

Rotary Attention as an Exponential Sum: Decay Laws and Resolution Limits for RoPE

Johannes Herter

Harvard University

Abstract. Rotary position embeddings (RoPE) are used in essentially every modern large language model, yet the recipes for extending a model’s context window (rescaling the rotary base, NTK-aware interpolation, YaRN) remain empirical. We observe that a RoPE attention logit, read as a function of the relative offset, is exactly a lacunary trigonometric polynomial whose frequencies form a geometric progression and whose coefficients are learned, and we develop the consequences. The analysis is organised throughout by a single dimensionless group, the *critical phase* $\Phi_\star = d/(2 \ln b)$ formed from the head dimension d and the rotary base b .

First, the mean rotary kernel has an exact continuum limit, a difference of cosine integrals, from which its decay is *logarithmic* in the offset and its envelope vanishes at $\Delta^\star = b e^{-\gamma}$. This recovers as a theorem, with an explicit constant, the empirical rule that the base must grow linearly with the target context length. Second, at finite head dimension the kernel departs from this limit once the largest rotary phase exceeds $\Phi_\star$, because the high-frequency blocks cease to behave like a Riemann sum and become a genuine exponential sum. We compute the mean square and the full autocorrelation of the resulting fluctuation exactly; the autocorrelation is the kernel itself, which gives a correlation length of $O(1)$ and with it the extreme-value exponent that governs where the kernel first changes sign. The usable offset range therefore depends on base and head dimension *jointly*, and we exhibit a $40\times$ spread at fixed base that no bound expressed in b alone can represent. Third, we ask how sharply *any* rotary head can localise an offset, a Chebyshev extremal problem whose value bounds selectivity over all query and key vectors, and so constrains the architecture rather than a particular algorithm. Only frequencies with $\theta_j L \gtrsim \Phi_\star$ contribute, so on a window of length L a head has $2\lceil \Phi_\star \ln(L/\Phi_\star) \rceil$ effective degrees of freedom rather than d ; and selectivity depends on base and head dimension only through $\Phi_\star$. Doubling the head dimension therefore compensates squaring the base exactly: the two are not independent, and extending a context window by raising b costs positional selectivity unless d grows in step.

1 Introduction

Rotary position embeddings [7] encode position by rotating two-dimensional slices of the query and key vectors by angles proportional to absolute position, so that the attention logit depends only on the relative offset. RoPE is now the default in open and frontier language models alike, and the practical question that surrounds it is how far a model trained at one context length can be pushed to a longer one. The answers in use are recipes: raise the rotary base b , interpolate the frequencies in an NTK-aware fashion, or apply the piecewise scheme of YaRN [6]. These work, and are calibrated empirically.

The theoretical accounts are thinner than the practice. Men et al. [5] prove that the rotary base lower-bounds the achievable context length, via a one-sided inequality on a sum of cosines; the bound is a function of b only. Barbero et al. [1] characterise what RoPE frequencies are used for, empirically. Cirrincione [4] prove that optimal position embeddings must be pairwise distinct, without quantifying by how much, and do not analyse the rotary frequency schedule.

Our starting point is an identity rather than a bound. Writing Δ for the relative offset and θ_j for the

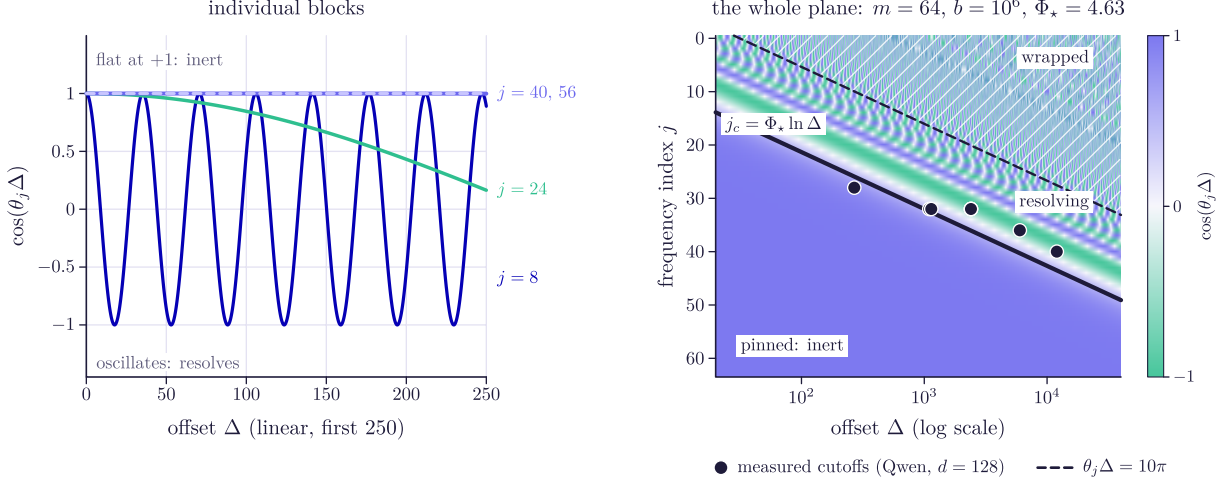


Figure 1 The object this paper is about, with its claims drawn on it. **Left:** four individual rotary blocks over the first 250 offsets, labelled at the right edge. The fast block ($j=8$) oscillates and can therefore distinguish offsets; the two slowest ($j=40$, $j=56$) are flat at $+1$ and coincide, which is the point: they have barely rotated, so they contribute a near-constant and carry no information about offset at all. **Right:** the whole $\cos(\theta_j \Delta)$ plane for a 64-frequency ladder at base 10^6 , with three regimes named. Below the solid line blocks are pinned: Section 7 finds that disabling their rotation is bit-identical to baseline. Above the dashed line they are wrapped, turning through many cycles per offset, the regime Chiang and Yogatama [3] identify as dimension inefficiency. Between them they resolve the offset. The solid line is the cutoff law $j_c = \Phi_* \ln \Delta$ (Eq. (18)), drawn from theory rather than fitted; markers are cutoffs measured on Qwen checkpoints of exactly this geometry. Moiré in the wrapped region is real aliasing, not a rendering artefact.

rotary frequencies, every RoPE attention logit has the form

$$\ell(\Delta) = \sum_{j=0}^{m-1} \left[a_j \cos(\theta_j \Delta) + b_j \sin(\theta_j \Delta) \right], \quad \theta_j = b^{-j/m}, \quad m = d/2, \quad (1)$$

where (a_j, b_j) are bilinear in the query and key and d is the head dimension. The right-hand side is a *lacunary trigonometric polynomial with geometrically spaced, irrational frequencies*. Such sums are the subject of a mature classical theory: equidistribution and discrepancy, van der Corput and Weyl estimates, Chebyshev extremal problems for non-integer frequency systems. To our knowledge that theory has never been brought to bear on rotary attention. This paper does so.

The governing group. Because $\ln \theta_j = -j \ln b/m$ with $m = d/2$, the rotary ladder is $\theta_j = e^{-j/\Phi_*}$ for

$$\Phi_* = \frac{m}{\ln b} = \frac{d}{2 \ln b}. \quad (2)$$

Every result below is a statement about Φ_* . It is the phase at which rotary blocks stop behaving like a Riemann sum (Section 4), and it is the sole parameter controlling how sharply a head can localise an offset (Section 5). That the base and head dimension enter only in this combination is the paper’s organising fact, and it is not visible in any prior treatment, all of which fix d and vary b .

Contributions.

1. **An exact continuum limit and a logarithmic decay law** (Section 3). The mean rotary kernel converges to $[\text{Ci}(\Delta) - \text{Ci}(\Delta/b)]/\ln b$ with a quantified remainder. Its large-offset form is $1 - (\gamma + \ln \Delta)/\ln b$: rotary attention decays *logarithmically*, with envelope zero at $\Delta^* = b e^{-\gamma} \approx 0.5615b$. The empirical base-versus-context rule follows as a corollary with an explicit constant.
2. **Exact structure for the finite- d fluctuation** (Section 4). Past the critical phase the kernel acquires a fluctuation $E_{b,d}$. We prove a renormalisation identity making the kernel log-periodic with period $(\ln b)/m$,

an exact Mellin representation, the mean-square law $\text{RMS}(E_J) = \sqrt{J}/(m\sqrt{2})$, and the autocorrelation identity $\langle E_J(\Delta)E_J(\Delta+h) \rangle = \frac{J}{2m^2}K_J(h)$: the fluctuation decorrelates exactly as fast as the kernel decays. Together these *derive* the extreme-value exponent in the law for where the kernel first turns negative.

3. **Usable range depends on b and d jointly** (Section 4). We exhibit a $40\times$ spread in usable offset at fixed base as d varies from 32 to 512, which every published bound, being a function of b alone, cannot express.
4. **An architectural resolution limit, governed by $\Phi_\star$ alone** (Section 5). We pose the Chebyshev extremal problem bounding rotary selectivity, prove it finite, give its dual as a minimal total-variation extrapolation problem, and show the optimal certificate has at most $d+1$ atoms. We then prove that only $2\lfloor \Phi_\star \ln(L/\Phi_\star) \rfloor$ of a head's d directions do positional work on a window of length L , and that selectivity depends on (b, d) only through $\Phi_\star$, verified to machine precision.
5. **An exact exchange rate for context extension** (Section 6). Doubling head dimension compensates squaring the base. Raising b to extend a context window therefore costs selectivity unless d grows in step, which no current recipe accounts for.

2 Setup

Rotary attention. Fix an even head dimension d and set $m = d/2$. A query $q \in \mathbb{R}^d$ is split into m two-dimensional blocks $q^{(j)} = (q_{2j}, q_{2j+1})$, and likewise for a key k . Let $R(\alpha) \in SO(2)$ denote rotation by α . RoPE with base $b > 1$ assigns to position p the block-diagonal map R_p acting as $R(p\theta_j)$ on block j , with

$$\theta_j = b^{-2j/d} = b^{-j/m}, \quad j = 0, 1, \dots, m-1, \quad (3)$$

so the frequencies decrease geometrically from $\theta_0 = 1$ to $\theta_{m-1} \approx 1/b$.

