1)(q-1)$ is $\alpha p + \beta q$ with $\alpha, \beta \geq 0$; push N past the Frobenius number

1972 B5

1972 B5. If the opposite angles of a skew (non-planar) quadrilateral are equal in pairs, prove that the opposite sides are equal in pairs.

Fields: geometry (solid geometry, law of cosines, Ptolemy's inequality) **Position:** B5 – fifth problem of the second session **Answer:** Eliminating the two angles with the law of cosines gives $BD^2(ad-bc) = (ab-cd)(bd-ac)$ and $AC^2(ab-cd) = (ad-bc)(bd-ac)$; if $(ab-cd)(ad-bc) \neq 0$ these force equality in Ptolemy's inequality, hence coplanarity, which is excluded; the remaining cases force $a=c$, $b=d$.

Layer 1 THE WHAT — surface analysis

Notation, fixed once: the quadrilateral is $ABCD$ with sides $a = AB$, $b = BC$, $c = CD$, $d = DA$ and diagonals $x = AC$, $y = BD$. The hypothesis is $\angle DAB = \angle BCD =: \alpha$ and $\angle ABC = \angle CDA =: \beta$. Since the four points are not coplanar, they are distinct and no three are collinear, so all six lengths are positive.

The natural attempt. In the plane this is a familiar fact: opposite angles equal means $2\alpha + 2\beta = 360^\circ$, so $\alpha + \beta = 180^\circ$, so adjacent sides are parallel, so $ABCD$ is a parallelogram. The strong student tries to lift this to space and finds that the angle sum is gone – a skew quadrilateral's angles add up to less than 360°, and there are no parallel sides to speak of. So they switch to coordinates or vectors: put A at the origin, write $\vec{AB} = \mathbf{u}$, $\vec{AD} = \mathbf{v}$, $\vec{AC} = \mathbf{w}$, and turn the angle conditions into dot-product equations. That gives two equations in the nine coordinates of three vectors, each involving norms in the denominators; squaring to clear them produces quartics, and there is no visible route from there to $|\mathbf{u}| = |\mathbf{w} - \mathbf{v}|$. Six lengths determine a tetrahedron, so the problem is really about lengths, and a vector set-up buries the lengths under angles and coordinates.

The model chain. Each angle appears in two triangles that share a diagonal: α sits in $\triangle DAB$ and $\triangle BCD$, both with side BD. The law of cosines on the shared side gives

$$y^2 = a^2 + d^2 - 2ad\cos\alpha = b^2 + c^2 - 2bc\cos\alpha.$$

Eliminate $\cos\alpha$: multiply the first form by bc, the second by ad, subtract:

$$y^2(ad-bc) = ad(b^2+c^2) - bc(a^2+d^2) = (ab-cd)(bd-ac). \tag{1}$$

(Expand the right side to check: $ab^2d - a^2bc - bcd^2 + ac^2d$.) The same with β in $\triangle ABC$ and $\triangle CDA$, which share AC (or relabel $A \to B \to C \to D \to A$ in (1)):

$$x^2(ab-cd) = (ad-bc)(bd-ac). \tag{2}$$

Case 1: $(ab-cd)(ad-bc) \neq 0$. Multiply (1) by (2) and cancel the non-zero factor $(ab-cd)(ad-bc)$: $x^2y^2 = (bd-ac)^2$, so $xy = |ac-bd|$, i.e.

$$AB\cdot CD = BC\cdot DA + AC\cdot BD \quad \text{or} \quad BC\cdot DA = AB\cdot CD + AC\cdot BD.$$

Both are equality cases of Ptolemy's inequality, which says that for any four points P, Q, R, S in space, $PR\cdot QS \leq PQ\cdot RS + PS\cdot QR$, with equality only when the points are coplanar (indeed concyclic, or collinear). Proof: invert in the sphere of radius 1 about P. Distances transform as $Q'R' = QR/(PQ\cdot PR)$, and similarly for the other pairs. The triangle inequality $Q'S' \leq Q'R' + R'S'$ becomes, after multiplying by $PQ\cdot PR\cdot PS$, exactly $PR\cdot QS \leq PQ\cdot RS + PS\cdot QR$. Equality needs Q', R', S' collinear. The inverse of a line is a circle through P (or the line itself, if it passes through P), so P, Q, R, S lie in one plane. Our quadrilateral is skew, so Case 1 is impossible.