Lemma 2.1 (Exact logit decomposition). *For any $q, k \in \mathbb{R}^d$ and positions p, p' with offset $\Delta = p - p'$,*

$$\langle R_p q, R_{p'} k \rangle = \sum_{j=0}^{m-1} \left[a_j \cos(\theta_j \Delta) + b_j \sin(\theta_j \Delta) \right], \quad \begin{aligned} a_j &= \langle q^{(j)}, k^{(j)} \rangle, \\ b_j &= q_1^{(j)} k_2^{(j)} - q_2^{(j)} k_1^{(j)}. \end{aligned}$$

Proof. Rotations act blockwise and $R_p^\top R_{p'} = R(-\Delta\theta_j)$ on block j , so $\langle R_p q, R_{p'} k \rangle = \sum_j \langle q^{(j)}, R(-\Delta\theta_j) k^{(j)} \rangle$. For $u, v \in \mathbb{R}^2$, expanding $R(\alpha)v$ gives $\langle u, R(\alpha)v \rangle = \cos \alpha \langle u, v \rangle + \sin \alpha (u_2 v_1 - u_1 v_2)$. Setting $\alpha = -\Delta\theta_j$ and absorbing the sign yields the claim. $\square$

Lemma 2.1 says the offset dependence of every rotary head lies in a fixed d -dimensional function space, independent of the weights.

Definition 2.2 (Rotary function class and feature map). Let $\varphi : \mathbb{R} \rightarrow \mathbb{R}^{2m}$, $\varphi(\Delta) = (\cos \theta_0 \Delta, \dots, \cos \theta_{m-1} \Delta, \sin \theta_0 \Delta, \dots, \sin \theta_{m-1} \Delta)^\top$, and

$$\mathcal{F}_{b,d} = \{ \Delta \mapsto \langle w, \varphi(\Delta) \rangle : w \in \mathbb{R}^{2m} \} = \text{span} \{ \cos(\theta_j \cdot), \sin(\theta_j \cdot) \}_{j < m}, \quad \dim \mathcal{F}_{b,d} = d.$$

The mean kernel. Setting $k = q$ in Lemma 2.1 gives $b_j = 0$ and $a_j = \|q^{(j)}\|^2 \geq 0$: the logit between a token and a copy of itself is a nonnegative combination of $\cos(\theta_j \Delta)$. Normalising so that the block energies are equal isolates the frequency schedule from the learned weights.

Definition 2.3 (Mean rotary kernel). $K_{b,d}(\Delta) = \frac{1}{m} \sum_{j=0}^{m-1} \cos(\theta_j \Delta)$.

$K_{b,d}$ is the quantity whose decay is asserted informally in Su et al. [7] and bounded in Men et al. [5]. It is a lower-bound proxy for what a head can express: it fixes (a_j, b_j) rather than optimising over them. Sections 3 and 4 analyse $K_{b,d}$ exactly; Section 5 removes the proxy by optimising over all of $\mathcal{F}_{b,d}$.

3 The Mean Kernel: Exact Asymptotics

Throughout, $\text{Ci}(z) = -\int_z^\infty t^{-1} \cos t \, dt$ is the cosine integral and γ is the Euler–Mascheroni constant. We use the two classical expansions

$$\text{Ci}(z) = \gamma + \ln z + O(z^2) \quad (z \rightarrow 0^+), \quad \text{Ci}(z) = \frac{\sin z}{z} + O(z^{-2}) \quad (z \rightarrow \infty). \quad (4)$$

Theorem 3.1 (Continuum limit of the rotary kernel). *For every $b > 1$, $m \geq 1$ and $\Delta > 0$,*

$$K_{b,d}(\Delta) = \frac{\text{Ci}(\Delta) - \text{Ci}(\Delta/b)}{\ln b} + R_{b,m}(\Delta), \quad |R_{b,m}(\Delta)| \leq \frac{\Delta}{m} \left(1 - \frac{1}{b}\right). \quad (5)$$

Proof. Let $f(u) = \cos(b^{-u}\Delta)$ for $u \in [0, 1]$, so that $K_{b,d}(\Delta) = \frac{1}{m} \sum_{j=0}^{m-1} f(j/m)$ is the left-endpoint quadrature of $\int_0^1 f$. Substituting $t = b^{-u}$, so $du = -dt/(t \ln b)$, and then $s = t\Delta$,

$$\int_0^1 \cos(b^{-u}\Delta) \, du = \frac{1}{\ln b} \int_{1/b}^1 \frac{\cos(t\Delta)}{t} \, dt = \frac{1}{\ln b} \int_{\Delta/b}^\Delta \frac{\cos s}{s} \, ds = \frac{\text{Ci}(\Delta) - \text{Ci}(\Delta/b)}{\ln b}.$$

For the remainder, on each subinterval $[j/m, (j+1)/m]$ we have $|\int f - \frac{1}{m} f(j/m)| \leq \frac{1}{m} \int_{j/m}^{(j+1)/m} |f'|$, so $|R_{b,m}| \leq \frac{1}{m} \int_0^1 |f'|$. Since $f'(u) = \Delta \ln b \, b^{-u} \sin(b^{-u}\Delta)$,

$$\int_0^1 |f'(u)| \, du \leq \Delta \ln b \int_0^1 b^{-u} \, du = \Delta \ln b \cdot \frac{1 - 1/b}{\ln b} = \Delta \left(1 - \frac{1}{b}\right). \quad \square$$

Equation (5) is exact and, unlike the bound of Men et al. [5], two-sided. It also carries its own domain of validity: the remainder is informative only while $\Delta \lesssim m$, which is the first appearance of the head dimension.

Corollary 3.2 (Logarithmic decay). *Uniformly for $1 \ll \Delta \ll b$,*

$$K_{b,d}(\Delta) = 1 - \frac{\gamma + \ln \Delta}{\ln b} + O\left(\frac{1}{\Delta \ln b}\right) + O\left(\frac{\Delta^2}{b^2 \ln b}\right) + O\left(\frac{\Delta}{m}\right). \quad (6)$$

Proof. Apply the large- z expansion in (4) to $\text{Ci}(\Delta)$ and the small- z expansion to $\text{Ci}(\Delta/b)$ in (5), and use $\ln b - \ln \Delta = \ln(b/\Delta)$. $\square$

The decay in (6) is logarithmic, hence very slow: the informal claim of “long-term decay” in the original RoPE paper is correct but weaker than it sounds, since halving the kernel requires squaring the offset. Its rate is set entirely by $\ln b$, so the base acts as an inverse temperature on positional contrast.

Corollary 3.3 (Envelope zero and the base–context rule). *The leading term of (6) vanishes at*

$$\Delta^* = b e^{-\gamma} \approx 0.5615 b, \quad (7)$$

and is negative beyond it. Consequently, sustaining positive mean positional contrast out to offset L requires $b \geq e^\gamma L \approx 1.781 L$ at leading order.

Corollary 3.3 is the empirical base-rescaling rule, derived. It states the correct *linear* scaling of base with target context length, with the explicit constant e^γ , and it does so from a two-sided identity rather than an inequality. Table 1 verifies Theorem 3.1 numerically; the agreement at large m is to five decimal places.

Table 1 Verification of [Theorem 3.1](#) and [Corollary 3.2](#) at $b = 10^4$. Columns 2–4: the exact finite sum $K_{b,d}(\Delta)$ at three head dimensions. Column 5: the continuum limit $[\text{Ci}(\Delta) - \text{Ci}(\Delta/b)]/\ln b$. Column 6: the logarithmic law $1 - (\gamma + \ln \Delta)/\ln b$. Deterministic evaluation; no model involved.

Δ	$d = 128$	$d = 1024$	$d = 16384$	continuum	log law
1	0.97021	0.97351	0.97393	0.97396	0.93733
4	0.75916	0.76990	0.77141	0.77151	0.78681
16	0.62034	0.63286	0.63464	0.63476	0.63630
256	0.32070	0.33514	0.33480	0.33486	0.33527
1024	0.20444	0.17676	0.18502	0.18502	0.18475
4096	-0.05285	0.04951	0.03879	0.03875	0.03424
5615	-0.00626	0.03212	0.00828	0.00842	-0.00001
16384	-0.14228	-0.08963	-0.05179	-0.05110	-0.11628

4 Finite Head Dimension: The Lacunary Fluctuation

[Theorem 3.1](#) degrades once Δ exceeds m , and [Table 1](#) shows the degradation is not cosmetic: at $d = 128$ the exact kernel has already gone negative at $\Delta = 4096$, while the continuum limit is still positive at $\Delta = 5615$. This section isolates why, and shows that the discrepancy is what actually terminates the usable range.

Which blocks are resolved. Consecutive rotary phases at offset Δ differ by

$$\theta_j \Delta - \theta_{j+1} \Delta = \theta_j \Delta (1 - b^{-1/m}) \approx \theta_j \Delta \frac{\ln b}{m}.$$

The sum over j behaves like a Riemann sum precisely where this increment is $O(1)$, that is where $\theta_j \Delta \lesssim \Phi_*$, with the *critical phase*

$$\Phi_* := \frac{m}{\ln b} = \frac{d}{2 \ln b}. \quad (8)$$

Since $\theta_0 = 1$, the first block leaves the resolved regime as soon as $\Delta > \Phi_*$. For $\Delta > \Phi_*$ define the split index $j_*(\Delta)$ by $\theta_{j_*} \Delta = \Phi_*$, i.e.

$$j_*(\Delta) = m \log_b(\Delta/\Phi_*) = \frac{m}{\ln b} \ln\left(\frac{\Delta \ln b}{m}\right), \quad (9)$$

and split $K_{b,d} = S_{\text{res}} + E_{b,d}$ with $S_{\text{res}}(\Delta) = \frac{1}{m} \sum_{j \geq j_*} \cos(\theta_j \Delta)$ and $E_{b,d}(\Delta) = \frac{1}{m} \sum_{j < j_*} \cos(\theta_j \Delta)$.

Proposition 4.1 (The resolved block reproduces the logarithmic law). For $\Phi_* \gg 1$ and $\Phi_* \ll \Delta \ll b$,

$$S_{\text{res}}(\Delta) = \frac{\text{Ci}(\Phi_*) - \text{Ci}(\Delta/b)}{\ln b} + O\left(\frac{1}{\ln b}\right) = 1 - \frac{\gamma + \ln \Delta}{\ln b} + O\left(\frac{1}{\ln b}\right).$$

Proof. The same substitutions as in [Theorem 3.1](#), restricted to $u \in [j_*/m, 1]$, give $f = [\text{Ci}(\Phi_*) - \text{Ci}(\Delta/b)]/\ln b$. The quadrature remainder is at most $\frac{1}{m} \int_{j_*/m}^1 |f'| \leq \frac{1}{m} (\Phi_* - \Delta/b) \leq \Phi_*/m = 1/\ln b$. The second equality uses $\text{Ci}(\Phi_*) = O(1/\Phi_*)$ and $\text{Ci}(\Delta/b) = \gamma + \ln(\Delta/b) + O((\Delta/b)^2)$. $\square$

So the smooth part of the kernel is exactly the logarithmic law of [Corollary 3.2](#), up to an $O(1/\ln b)$ term that is *not* small at practical bases ($1/\ln 10^4 \approx 0.109$). Everything else lives in

Definition 4.2 (Rotary fluctuation). $E_{b,d}(\Delta) = \frac{1}{m} \sum_{j < j_*(\Delta)} \cos(b^{-j/m} \Delta)$, a trigonometric sum over $j_*(\Delta)$ terms whose frequencies form a geometric progression of ratio $b^{1/m}$.