Case 2: $ab = cd$. Then (2) gives $(ad-bc)(bd-ac) = 0$. If $ad = bc$: dividing $ab = cd$ by $ad = bc$ gives $b/d = d/b$, so $b = d$ and then $a = c$. Done. If $bd = ac$: dividing $ab = cd$ by

$bd = ac$ gives $a/d = d/a$, so $a = d$ and $b = c$. Then (1) reads $y^2(a^2 - b^2) = 0$, so $a = b$, and all four sides are equal. Done again.

Case 3: $ad = bc$, $ab \neq cd$. Then (1) gives $bd = ac$; with $ad = bc$ this gives $a/b = b/a$, so $a = b$, $c = d$, and (2) then gives $x^2(a^2 - c^2) = 0$, so $a = c$. All four sides equal.

In every surviving case $AB = CD$ and $BC = DA$.

Caveat. The book's first case ("$ab - cd = 0$; then $ad - bc = 0$ and $a = c$, $b = d$") skips the sub-case $ab = cd$, $bd = ac$, i.e. the kite $a = d$, $b = c$. The sub-case is harmless – equation (1) forces the kite to be a rhombus – but it has to be closed, and the analysis above does so.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains know the answer must be about lengths. The split is at the first equation written. The natural attempt encodes the angles as dot products of position vectors; the model notices that each angle and its opposite partner sit in two triangles sharing a diagonal, writes the law of cosines on that diagonal twice, and eliminates the angle. After that the whole problem is the pair of identities (1), (2), and the only non-routine step is recognising $xy = |ac - bd|$ as Ptolemy.

> *The two equal angles A and C both look across at the same diagonal BD. So I have two formulas for BD^2 with the same cosine in them – I eliminate the cosine and I'm left with a relation among lengths only.*

> *$AC \cdot BD = |AB \cdot CD - BC \cdot DA|$: products of opposite sides and of diagonals – that's Ptolemy with equality, and equality in Ptolemy means the four points are concyclic, so coplanar. That's where "skew" gets used.*

Layer 3 THE WHY

Why should the law of cosines be the first tool rather than vectors? A tetrahedron $ABCD$ is determined up to congruence by its six edge lengths, and the conclusion is a statement about four of them. An angle of the quadrilateral is an angle of a face triangle, and the law of cosines is the dictionary between face angles and edge lengths. So the natural question is "in which triangle does each given angle live, and which edge does it face?" The answer – both α's face BD, both β's face AC – hands you two equations with one unknown angle each, and elimination is forced.

Why does Ptolemy appear, rather than being imported? Look at what (1) and (2) say when you multiply them: the three pairings of the four points into two pairs – $\{AB, CD\}$, $\{BC, DA\}$, $\{AC, BD\}$ – produce three products, and the hypothesis forces one of them to equal the sum of the other two. Any relation among exactly those three products is Ptolemy territory. Ptolemy's inequality is to four points what the triangle inequality is to three (the inversion proof shows it literally is the triangle inequality, seen from one of the points), and equality in it is the four-point version of "three points are collinear": it pins the configuration into a plane.

That also explains why the word "skew" is in the problem, and it is worth seeing the counterexample it excludes. Put A, C, B, D on a circle in that cyclic order and join them as the crossed quadrilateral $ABCD$. The angles at A and C are inscribed angles on chord BD from the same side, so they are equal; the angles at B and D are inscribed angles on chord AC from the same side, so they are equal. Yet the sides need not match: a numerical instance with the four points at angular positions 0.3, 2.9, 1.4, 4.6 radians on the unit circle (for A, B, C, D) gives $AB \approx 1.93$ but $CD \approx 2.00$. So the planar statement is false, and it fails in exactly the way Case 1 predicts: concyclic points with Ptolemy's equality. The algebra in Case 1 is not a detour; it is a precise description of all the counterexamples, and the skew hypothesis kills them all at once.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that the given equal angles are *opposite*, so each pair faces a common diagonal. Write down the law of cosines for BD from A and from C, and for AC from B and from D.
2. Elimination of the shared cosine, leaving polynomial relations among the six lengths.
3. Hesitation occurs at $x^2y^2 = (ac - bd)^2$: it looks like a dead end until you read it as a statement about the three pairings of four points. Reinforce by memorising Ptolemy's inequality as "the triangle inequality for four points", with its equality case.
4. Two traps. First, trying to use the planar angle sum, which does not exist in space. Second, the degenerate cases: dividing by $ab - cd$ or $ad - bc$ without checking they are non-zero, and then skipping the kite sub-case (as the book's solution does).
5. Any hypothesis on angles of a spatial figure whose conclusion is about lengths: move to face triangles with the law of cosines immediately. Any appearance of the three products $AB \cdot CD$, $AC \cdot BD$, $AD \cdot BC$: think Ptolemy, and its equality case means concyclic.
6. Reusable rule: when a problem says "non-planar" (or "non-degenerate"), find the planar counterexample first. The proof then has to use the hypothesis at exactly the point where that counterexample would slip through, and knowing the counterexample tells you which theorem's equality case to invoke.

▷ **Cue:** two equal angles face the same segment (a shared diagonal) $\longrightarrow$ **Move:** write the law of cosines for that segment twice and eliminate the cosine
▷ **Cue:** a relation among $AB \cdot CD$, $BC \cdot DA$, $AC \cdot BD$ $\longrightarrow$ **Move:** Ptolemy's inequality; equality forces the four points to be concyclic, hence coplanar
▷ **Cue:** a hypothesis "non-planar" or "non-degenerate" in a geometry statement $\longrightarrow$ **Move:** construct the degenerate counterexample first; it shows which equality case the proof must exclude

1972 B6

1972 B6. Let $n_1 < n_2 < n_3 < \cdots < n_k$ be a set of positive integers. Prove that the polynomial $1 + z^{n_1} + z^{n_2} + \cdots + z^{n_k}$ has no roots inside the circle $|z| < (\sqrt{5} - 1)/2$.

Fields: complex analysis (modulus estimates) / power series **Position:** B6 – last problem of the second session **Answer:** Write $2P(z) = \left(1 + \frac{1}{1-z}\right) + E(z)$ where E has coefficients ± 1 from z^1 on; then $|2P(z)| \geq 1 + \frac{1}{1+r} - \frac{r}{1-r} = \frac{2(1-r-r^2)}{1-r^2} > 0$ for $r = |z| < (\sqrt{5} - 1)/2$. The constant is sharp.

Layer 1 THE WHAT — surface analysis

Notation: $P(z) = 1 + z^{n_1} + \cdots + z^{n_k}$, $S = \{n_1, \ldots, n_k\}$, $r = |z|$, and $\rho = (\sqrt{5} - 1)/2 \approx 0.618$, the positive root of $\rho^2 + \rho = 1$.

The natural attempt. The reflex for "no roots in a small disc" is to show the constant term dominates. By the triangle inequality,

$$|P(z)| \geq 1 - \sum_i r^{n_i} \geq 1 - (r + r^2 + r^3 + \cdots) = 1 - \frac{r}{1-r} = \frac{1-2r}{1-r}.$$

That is positive for $r < \frac{1}{2}$ – and stops there. The bound $\frac{1}{2}$ is not ρ, and no amount of tidying fixes it: at $r = \frac{1}{2}$ the estimate is exactly zero, because it assumes every power $z, z^2, z^3, \ldots$ is present *and* every one of them points in the direction opposite to 1. Those two assumptions cannot hold together. If z points in the direction -1 (so $z = -r$), then z^2 points in direction $+1$, and it helps rather than hurts. The triangle inequality has thrown away exactly the information the problem is testing. A student who stops here has a correct theorem with the wrong constant.

The model chain. First find the true worst case; then prove it is the worst.

Worst case. To make $1+\sum_i z^{n_i}$ small, pick $z=-r$ and include only terms that are negative there: the odd powers. With $S=\{1,3,5,\dots,2m-1\}$,

$$P(-r)=1-r-r^3-\cdots-r^{2m-1}\ \longrightarrow\ 1-\frac{r}{1-r^2}\quad(m\to\infty),$$

which vanishes when $1-r^2=r$, i.e. at $r=\rho$. So ρ is the right constant (the roots of $1+z+z^3+\cdots+z^{2m-1}$ approach $-\rho$), and the proof must capture the fact that the coefficients are 0 or 1, never -1.