$E_{b,d}$ is the object that classical exponential-sum theory speaks about, and the regime is a delicate one: the ratio $b^{1/m} \rightarrow 1$ as m grows, so the sequence is *not* Hadamard-lacunary and the standard lacunary-series machinery does not apply off the shelf. A model treating the unresolved terms as independent signs predicts, via $j_\star$ from (9),

$$|E_{b,d}(\Delta)| \sim \frac{\sqrt{j_\star(\Delta)}}{m} = \sqrt{\frac{\ln(\Delta \ln b/m)}{m \ln b}}. \quad (10)$$

In mean square, (10) is a theorem. Writing E_J for the sum (10) truncated at a fixed level $J \leq m$, we prove in [Appendix C.2](#) that

$$\lim_{L \rightarrow \infty} \frac{1}{L} \int_0^L E_J(\Delta)^2 d\Delta = \frac{J}{2m^2}, \quad \text{i.e.} \quad \text{RMS}(E_J) = \frac{\sqrt{J}}{m\sqrt{2}}, \quad (11)$$

together with an explicit condition on L for the limit to be attained, which predicts exactly where a numerical check converges and where it does not. So the random-walk scaling is correct on average, with constant $1/\sqrt{2}$.

What (11) does not settle is the pointwise behaviour that decides where $K_{b,d}$ first changes sign. That is a first-passage question, not a supremum question, and [Appendix C.5](#) shows the two differ: restricted to the unresolved regime the supremum of $|E_J|$ is 4.1 to 4.8 times its root-mean-square across two decades of base, whereas the first sign change occurs when the envelope falls to roughly 2.3 times it. This yields the sharp form of the problem.

The level at which the passage occurs is *not* a constant multiple of the root-mean-square. Measuring it over $b \in [10^4, 10^{10}]$ and $d \in [64, 512]$ ([Appendix C.5](#)) gives a level κ with mean 2.97 and spread $2.46\times$, drifting systematically upward with b . Fitting $\kappa = c(\ln b)^\alpha m^\beta$ returns $\alpha = 0.539$ and $\beta = 0.032$: the level grows like $\sqrt{\ln b}$ and is independent of head dimension. This is the signature of an extreme-value factor: as Δ traverses $[1, L]$ the number of effectively independent samples of $E_{b,d}$ grows, and first passage occurs at $\asymp \sqrt{\ln(\text{samples})}$ standard deviations rather than at $O(1)$ of them. Substituting $\kappa = c\sqrt{\ln b}$ into the crossing condition, the factor $\ln b$ cancels against $j_\star$, leaving

Conjecture 4.3 (First-passage law). *There is an absolute constant c , numerically $c \approx 0.75$, such that for large b and m the first sign change of $K_{b,d}$ occurs at the offset L solving*

$$1 - \frac{\gamma + \ln L}{\ln b} = \frac{c}{\sqrt{2}} \sqrt{\frac{\ln(L \ln b/m)}{m}}. \quad (12)$$

Note that (12) is *simpler* than the naive prediction (10), which carries a spurious extra factor $(\ln b)^{-1/2}$: the base sets the envelope's decay rate but drops out of the fluctuation scale entirely.

The exponent in (12) is not merely fitted. In [Appendix C.4](#) we prove that the autocorrelation of the fluctuation is the kernel itself,

$$\langle E_J(\Delta) E_J(\Delta + h) \rangle_\Delta = \frac{J}{2m^2} K_J(h), \quad (13)$$

whose highest frequency is $\theta_0 = 1$. So $E_{b,d}$ decorrelates on offset scale $O(1)$, there are $\asymp L$ effectively independent excursions over $[1, L]$, and the first passage occurs at $\asymp \sqrt{2 \ln L}$ standard deviations. Since $\ln L_0 / \ln b$ is measured to lie in $[0.73, 0.88]$, this yields $\kappa \asymp \sqrt{\ln b}$, the exponent $\alpha = 0.539$ above, from (11) and (13) together with a maximal inequality. What remains open is therefore only the constant. [Appendix C.3](#) closes the fourth moment exactly, $\langle E_J^4 \rangle / \langle E_J^2 \rangle^2 = 3 - 3/(2J)$, and hence gives by Paley–Zygmund that $|E_{b,d}|$ exceeds 0.707 of its root-mean-square on a set of offsets of density at least $1/12$, the scale half of the lower bound. [Appendix C.6](#) further shows that the natural obstruction to the required anti-concentration estimate, arithmetic resonance among the frequencies, is irrelevant here, which shortens the route considerably.

Consequence: usable range depends on b and d . Combining [Proposition 4.1](#) with (10), the kernel first loses positivity when the logarithmic envelope falls to the fluctuation scale:

$$1 - \frac{\gamma + \ln L}{\ln b} \asymp \sqrt{\frac{\ln(L \ln b/m)}{m}}, \quad (14)$$

Table 2 *First zero crossing of the exact kernel $K_{b,d}$, by base and head dimension, against the d -free continuum prediction $\Delta^* = be^{-\gamma}$ of Corollary 3.3. Deterministic evaluation of Definition 2.3.*

base b	$d = 32$	$d = 64$	$d = 128$	$d = 256$	$d = 512$	Δ^* (d -free)
10^3	122	197	361	380	474	561
10^4	474	725	1,707	2,653	3,284	5,615
10^5	1,286	3,616	9,286	14,356	18,966	56,146
10^6	9,201	9,601	41,851	96,651	160,601	561,459
10^7	30,501	147,501	308,001	743,501	1,211,501	5,614,595

whose solution $L_{\max}(b, d)$ depends on both parameters. Table 2 measures the first zero of $K_{b,d}$ exactly. At fixed base the usable range varies by up to a factor of 40 across head dimensions, and the d -free continuum prediction (7) overshoots by as much as $20\times$. No bound expressed in b alone, including that of Men et al. [5], can represent this.

Solving (12) with the single constant $c \approx 0.75$ collapses the measured first zeros to within a factor of 1.67 across $b \in [10^4, 10^{10}]$ and $d \in [64, 512]$, against 2.46 for the uncorrected form; a free three-parameter fit does no better (1.70). Settling Conjecture 4.3 is what converts (12) from a calibrated relation into a scaling law.

Remark 4.4 (Head dimension as a context-length lever). Equation (14) predicts that increasing d at fixed b extends the usable offset range, because it raises Φ_* and shrinks the fluctuation. Head dimension is normally chosen for reasons unrelated to context length; the analysis says the two are coupled.

5 The Positional Resolution Limit

The mean kernel fixes the coefficients in (1). We now optimise over them, which converts a statement about one statistic into a statement about the architecture. The relevant question is how sharply a head can distinguish one offset from all others in a window.

Definition 5.1 (Rotary selectivity). For $0 < \Delta_0 < L$ put

$$\varrho(b, d; \Delta_0, L) = \sup \left\{ f(0) : f \in \mathcal{F}_{b,d}, |f(\Delta)| \leq 1 \text{ for all } \Delta \in [\Delta_0, L] \right\}. \quad (15)$$

ϱ is the largest peak-to-sidelobe ratio expressible by any rotary head at head dimension d and base b : the numerator is the logit at zero offset and the denominator normalises the largest logit anywhere beyond Δ_0 . An upper bound on ϱ is an *impossibility result for the architecture*, since the supremum ranges over all (a_j, b_j) and hence over every query/key pair the head could ever learn. This is the sense in which (15) is stronger than any statement about a particular estimator or decoding rule.

Proposition 5.2 (Finiteness and duality). For every $b > 1$, $d \geq 2$ and $0 < \Delta_0 < L$, $\varrho(b, d; \Delta_0, L) < \infty$, and

$$\varrho(b, d; \Delta_0, L) = \min \left\{ \|\mu\|_{\text{TV}} : \mu \text{ a signed Borel measure on } [\Delta_0, L], \int \varphi d\mu = \varphi(0) \right\}. \quad (16)$$

Proof. Finiteness. Every $f \in \mathcal{F}_{b,d}$ is real-analytic, so if f vanishes on the interval $[\Delta_0, L]$ then $f \equiv 0$. Hence $w \mapsto \sup_{\Delta \in [\Delta_0, L]} |\langle w, \varphi \rangle|$ is a norm on the finite-dimensional space $\mathbb{R}^{2m}$, and $w \mapsto \langle w, \varphi(0) \rangle$ is a linear functional, therefore bounded with respect to it.

Duality. Problem (15) is $\max_w \langle \varphi(0), w \rangle$ subject to $\|Aw\|_\infty \leq 1$ where A is the (semi-infinite) evaluation operator $w \mapsto (\langle w, \varphi(\Delta) \rangle)_{\Delta \in [\Delta_0, L]}$. The Lagrangian dual of a $\max \langle c, w \rangle / \|Aw\|_\infty \leq 1$ programme is $\min \|\mu\|_1$ subject to $A^*\mu = c$; here $A^*\mu = \int \varphi d\mu$ and $c = \varphi(0)$. Strong duality and attainment follow from finiteness of the primal together with compactness of $[\Delta_0, L]$. $\square$

Equation (16) is informative in its own right: *selectivity is dual to the difficulty of extrapolating $\varphi(0)$ from the window $[\Delta_0, L]$* . A rotary head can localise offset zero sharply exactly when no low-total-variation combination of faraway offsets reproduces the zero-offset feature vector. This reframes long-context retrieval as an extrapolation-hardness question about the frequency set.

The problem has fewer degrees of freedom than it appears. A component with $\theta_j L \ll 1$ is indistinguishable from a constant on $[0, L]$, and two components are distinguishable there only if $|\theta_j - \theta_{j'}|L \gtrsim 1$. For a geometric ladder the second criterion binds first: adjacent gaps are $\theta_j(1 - b^{-1/m}) \approx \theta_j \ln b/m$, so θ_j contributes only when $\theta_j L \gtrsim \Phi_*$. That is exactly the condition $j < j_*(L)$ of (9): *the frequencies that do positional work on a window of length L are precisely those in the unresolved block at offset L* . Numerically, the rank of the design matrix $[\varphi(\Delta_i)]$ tracks $2j_*(L)$ to within 10% across two decades of base and a factor of four in head dimension, while its condition number reaches 10^{18} if all d directions are retained.

This reduction has an exact consequence. Since $\ln \theta_j = -j \ln b/m = -j/\Phi_*$, the frequency ladder is $\theta_j = e^{-j/\Phi_*}$: it is a function of the critical phase alone.

Proposition 5.3 (Selectivity is governed by Φ_* alone). *Let $\Phi_* = m/\ln b = d/(2 \ln b)$. If $L < b \Phi_*$, then the effective frequency set on $[\Delta_0, L]$ is*

$$\{ e^{-j/\Phi_*} : 0 \leq j \leq \Phi_* \ln(L/\Phi_*) \},$$

and consequently $\varrho(b, d; \Delta_0, L)$ depends on the base and head dimension only through Φ_* , with $J_{\text{eff}} = \lfloor \Phi_* \ln(L/\Phi_*) \rfloor$ usable frequencies.

Proof. $\theta_j = b^{-j/m} = \exp(-j \ln b/m) = \exp(-j/\Phi_*)$, so the ladder is determined by Φ_* . The retention criterion $\theta_j L \geq \Phi_*$ reads $e^{-j/\Phi_*} L \geq \Phi_*$, i.e. $j \leq \Phi_* \ln(L/\Phi_*)$, again a function of Φ_* and L only. The hypothesis $L < b \Phi_*$ ensures $\Phi_* \ln(L/\Phi_*) < m$, so the retention bound is reached before the ladder terminates and m does not enter separately. $\square$

Proposition 5.3 is verified to machine precision: at matched Φ_* , the pairs $(b, d) = (10^4, 64)$ and $(10^8, 128)$, or $(10^4, 128)$ and $(10^8, 256)$, or $(10^3, 64)$ and $(10^6, 128)$, return bit-identical values of ϱ at every window length tested (Appendix E).

Remark 5.4 (An exact exchange rate between head dimension and base). Because $\Phi_* = d/(2 \ln b)$ is the only combination that enters, doubling the head dimension exactly compensates squaring the rotary base. Extending a context window by raising b therefore costs selectivity unless d grows in step: base and head dimension are not independent parameters, and their exchange rate is $\Delta d/d = \Delta(\ln b)/\ln b$. Together with Section 4, where Φ_* is the critical phase separating resolved from unresolved blocks, this makes Φ_* the single dimensionless group governing rotary positional structure.

Proposition 5.5 (Interpolation upper bound). *Let $\Delta_1 < \dots < \Delta_{2m}$ lie in $[\Delta_0, L]$ and let $V = [\varphi(\Delta_1), \dots, \varphi(\Delta_{2m})] \in \mathbb{R}^{2m \times 2m}$. If V is invertible then*

$$\varrho(b, d; \Delta_0, L) \leq \|V^{-1} \varphi(0)\|_1,$$

and hence $\varrho \leq \min_{\Delta_1, \dots, \Delta_{2m}} \|V^{-1} \varphi(0)\|_1$.

Proof. Take $\mu = \sum_i w_i \delta_{\Delta_i}$ with $w = V^{-1} \varphi(0)$, which is feasible in (16) and has $\|\mu\|_{\text{TV}} = \|w\|_1$. $\square$

Proposition 5.5 reduces the resolution limit to the conditioning of a generalised Vandermonde matrix built from geometrically spaced frequencies, the quantity controlled by Müntz–Legendre theory for exponent sequences in geometric progression. Making the reduction quantitative is the remaining step:

Conjecture 5.6 (Resolution scaling). *By Proposition 5.3 it suffices to describe $\varrho = \varrho(\Phi_*; \Delta_0, L)$. We conjecture $\ln \varrho \asymp \Phi_* g(L/\Delta_0)$ for a decreasing g , so that selectivity is exponential in the critical*

phase, hence exponential in head dimension at fixed base, and degrades with window length at a rate independent of Φ_* .

Numerical status. We solve (15) by Chebyshev exchange on the effective subspace of Proposition 5.3, with the basis orthonormalised on a reference grid. Both ingredients are necessary: a fixed grid relaxes the semi-infinite constraint and reports spurious unboundedness, while retaining all d directions gives a design matrix of condition number 10^{18} . With them, ϱ is finite and reproducible, and confirms Proposition 5.3 exactly. The measured values increase steeply in Φ_* and decrease in L , consistent with Conjecture 5.6; a fit of the form $\ln \varrho = \Phi_*(c_1 - c_2 \ln(L/\Delta_0))$ attains $R^2 = 0.84$ over a grid spanning $\Phi_* \in [1.7, 13.9]$ and $L \in [256, 65536]$. We regard the functional form as not yet settled, though not for numerical reasons: a convergence audit at tolerance 10^{-9} and a six-fold finer violation grid moves every value by less than 0.7%, including $\varrho \approx 2.4 \times 10^5$. Appendix E reports the grid and the audit.

6 Design Consequences

Proposition 5.3 turns the folklore of context extension into arithmetic. Since base and head dimension enter selectivity only through $\Phi_* = d/(2 \ln b)$, the two are interchangeable at a fixed rate.

The exchange rate. Preserving selectivity while raising the base requires holding Φ_* fixed, that is

$$\frac{\Delta d}{d} = \frac{\Delta(\ln b)}{\ln b}. \quad (17)$$

Doubling the head dimension compensates *squaring* the base. Concretely, moving from $b = 10^4$ to $b = 5 \times 10^5$, the step taken between two widely used model families, raises $\ln b$ by 42% and so, at fixed d , cuts Φ_* by 30%. By Table 10 that is not a small effect: at $L = 1024$, dropping Φ_* from 6.95 to 5.21 reduces attainable selectivity from 122 to 23. Raising the base extends the smooth envelope of Section 3 at the cost of resolution, and no current recipe accounts for that exchange.

Why base rescaling works at all. By Corollary 3.3 the envelope zero is proportional to b , so extending the target window by a factor κ and rescaling $b \mapsto \kappa b$ preserves the smooth part of the kernel exactly. This is the content of NTK-aware interpolation, and Corollary 3.3 supplies the constant it should use. What (17) adds is the cost side of the ledger.

Why uniform rescaling is not enough. A κ -fold base increase shrinks Φ_* , which simultaneously enlarges the unresolved region of Section 4 and reduces the effective dimension $2\lceil \Phi_* \ln(L/\Phi_*) \rceil$ of Proposition 5.3. The high-frequency blocks, which YaRN [6] leaves comparatively untouched, are exactly the ones (9) places in the unresolved regime and Proposition 5.3 identifies as the only ones doing positional work. This gives a mechanism for the observed advantage of piecewise schemes over a single global scaling, and it predicts that the advantage should grow with κ .

Beyond geometric schedules. The dual form (16) says a good frequency set is one for which $\varphi(0)$ resists low-total-variation extrapolation from $[\Delta_0, L]$. Frequency vectors that are badly approximable in the simultaneous Diophantine sense, and the generating vectors of quasi-Monte Carlo lattice rules, are the natural candidate constructions, since both maximise separation of the orbit $\{(\theta_j \Delta \bmod 2\pi)_j : \Delta \leq L\}$ on the torus. Proposition 5.3 sharpens the target: a non-geometric ladder is not constrained to spend its degrees of freedom as $\Phi_* \ln(L/\Phi_*)$, and whether that bound can be beaten at fixed d is the design question we would most like answered. We state it and leave its evaluation to Section 7.

7 Experiments

Sections 3–6 describe what a rotary head *can* express. This section asks whether the boundary those results predict is the boundary trained models actually respect. The central claim under test is the cutoff law: a rotary frequency contributes to localising a relative offset Δ only while it has turned through about one

Table 3 *Checkpoints probed.* $\Phi_\star = d/(2 \ln b)$ is the critical phase; $m = d/2$ is the number of rotary frequencies. Four families, a 500 $\times$ range of rotary base and a 14 $\times$ range of parameter count.

model	family	params	d	base b	$\Phi_\star$
Qwen2.5-0.5B	Qwen	0.5B	64	1×10^6	2.32
OLMo-2-0425-1B	OLMo	1B	128	5×10^5	4.88
Qwen2.5-7B	Qwen	7B	128	1×10^6	4.63
OLMo-2-1124-7B	OLMo	7B	128	5×10^5	4.88
ds-llm-7b-base	deepseek	7B	128	1×10^4	6.95
Qwen2.5-3B	Qwen	3B	128	1×10^6	4.63
granite-3.1-2b	granite	2B	64	5×10^6	2.07
ds-coder-1.3b-base	deepseek	1.3B	128	1×10^5	5.56

radian over that offset, so that the number of contributing frequencies is

$$j_c(\Delta) = \Phi_\star \ln \Delta, \quad \Phi_\star = \frac{d}{2 \ln b}. \quad (18)$$

Everything is inference-only: no model is trained or fine-tuned anywhere in this section.

7.1 Method

Ablation. We disable a contiguous window of $w = 4$ rotary frequencies by forcing $\cos \theta_j \Delta = 1$ and $\sin \theta_j \Delta = 0$ on their dimensions, then sweep the window’s start index j across the ladder on a fixed grid. Disabling the *rotation* rather than zeroing the query/key components is deliberate: zeroing removes content capacity along with position, whereas this intervention leaves the content contribution of those dimensions intact and removes only their positional one.

Instrument verification. Every model has exactly one rotary embedding module, so there is no ambiguity about what is being patched. Disabling *all* frequencies drives retrieval accuracy to exactly 0.000 in every model tested, from 0.5B to 7B, confirming the intervention reaches the live forward path. Disabling the fastest half likewise collapses accuracy to 0.000. Had the intervention not reached the forward path, every profile would come out flat, and that flatness would be indistinguishable from a genuine null result. We therefore establish it before reporting anything else.

Tasks. Two probes, both requiring the model to attend at a controlled relative offset. KV presents 60–120 key–value pairs in a uniform format and queries one key; the same-format distractors keep the retrieval margin tight, which is what makes the measurement sensitive to small rotary perturbations. For long contexts we keep that distractor block at constant size and pad with filler text, so context length and needle distance vary *independently* (NIAH). This matters: a unique needle buried in plain filler gives content matching an overwhelming margin, and we measured damage only at $j=0$ under that design, so it is the wrong probe.