Proof. Each coefficient of P is 0 or 1. Centre it: $\varepsilon=\frac{1}{2}+\frac{1}{2}\delta$ with $\delta=\pm1$. Concretely, $\frac{1}{1-z}=\sum_{m\ge0}z^m$ for $|z|<1$, and

$$\frac{1}{1-z}-2P(z)=\sum_{m\ge0}\delta_m z^m,\qquad \delta_m=\begin{cases}-1, & m=0 \text{ or } m\in S,\\ +1, & \text{otherwise.}\end{cases}$$

Move the constant term $\delta_0=-1$ to the other side:

$$2P(z)=\left(1+\frac{1}{1-z}\right)-E(z),\qquad E(z)=\sum_{m\ge1}\delta_m z^m,\qquad |E(z)|\le\sum_{m\ge1}r^m=\frac{r}{1-r}.$$

Now bound the main term from below by its real part. Write $z=u+iv$, $u^2+v^2=r^2$. Then $\operatorname{Re}\frac{1}{1-z}=\frac{1-u}{|1-z|^2}=\frac{1-u}{1-2u+r^2}$, and

$$\frac{1-u}{1-2u+r^2}\ge\frac{1}{1+r}\iff(1-u)(1+r)\ge1-2u+r^2\iff r+u(1-r)\ge r^2,$$

which holds because $u\ge-r$ gives $r+u(1-r)\ge r-r(1-r)=r^2$. (Geometrically: $w=1/(1-z)$ maps the circle $|z|=r$ to the circle on diameter $[\frac{1}{1+r},\frac{1}{1-r}]$, whose leftmost point is $\frac{1}{1+r}$.) Hence

$$|2P(z)|\ge\left|1+\frac{1}{1-z}\right|-|E(z)|\ge1+\frac{1}{1+r}-\frac{r}{1-r}=\frac{(1-r^2)+(1-r)-r(1+r)}{1-r^2}=\frac{2(1-r-r^2)}{1-r^2}.$$

For $0\le r<\rho$ we have $r^2+r<1$, so the right side is positive and $P(z)\neq0$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains split P into "1" and "everything else" and bound the rest by $\sum r^m$. The natural attempt compares the rest against 0 as its centre, so its uncertainty per coefficient is 1 (a coefficient in $\{0,1\}$ is at most 1 away from 0). The model compares it against $\frac{1}{2}$, so its uncertainty per coefficient is $\frac{1}{2}$, and the $\frac{1}{2}\sum z^m$ it has taken out is a known function, $\frac{1}{2(1-z)}$, whose real part is positive on the whole disc.

> *The triangle inequality gives me $\frac{1}{2}$, but the extremal case – $z=-r$ with the odd powers – gives ρ. The gap is that my bound lets every coefficient be -1 in effect, and the real coefficients are 0 or 1. So I should measure the coefficients from the midpoint $\frac{1}{2}$: then the unknown part has coefficients $\pm\frac{1}{2}$ and the known part $\frac{1}{2}\sum z^m$ sums to $\frac{1}{2(1-z)}$.*

> *I don't need $|1+\frac{1}{1-z}|$ exactly, only a lower bound – and its real part is at least $1+\frac{1}{1+r}$, attained at $z=-r$, the same point as the extremal example. If the bound and the example agree at the same point, the constant will come out right.*

Layer 3 THE WHY

Why is centring the coefficients the natural move and not a rabbit out of a hat? Start with what the triangle inequality actually bounds. For a power series $\sum c_m z^m$ with each c_m confined to a set K, the crude bound $|\sum_{m\geq 1} c_m z^m| \leq \max_K |c| \cdot \frac{r}{1-r}$ is the best you can say if K is symmetric about 0 – because then the adversary can choose each c_m to align $c_m z^m$ with -1. Here $K = \{0, 1\}$ is not symmetric about 0. It is symmetric about $\frac{1}{2}$. So the honest version of the triangle inequality for this K is: subtract the centre of K from every coefficient, sum the centre's contribution exactly, and apply the triangle inequality only to what is left, which now ranges over a set symmetric about 0 with radius $\frac{1}{2}$. That is the book's "$1/(1-z) - 2P(z)$ has coefficients ± 1" – a factor of 2 applied to avoid halves. The book states it without comment. Here is where it comes from: a coefficient set that is not centred at 0 is being bounded as if it were, and centring it halves the error.