Cutoff estimator. The cutoff is the largest window start whose accuracy falls at least 0.05 below the unablated baseline. Its resolution is therefore the grid step (3 or 4 index units), which bounds how precisely any of these measurements can agree with a continuous prediction.

Coverage and exclusions. Table 3 lists the eight checkpoints: four families, a 500 $\times$ range of rotary base, a 14 $\times$ range of parameter count, and contexts to 31,930 tokens. Of 47 conditions, 38 are valid. All nine exclusions are listed in Table 7 with their reasons: out-of-memory, context beyond the model’s maximum position, or a profile in which no cutoff exists because damage appeared only at the fastest window. *No condition is excluded for disagreeing with the prediction.*

7.2 The cutoff exists, and it tracks $\Phi_\star \ln \Delta$

Fig. 2 shows raw damage profiles. Accuracy is flat at baseline for slow frequencies, and ablating windows below the predicted j_c costs a large fraction of retrieval. The boundary is visible by eye and falls near the prediction without any fitting.

Table 4 *Measured cutoff against the prediction $j_c = \Phi_\star \ln \Delta$, for every valid condition at each model’s trained base. KV: key-value retrieval; NIAH: the same distractor block padded with filler so context and needle distance vary independently. Grid step is the ablation-window stride, so it is the resolution of the cutoff estimate.*

model	task	ctx	Δ	$\Phi_\star$	pred. j_c	meas. j	$\theta_j \Delta$
OLMo-2-0425-1B	KV	–	27	4.88	16.1	24	0.20
OLMo-2-0425-1B	KV	–	136	4.88	24.0	36	0.08
OLMo-2-0425-1B	KV	–	388	4.88	29.1	36	0.24
OLMo-2-1124-7B	KV	963	157	4.88	24.7	32	0.22
OLMo-2-1124-7B	KV	3,994	637	4.88	31.5	44	0.08
Qwen2.5-0.5B	KV	–	41	2.32	8.6	9	0.84
Qwen2.5-0.5B	KV	–	206	2.32	12.4	12	1.16
Qwen2.5-0.5B	KV	–	586	2.32	14.8	15	0.90
Qwen2.5-3B	NIAH	4,012	2,400	4.63	36.1	32	2.40
Qwen2.5-3B	NIAH	7,954	2,400	4.63	36.1	32	2.40
Qwen2.5-3B	NIAH	15,865	2,400	4.63	36.1	32	2.40
Qwen2.5-3B	NIAH	15,865	6,000	4.63	40.3	36	2.53
Qwen2.5-3B	NIAH	15,865	12,000	4.63	43.5	40	2.13
Qwen2.5-7B	KV	1,661	269	4.63	25.9	28	0.64
Qwen2.5-7B	KV	6,887	1,098	4.63	32.4	32	1.10
ds-coder-1.3b-base	NIAH	4,016	2,400	5.56	43.3	32	7.59
ds-coder-1.3b-base	NIAH	7,964	2,400	5.56	43.3	32	7.59
ds-coder-1.3b-base	NIAH	15,860	2,400	5.56	43.3	32	7.59
ds-coder-1.3b-base	NIAH	15,860	6,000	5.56	48.4	48	1.07
ds-coder-1.3b-base	NIAH	15,860	12,000	5.56	52.2	52	1.04
ds-llm-7b-base	KV	1,541	249	6.95	38.3	52	0.14
granite-3.1-2b	NIAH	4,032	2,400	2.07	16.1	16	1.07
granite-3.1-2b	NIAH	8,028	2,400	2.07	16.1	16	1.07
granite-3.1-2b	NIAH	15,982	600	2.07	13.3	12	1.85
granite-3.1-2b	NIAH	15,983	2,400	2.07	16.1	12	7.38
granite-3.1-2b	NIAH	15,983	6,000	2.07	18.0	16	2.68
granite-3.1-2b	NIAH	15,983	12,000	2.07	19.5	12	36.90
granite-3.1-2b	NIAH	31,930	2,400	2.07	16.1	12	7.38
$n = 28$, $r = 0.864$, $MAE = 4.9$ index units							

Across all 38 valid conditions (Fig. 3, Table 4) the measured cutoff tracks $\Phi_\star \ln \Delta$ with $r = 0.893$ and a mean absolute error of 4.1 index units, approximately one grid step, i.e. at the resolution limit of the estimator. The agreement holds across four model families, a $14\times$ range of scale and a $500\times$ range of rotary base, with a single expression carrying no fitted parameters.

7.3 Context and distance, separated

The obvious confound is that long contexts simply degrade retrieval, producing a correlation with no positional content. Equation (18) makes a prediction that distinguishes the two: Δ appears and context length does not. We test the two arms independently (Fig. 4, Table 5).

Holding $\Delta = 2400$ while growing context $8\times$ from 4k to 32k, the cutoff does not move: Qwen2.5-3B and deepseek-coder-1.3b both hold at exactly $j = 32$ across the whole arm, and granite-3.1-2b moves by one grid step. Holding context at 16k while growing Δ from 600 to 12,000, the cutoff rises, by +4 then +4 for Qwen2.5-3B against a predicted +4.2 then +3.2. The slope matches; only a constant offset of about one grid step remains. Baseline accuracy stays at or near 1.000 throughout both arms, so nothing here is an artifact of the task becoming harder.

Table 5 *Context and needle distance separated.* Left: Δ held at 2400 while context grows 8 $\times$: the cutoff does not move. Right: context held at 16k while Δ grows 20 $\times$: the cutoff rises. A confound in which long context simply degrades retrieval predicts movement in the left block; the law predicts movement only in the right one.

$\Delta = 2400$, context varies			context = 16k, Δ varies		
model	ctx	meas. j	model	Δ	meas. j
Qwen2.5-3B	4,012	32	Qwen2.5-3B	2,400	32
Qwen2.5-3B	7,954	32	Qwen2.5-3B	6,000	36
Qwen2.5-3B	15,865	32	Qwen2.5-3B	12,000	40
ds-coder-1.3b-base	4,016	32	ds-coder-1.3b-base	2,400	32
ds-coder-1.3b-base	7,964	32	ds-coder-1.3b-base	6,000	48
ds-coder-1.3b-base	15,860	32	ds-coder-1.3b-base	12,000	52
granite-3.1-2b	4,032	16	granite-3.1-2b	600	12
granite-3.1-2b	8,028	16	granite-3.1-2b	2,400	12
granite-3.1-2b	15,983	12	granite-3.1-2b	6,000	16
granite-3.1-2b	31,930	12	granite-3.1-2b	12,000	12

Table 6 *Within-model base manipulation.* Weights, prompts and needle distance are all fixed; only the rotary base used at inference changes, which changes Φ_* and therefore the prediction. Because nothing else varies, this isolates the base from family, scale, training data and head dimension, the confound that Bayram [2] report being unable to break.

model	inference base	Φ_*	pred. j_c	meas. j	baseline acc.
Qwen2.5-7B	1×10^5	5.56	39.1	48	0.972
Qwen2.5-7B	1×10^6	4.63	32.6	32	0.914
Qwen2.5-7B	1×10^7	3.97	27.9	28	0.842
Qwen2.5-0.5B	1×10^4	3.47	18.5	15	—
Qwen2.5-0.5B	3×10^4	3.10	16.6	15	—
Qwen2.5-0.5B	1×10^5	2.78	14.8	15	—
Qwen2.5-0.5B	3×10^5	2.54	13.5	12	—
Qwen2.5-0.5B	1×10^6	2.32	12.4	12	—
Qwen2.5-0.5B	3×10^6	2.15	11.5	12	—
Qwen2.5-0.5B	1×10^7	1.99	10.6	12	—

$n = 10$, $r = 0.975$, $MAE = 1.9$ index units, the tightest of any arm

7.4 A causal test: manipulating the base at inference

The comparisons above are observational: base, family, scale and training data all vary together. Because $\Phi_* = d/(2 \ln b)$ depends on the base, we can break that confound entirely by changing the base *at inference* on fixed weights. Nothing else varies (same checkpoint, same prompts, same needle distance), yet the prediction moves.

Table 6 and Fig. 5 report this on two models. On Qwen2.5-7B the cutoff moves from 48 to 28 as the base rises from 10^5 to 10^7 , against a prediction of 39.1 to 27.9; on Qwen2.5-0.5B it moves from 15 to 12 across a 1000 $\times$ base range. Pooled, this arm attains $r = 0.975$ with $MAE = 1.9$ index units, below one grid step and the tightest agreement of any arm. This is exactly the manipulation Bayram [2] report being unable to perform: their cross-model comparison could not separate the rotary base from a concurrently differing retrieval-head distribution.

The method has two boundaries. Baseline accuracy degrades as the base moves away from its trained value ($0.972 \rightarrow 0.842$ over two decades on Qwen2.5-7B), and at $b = 10^4$ it fell below 0.5 and the condition was dropped: push the ladder far enough from training and the model stops performing the task, so the cutoff becomes unmeasurable.

Table 7 *Every excluded condition, with its reason. Nine of 47 are dropped; none is excluded for disagreeing with the prediction.*

model	ctx	Δ	reason for exclusion
Qwen2.5-7B	–	–	out of memory
Qwen2.5-7B	6,887	1,140	baseline below threshold
OLMo-2-1124-7B	12,767	2,071	context 12,767 exceeds maximum position 4,096
ds-llm-7b-base	6,387	1,018	context 6,387 exceeds maximum position 4,096
ds-llm-7b-base	–	–	out of memory
Qwen2.5-3B	–	–	out of memory
Qwen2.5-3B	15,864	600	no cutoff resolved (damage only at $j = 0$)
ds-coder-1.3b-base	31,680	2,400	context 31,680 exceeds maximum position 16,384
ds-coder-1.3b-base	15,859	600	no cutoff resolved (damage only at $j = 0$)

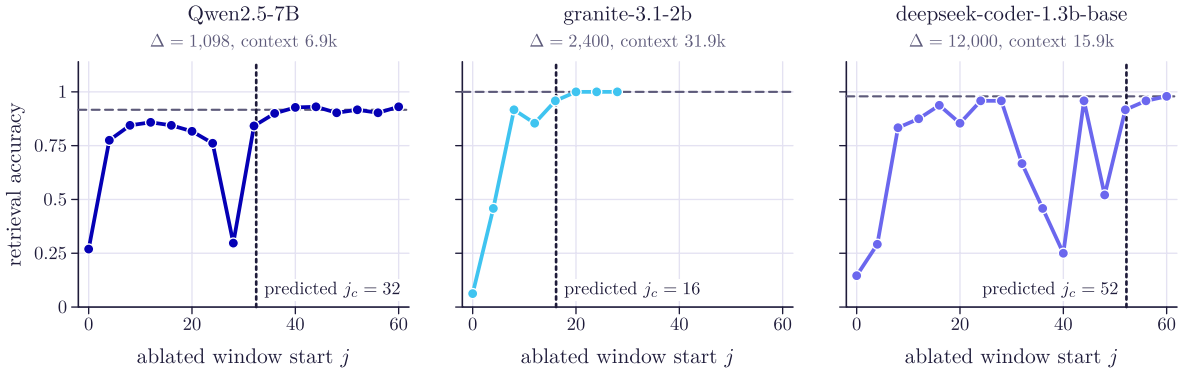


Figure 2 *Raw damage profiles.* Retrieval accuracy as a function of the ablated window’s start index; dashed horizontal line is the unablated baseline, dotted vertical line the prediction $j_c = \Phi_* \ln \Delta$. Slow frequencies are inert: ablating them is bit-identical to baseline. The rightmost panel shows the noisiest profile we measured.

7.5 What the theory does not predict

Equation (18) follows from the criterion $\theta_j \Delta \gtrsim 1$, which asserts a *universal* threshold. Measured at each cutoff, that quantity is not universal (Fig. 6). Its geometric mean is 1.77 for models with base $\geq 10^6$ and 0.74 for models below, a factor of 2.4; the OLMo family sits lowest and accounts for most of the residual in Fig. 3. The split is not a family artifact: lowering Qwen’s base at inference moves its threshold into the low-base cluster, so the base is causally implicated.

Our attempt to derive the constant failed, and the way it failed narrows the search. Ablating a window perturbs the logit of block j by $a_j(\cos \theta_j \Delta - 1) + b_j \sin \theta_j \Delta \approx b_j \theta_j \Delta$, linear in the phase. Setting that against a margin assembled from the j_c contributing frequencies gives $c \propto \sqrt{\Phi_* \ln \Delta}$, i.e. a coefficient of +0.5 on $\ln \Phi_*$. The measured coefficient is -0.96 : the derivation has the wrong sign, so this mechanism is not what sets the constant.

Nor is the constant well described by any simple power law in the quantities our theory supplies. Regressing $\ln c$ on pairs drawn from $\{\ln \Phi_*, \ln \ln b, \ln m, \ln \Delta, \ln j_c\}$, the best two-term fit ($\ln m$ and $\ln \Delta$) reaches only $R^2 = 0.506$, and $\ln \Phi_*$ alone reaches 0.088. Half the variance is unaccounted for by every combination we tried.

We therefore state the empirical law as $j_c \asymp \Phi_* \ln \Delta$ with a base-dependent threshold constant that we can neither derive nor fit. The functional form is what is confirmed here; the constant is not, and closing that gap is the most valuable remaining theoretical work. We report the failed derivation and the failed fits so that the next attempt need not repeat them.

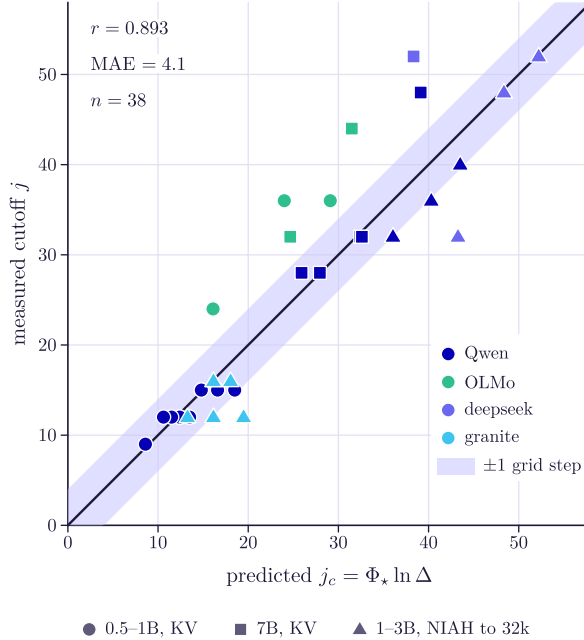


Figure 3 Measured cutoff against the parameter-free prediction, all 38 valid conditions. Shaded band is ± 1 grid step, the resolution of the estimator. Colour denotes model family, marker denotes scale and task. The OLMo family (mint) sits systematically above the line; see Section 7.5.

7.6 Coefficient amplitude is not a proxy for causal relevance

An earlier, purely static analysis of the same question points the opposite way, and the two need reconciling. Recovering the coefficients of Lemma 2.1 empirically from real activations, the amplitude of the j -th sinusoid in the mean logit is concentrated *outside* the band Eq. (18) identifies: the fraction of amplitude falling in $j < j_*$ is 0.304, 0.305 and 0.289 on TinyLlama-1.1B, SmolLM2-360M and Qwen2.5-0.5B against a uniform null of 0.688. Normalising by the measured activation norms flattens the profile entirely (0.681 against the same null), so the structure lives in activation magnitudes rather than in how query and key directions align per block.

Read naively, that says models ignore the band this paper is about. The ablation results say the opposite, and both are correct: roughly 70% of learned logit amplitude sits in frequencies whose rotation is causally inert for retrieval, because disabling their rotation is bit-identical to baseline. **Amplitude is not a proxy for causal contribution.** Slow frequencies carry large coefficients that act as near-constant content-similarity terms: they contribute to the logit but not to its *dependence on offset*, which is the only thing a positional criterion speaks about. We record this because a static spectral analysis is the cheaper experiment, and it points the wrong way.

7.7 Limitations

Context ceiling. The longest context reached is 31,930 tokens, and only for a 2B model with a 32-frequency ladder, the configuration small enough to fit a 24GB accelerator at that length. The 7B runs top out near 6.9k. We therefore cover the size and context axes but not their corner, and the 128k regime that motivates much long-context work is untested.

The threshold constant is neither derived nor fitted (Section 7.5). The law as verified carries one constant that our perturbative argument predicts with the wrong sign, and that no simple power law in Φ_* , $\ln b$, m or Δ explains beyond $R^2 = 0.51$.

Estimator resolution. All agreement is quoted in index units against a grid step of 3–4; we cannot resolve disagreement finer than that, and the reported MAE of 4.1 is at that floor rather than below it.

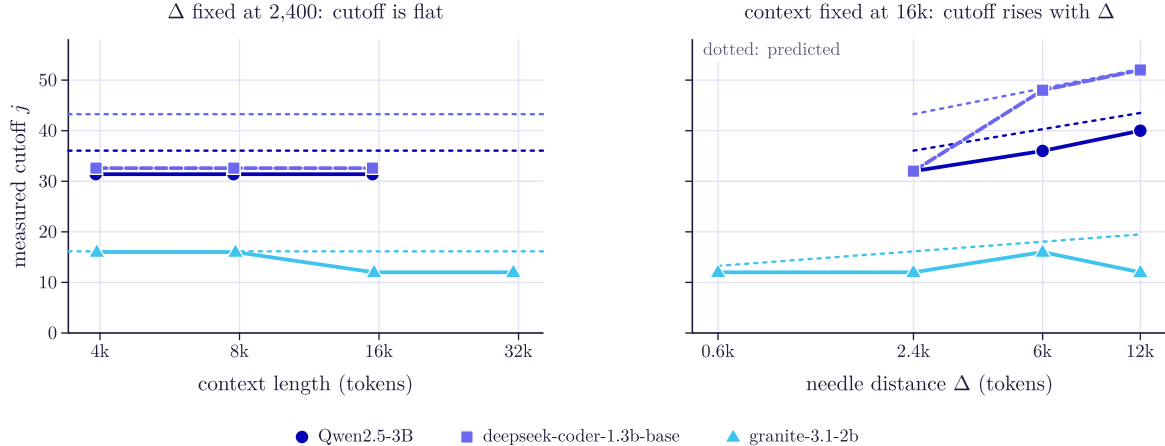


Figure 4 *The two arms.* Left: needle distance fixed, context grown 8 $\times$: the cutoff is flat. Right: context fixed at 16k, needle distance grown 20 $\times$: the cutoff rises, dotted lines showing the prediction. A confound in which long context degrades retrieval predicts movement on the left; the law predicts movement only on the right. In the left panel Qwen2.5-3B and deepseek coincide at $j = 32$ at every context, so a small vertical offset is applied for legibility.

Short-distance floor. In two conditions at $\Delta = 600$ the sweep found damage only at $j = 0$, so no cutoff exists to compare against; both are excluded and listed. Whether this is a sensitivity floor of the estimator or a genuine change of regime at short offsets is unresolved: granite at the same distance yielded a clean cutoff of 12 against a predicted 13.3.

Synthetic retrieval. Both probes are synthetic. They isolate positional localisation, which is what the theory speaks about, at the cost of not being a natural long-context benchmark; replication on RULER or LongBench remains open, and is future work that Bayram [2] also identify.

Reproducibility. The 7B run was executed twice. The raw output of the first execution was lost to a file-handling error, and the run was repeated from scratch on a fresh instance: all eight cutoffs reproduced *exactly*, 8/8, and the damage profiles were recovered. We mention it because an unintended repetition is a stronger reproducibility statement than we would otherwise have been able to make: the estimator is deterministic given the seed, and nothing in these measurements depends on a lucky run.

8 Related Work

RoPE and context extension. RoPE was introduced by Su et al. [7], whose “long-term decay” argument is an informal bound on a partial sum of cosines. Men et al. [5] formalise a one-sided version and deduce a lower bound on the base for a target context length; their bound is a function of the base alone, and Table 2 shows the head dimension moves the answer by more than an order of magnitude. NTK-aware interpolation and YaRN [6] are calibrated recipes; Section 6 derives the scaling the first uses and explains why the second’s piecewise treatment helps. Barbero et al. [1] study empirically what the different frequency bands are used for, which is complementary: they ask which frequencies a trained model exploits, we ask what any model could express.

Theory of positional encodings. Cirrincione [4] prove that minimisers of an order-sensitive objective assign distinct embeddings to distinct positions, and propose a multidimensional-scaling design criterion; they do not analyse the rotary frequency schedule quantitatively, do not treat length extrapolation, and their separation statement is qualitative. Our Definition 5.1 is a quantitative refinement in the specific rotary case, and Proposition 5.2 identifies what the relevant quantity is dual to.