Why must the answer be $\rho = (\sqrt{5}-1)/2$, and why does the model chain reach it with no slack? The extremal example $z = -r$, $S =$ odd numbers, has

$$2P(-r) = \left(1 + \frac{1}{1+r}\right) - \sum_{m\geq 1} r^m = 1 + \frac{1}{1+r} - \frac{r}{1-r},$$

in the limit where S is the set of all odd numbers, because then every term of $E(-r)$ equals $+r^m$: $\delta_m = -1$ exactly on odd m (where $(-r)^m < 0$) and $\delta_m = +1$ on even m (where $(-r)^m > 0$). So in the limit every inequality in the model chain is an equality: $|E|$ attains $\frac{r}{1-r}$, the real-part bound is attained at $u = -r$, and E points exactly against the main term. That is why the final expression $\frac{2(1-r-r^2)}{1-r^2}$ vanishes precisely at ρ: the proof is tight on the extremal configuration. Finding that configuration first is what told you the naive estimate was lossy and where the loss lay.

The golden-ratio constant is thus not decorative. It is the solution of $1 = \frac{r}{1-r^2}$: "the constant term is exactly cancelled by the sum of the odd powers". Once you see the extremal example, the number in the statement is a confirmation, not a mystery.

This is a hard B6 only in the sense that the first estimate every strong student writes down gives the wrong constant, and the problem gives no hint about what to do next. The way out is not a new tool; it is to take the wrong constant seriously and ask which configuration the estimate thinks is worst, and whether that configuration is actually allowed.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the strange constant $(\sqrt{5}-1)/2$. On seeing it, write $\rho^2 + \rho = 1$ and test the most hostile configuration: $z = -r$ with only odd exponents, which gives $1 - \frac{r}{1-r^2}$ and vanishes exactly at ρ.

2. A modulus lower bound $|P| \geq (\text{main term}) - (\text{error})$, with the main term chosen so that the error bound is tight on the extremal configuration.

3. Hesitation occurs when the triangle inequality delivers $\frac{1}{2}$. Reinforce it away by asking what the bound assumes about the coefficients, and comparing that with what the coefficients are actually allowed to be.

4. The trap is the triangle inequality around 0 applied to coefficients in $\{0, 1\}$: it is correct, it proves something, and it gives the wrong constant, so a hurried solver may think the problem's constant is wrong or reach for heavy machinery (Rouché, Eneström–Kakeya) that does not help.

5. Whenever unknown coefficients lie in a set not centred at 0 (such as $\{0, 1\}$, $[0, 1]$ or $\{1, 2\}$), subtract the centre first and bound only the remainder; and whenever a problem gives a specific constant, find the configuration that attains it before trying to prove anything.

6. Reusable rule: a triangle-inequality bound is only as good as the centre you measure from; take out the mean of the allowed values exactly and apply the inequality to the deviation.

Check sharpness by building the extremal case and confirming every inequality in the proof is an equality there.

- ▷ **Cue:** coefficients restricted to $\{0,1\}$ (or any set not centred at 0) $\longrightarrow$ **Move:** write each coefficient as centre plus deviation; sum the centre's contribution in closed form; bound only the deviation
- ▷ **Cue:** an unfamiliar exact constant in a bound, such as $(\sqrt{5}-1)/2$ $\longrightarrow$ **Move:** solve for the configuration that attains it (here $z=-r$ with odd exponents, where $1=\frac{r}{1-r^2}$ rearranges to $r^2+r=1$) before proving anything
- ▷ **Cue:** need a lower bound for $|1+\frac{1}{1-z}|$ on $|z|\le r$ $\longrightarrow$ **Move:** use the real part: $\mathrm{Re}\,\frac{1}{1-z}\ge\frac{1}{1+r}$, the image of the circle being the circle on diameter $[\frac{1}{1+r},\frac{1}{1-r}]$