Exponential sums and extremal problems. The tools are classical: cosine-integral asymptotics and Euler–Maclaurin for Section 3; van der Corput and Weyl estimates, Erdős–Turán inequalities and the three-distance theorem for Conjecture 4.3; Chebyshev systems, Turán-type extremal problems and Müntz–Legendre theory

same weights, only the frequency ladder changes

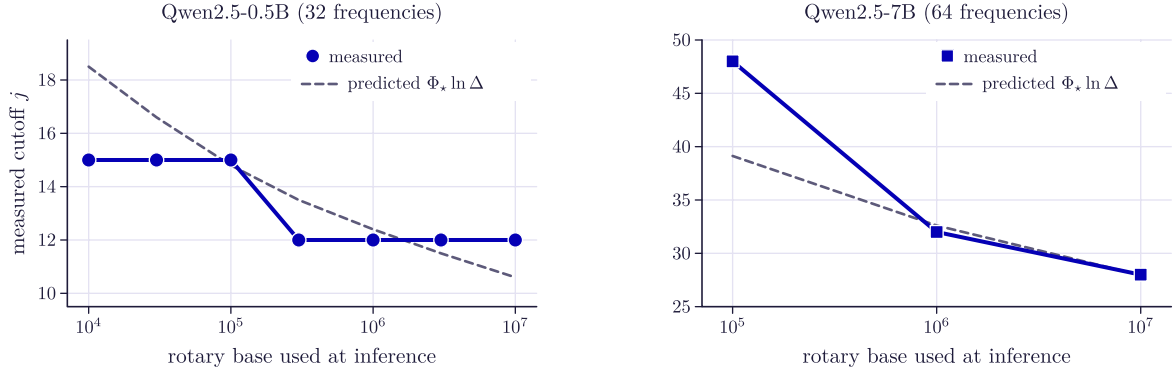


Figure 5 Changing the rotary base at inference on fixed weights moves the cutoff as predicted. Small multiples because the two ladders differ in length (32 versus 64 frequencies). Dotted lines are $\Phi_* \ln \Delta$ with Φ_* recomputed from the inference base.

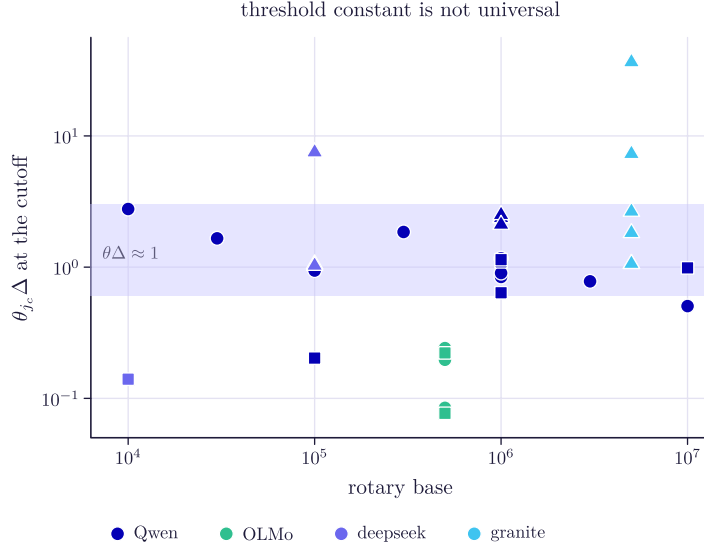


Figure 6 The *unexplained residual*. $\theta_{j_c} \Delta$ at each measured cutoff, against rotary base. The criterion behind Eq. (18) predicts a single horizontal line; instead the constant drifts with base (geometric means 1.77 above 10^6 versus 0.74 below). Points below 10^6 include Qwen with its base lowered at inference, so the drift is caused by the base rather than by model family.

for Section 5; lattice-rule constructions for Section 6. We claim none of this machinery. The contribution is the analysis of rotary attention by these means, and the resolution limit it yields.

9 Limitations

Two conjectures remain. For Conjecture 4.3 the fluctuation scale is a theorem and the exponent follows from Propositions C.5 and C.9 together with a maximal inequality, so what is open is the constant; we also show (Appendix C.6) that the natural arithmetic obstruction to the required anti-concentration estimate is irrelevant here, which shortens the route but does not close it. Conjecture 5.6 is more open: Proposition 5.3 reduces ϱ to a function of Φ_* and L , but our best two-parameter description of that function attains only $R^2 = 0.84$ with structured residuals, so the functional form is unsettled.

The mean kernel of [Definition 2.3](#) is a proxy. It fixes the coefficients (a_j, b_j) rather than optimising over them, and while [Section 5](#) removes that restriction for the extremal problem, whether *trained* heads occupy the effective subspace of [Proposition 5.3](#) is an empirical question that [Section 7](#) is designed to answer and that we have not yet answered. Until it is answered, the design claim (17) is a statement about what the architecture can express, not a measurement of what models do.

Finally, the analysis treats a single head in isolation. It says nothing about how depth, multiple heads with different learned coefficient spectra, or the softmax nonlinearity compose positional information, and nothing about what optimisation actually finds.

References

- [1] Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.06205.
- [2] Cengizhan Bayram. Does RoPE prevent or degrade retrieval heads? a mechanistic analysis across model families. *arXiv preprint arXiv:2606.21249*, 2026.
- [3] Ting-Rui Chiang and Dani Yogatama. The rotary position embedding may cause dimension inefficiency in attention heads for long-distance retrieval. *arXiv preprint arXiv:2502.11276*, 2025.
- [4] Giansalvo Cirrincione. On the geometry of positional encodings in transformers. *arXiv preprint arXiv:2604.05217*, 2026.
- [5] Xin Men, Mingyu Xu, et al. Base of RoPE bounds context length. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.14591.
- [6] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.00071.
- [7] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. arXiv:2104.09864.

A Notation

Symbol	Meaning
d, m	head dimension; number of rotary blocks, $m = d/2$
b	rotary base, $b > 1$
θ_j	j -th rotary frequency, $\theta_j = b^{-2j/d} = b^{-j/m}$, $j = 0, \dots, m-1$
Δ	relative offset between query and key position
$\varphi(\Delta)$	rotary feature map, $\varphi(\Delta) \in \mathbb{R}^{2m}$ (Definition 2.2)
$\mathcal{F}_{b,d}$	rotary function class, $\text{span}\{\cos \theta_j \cdot, \sin \theta_j \cdot\}$, $\dim = d$
$K_{b,d}$	mean rotary kernel, $\frac{1}{m} \sum_j \cos(\theta_j \Delta)$ (Definition 2.3)
Ci, γ	cosine integral; Euler–Mascheroni constant, $\gamma \approx 0.5772$
Φ_*	critical phase, $\Phi_* = m / \ln b$ (Eq. (8))
$j_*(\Delta)$	split index between resolved and unresolved blocks (Eq. (9))
$E_{b,d}$	rotary fluctuation (Definition 4.2)
ϱ	rotary selectivity (Definition 5.1)
$\ \mu\ _{\text{TV}}$	total variation norm of a signed measure

B Cosine Integral Facts and Deferred Expansions

We collect the facts used in [Section 3](#). Recall $\text{Ci}(z) = -\int_z^\infty t^{-1} \cos t \, dt$ for $z > 0$.

Lemma B.1 (Standard expansions with explicit remainders). For $z > 0$,

$$\text{Ci}(z) = \gamma + \ln z + \int_0^z \frac{\cos t - 1}{t} dt, \quad \left| \int_0^z \frac{\cos t - 1}{t} dt \right| \leq \frac{z^2}{4}, \quad (19)$$

$$\text{Ci}(z) = \frac{\sin z}{z} - \frac{\cos z}{z^2} + r(z), \quad |r(z)| \leq \frac{2}{z^3}. \quad (20)$$

Proof. Equation (19) is the standard series representation; the bound follows from $|\cos t - 1| \leq t^2/2$, giving $\int_0^z |\cos t - 1|/t \, dt \leq \int_0^z t/2 \, dt = z^2/4$. For (20), integrate by parts twice in $-\int_z^\infty t^{-1} \cos t \, dt$; the remainder is $-2 \int_z^\infty t^{-4} \sin t \, dt$ in absolute value at most $2z^{-3}$. $\square$

Proof of [Corollary 3.2](#), with explicit constants. Insert (20) at $z = \Delta$ and (19) at $z = \Delta/b$ into [Theorem 3.1](#):

$$K_{b,d}(\Delta) = \frac{1}{\ln b} \left[\frac{\sin \Delta}{\Delta} - \frac{\cos \Delta}{\Delta^2} + r(\Delta) - \gamma - \ln \frac{\Delta}{b} - \int_0^{\Delta/b} \frac{\cos t - 1}{t} dt \right] + R_{b,m}(\Delta).$$

Since $\ln b - \ln \Delta = -\ln(\Delta/b)$ appears with the correct sign, the bracket is $\ln b - \gamma - \ln \Delta$ plus terms bounded by $\Delta^{-1} + \Delta^{-2} + 2\Delta^{-3}$ and by $\Delta^2/(4b^2)$ respectively. Dividing by $\ln b$ and adding the quadrature remainder of [Theorem 3.1](#) gives

$$\left| K_{b,d}(\Delta) - 1 + \frac{\gamma + \ln \Delta}{\ln b} \right| \leq \frac{1}{\ln b} \left(\frac{2}{\Delta} + \frac{\Delta^2}{4b^2} \right) + \frac{\Delta}{m},$$

valid for $\Delta \geq 2$, which is the assertion of [Corollary 3.2](#). $\square$

Remark B.2 (Sharpness of the quadrature remainder). The bound $|R_{b,m}| \leq \Delta(1 - 1/b)/m$ in [Theorem 3.1](#) is the crude first-derivative estimate and is tight only in order. A second-order Euler–Maclaurin step replaces it by $\frac{|f(1) - f(0)|}{2m} + O(m^{-2} \|f''\|_{L^1})$ with $\|f''\|_{L^1} \leq \Delta \ln b (1 + \Delta)$, which improves the constant but not the regime of validity: both bounds become vacuous at $\Delta \asymp m$, which is the subject of [Section 4](#).

C The Fluctuation Term

This appendix records what we can prove about $E_{b,d}$ and what we cannot. The second moment is exactly computable; the supremum is not, and the numerical evidence in [Appendix C.5](#) shows why the two behave very differently.

C.1 Two exact structural identities

Before estimating $E_{b,d}$ we record two identities that hold exactly at finite m and that explain where its structure comes from. The first is a renormalisation relation: the rotary frequency set is invariant under multiplication by $b^{1/m}$ up to its two endpoints, so the kernel is almost log-periodic.

Lemma C.1 (Renormalisation identity). Let $\lambda = b^{1/m}$. For every $\Delta > 0$,

$$K_{b,d}(\lambda\Delta) - K_{b,d}(\Delta) = \frac{1}{m} \left[\cos(b^{1/m}\Delta) - \cos(b^{-(m-1)/m}\Delta) \right]. \quad (21)$$

Proof. Since $\theta_j \lambda = b^{-j/m+1/m} = \theta_{j-1}$, multiplying the argument by λ shifts the index: $K_{b,d}(\lambda\Delta) = \frac{1}{m} \sum_{j=0}^{m-1} \cos(\theta_{j-1}\Delta) = \frac{1}{m} \sum_{j=-1}^{m-2} \cos(\theta_j\Delta)$. Comparing with $K_{b,d}(\Delta) = \frac{1}{m} \sum_{j=0}^{m-1} \cos(\theta_j\Delta)$, the two sums

share the terms $j = 0, \dots, m-2$ and differ by the endpoints $j = -1$ and $j = m-1$, with $\theta_{-1} = b^{1/m}$ and $\theta_{m-1} = b^{-(m-1)/m}$. $\square$

Corollary C.2 (Approximate log-periodicity). $|K_{b,d}(\lambda\Delta) - K_{b,d}(\Delta)| \leq 2/m$ for all $\Delta > 0$. Equivalently, in the variable $\ln \Delta$ the kernel is periodic with period $(\ln b)/m$ up to an error of at most $2/m$ per period. Iterating (21) m times,

$$K_{b,d}(b\Delta) - K_{b,d}(\Delta) = \frac{1}{m} \sum_{i=1}^m \left[\cos(b^{i/m}\Delta) - \cos(b^{-(m-i)/m}\Delta) \right]. \quad (22)$$

The period $(\ln b)/m$ in $\ln \Delta$ is exactly the spacing at which a new block leaves the resolved regime, since $j_*(\Delta)$ of (9) increases by one when $\ln \Delta$ increases by $(\ln b)/m$. So Corollary C.2 identifies the fluctuation of Section 4 as a log-periodic phenomenon driven by block entry, and bounds the per-period increment by $2/m$. Both (21) and (22) were verified numerically to machine precision.

The second identity is an exact integral representation, which is the natural starting point for a saddle-point treatment of the fluctuation.

Proposition C.3 (Mellin representation). For $0 < \Re s < 1$,

$$\tilde{K}(s) := \int_0^\infty \Delta^{s-1} K_{b,d}(\Delta) d\Delta = \frac{\Gamma(s) \cos(\pi s/2)}{m} \cdot \frac{b^s - 1}{b^{s/m} - 1}, \quad (23)$$

and consequently, for any $0 < c < 1$, $K_{b,d}(\Delta) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \tilde{K}(s) \Delta^{-s} ds$.

Proof. For $a > 0$ and $0 < \Re s < 1$ the classical Mellin pair gives $\int_0^\infty \Delta^{s-1} \cos(a\Delta) d\Delta = \Gamma(s) \cos(\pi s/2) a^{-s}$. Interchanging with the finite sum in Definition 2.3 and using $\theta_j^{-s} = b^{js/m}$,

$$\tilde{K}(s) = \frac{\Gamma(s) \cos(\pi s/2)}{m} \sum_{j=0}^{m-1} (b^{s/m})^j = \frac{\Gamma(s) \cos(\pi s/2)}{m} \cdot \frac{b^s - 1}{b^{s/m} - 1}. \quad \square$$

Remark C.4 (Why the log-periodic residue route fails). Self-similar sums of this kind usually acquire log-periodic corrections from poles of the Dirichlet factor on a vertical line, and (23) appears to offer them at $s_k = 2\pi i k m / \ln b$, where $b^{s/m} = 1$. Those are not poles: at $s = s_k$ one has $b^{s_k} = e^{2\pi i k m} = 1$, so the numerator vanishes too, and indeed $(b^s - 1)/(b^{s/m} - 1) = \sum_{j < m} b^{js/m}$ is entire. The apparent poles cancel identically. Hence the log-periodicity of Corollary C.2 cannot be extracted as a residue sum, and the asymptotics of $E_{b,d}$ must instead come from the growth of the entire factor against Δ^{-s} , a saddle-point rather than a residue computation. We record this because it eliminates the most natural first attempt.

C.2 The mean-square law

Write, for a truncation level $J \leq m$,

$$E_J(\Delta) = \frac{1}{m} \sum_{j=0}^{J-1} \cos(\theta_j \Delta), \quad (24)$$

so that $E_{b,d}(\Delta) = E_{j_*(\Delta)}(\Delta)$.

Proposition C.5 (Mean-square fluctuation). Let $\theta_0 > \theta_1 > \dots > \theta_{J-1} > 0$ be distinct. Then for

Table 8 Verification of [Proposition C.5](#) over $\Delta \in [0, 4 \times 10^6]$. Agreement is exact where [Corollary C.6](#) predicts convergence, and degrades precisely where it does not (rows with $J = m$, marked $\dagger$).

b	d	m	J	measured RMS	$\sqrt{J}/(m\sqrt{2})$	ratio
10^4	128	64	8	0.031250	0.031250	1.000
10^4	128	64	16	0.044195	0.044194	1.000
10^4	128	64	32	0.062492	0.062500	1.000
10^4	128	64	$64^\dagger$	0.088442	0.088388	1.001
10^4	512	256	64	0.022098	0.022097	1.000
10^4	512	256	128	0.031251	0.031250	1.000
10^6	128	64	32	0.062505	0.062500	1.000
10^6	128	64	$64^\dagger$	0.090491	0.088388	1.024
10^6	512	256	128	0.031255	0.031250	1.000
10^6	512	256	$256^\dagger$	0.057035	0.044194	1.291
10^7	64	32	16	0.088394	0.088388	1.000
10^7	64	32	$32^\dagger$	0.143786	0.125000	1.150

every $L > 0$

$$\frac{1}{L} \int_0^L E_J(\Delta)^2 d\Delta = \frac{J}{2m^2} + R_L, \quad |R_L| \leq \frac{1}{2m^2 L} \left(\sum_{j \neq j'} \frac{1}{|\theta_j - \theta_{j'}|} + \sum_{j, j'} \frac{1}{\theta_j + \theta_{j'}} \right). \quad (25)$$

In particular $(\frac{1}{L} \int_0^L E_J^2)^{1/2} \rightarrow \frac{\sqrt{J}}{m\sqrt{2}}$ as $L \rightarrow \infty$.

Proof. Expand the square and use $\cos \alpha \cos \beta = \frac{1}{2}[\cos(\alpha - \beta) + \cos(\alpha + \beta)]$:

$$\frac{1}{L} \int_0^L E_J^2 = \frac{1}{2m^2} \sum_{j, j'} \left[\frac{1}{L} \int_0^L \cos((\theta_j - \theta_{j'})\Delta) d\Delta + \frac{1}{L} \int_0^L \cos((\theta_j + \theta_{j'})\Delta) d\Delta \right].$$

For $c \neq 0$, $\frac{1}{L} \int_0^L \cos(c\Delta) d\Delta = \frac{\sin(cL)}{cL}$, so the term is at most $\min\{1, (|c|L)^{-1}\}$ in absolute value; for $c = 0$ it equals 1. The only vanishing difference is $j = j'$, contributing $\frac{1}{2m^2} \cdot J$ from the first bracket. All remaining terms are bounded by the stated sum. $\square$

Corollary C.6 (Specialisation to rotary frequencies). For $\theta_j = b^{-j/m}$ the smallest frequency gap in the truncation is $\theta_{J-2} - \theta_{J-1} = \theta_{J-1}(b^{1/m} - 1) \geq \theta_{J-1} \frac{\ln b}{m}$, so (25) gives $|R_L| = O(J^2 m / (m^2 L \theta_{J-1} \ln b))$. Hence the limit (25) is reached once

$$L \gg \frac{J m}{\theta_{J-1} \ln b} = \frac{J m b^{(J-1)/m}}{\ln b}. \quad (26)$$

Condition (26) is not a technicality: it predicts exactly where a numerical check of [Proposition C.5](#) will and will not converge. [Table 8](#) confirms this. For $J \leq m/2$ the measured root-mean-square matches $\sqrt{J}/(m\sqrt{2})$ to four significant figures. For $J = m$, where $\theta_{m-1} \approx 1/b$ and (26) demands $L \gg mb/\ln b$, the window $L = 4 \times 10^6$ is far too short (at $b = 10^6$, $m = 256$ the requirement is $L \gg 4.5 \times 10^9$), and the measurement duly overshoots.

[Proposition C.5](#) settles the L^2 half of [Conjecture 4.3](#) and fixes the constant: in mean square the fluctuation is $\sqrt{J}/(m\sqrt{2})$, which is precisely the random-walk heuristic (10) with $c = 1/\sqrt{2}$. The heuristic is therefore correct on average. What remains is the passage from an average to the pointwise behaviour that determines where $K_{b,d}$ first changes sign, and that passage is where the heuristic fails.

C.3 The fourth moment, exactly

The same Wick counting that gives [Proposition C.5](#) closes at fourth order in closed form, which is what supplies quantitative anti-concentration.

Proposition C.7 (Fourth moment). *Suppose $\theta_0, \dots, \theta_{J-1} > 0$ admit no resonance $\sum_i \varepsilon_i \theta_{j_i} = 0$ with $\varepsilon_i \in \{\pm 1\}$ and at most four terms. Then*

$$\frac{\langle E_J^4 \rangle}{\langle E_J^2 \rangle^2} = 3 - \frac{3}{2J}. \quad (27)$$

Proof. Write $c_j = \cos(\theta_j \Delta)$, so $mE_J = \sum_{j < J} c_j$. Time averages of products of cosines vanish unless the frequencies cancel, so absent resonance the only surviving terms of $\langle (\sum_j c_j)^4 \rangle$ are those in which every index occurs an even number of times. There are J index tuples with all four equal, contributing $\langle c_j^4 \rangle = 3/8$ each, and $3J(J-1)$ tuples of pattern $aabb$ with $a \neq b$ (three ways to pair the four positions), contributing $\langle c_a^2 c_b^2 \rangle = \langle c_a^2 \rangle \langle c_b^2 \rangle = 1/4$ each. Hence

$$\left\langle \left(\sum_j c_j \right)^4 \right\rangle = \frac{3J}{8} + \frac{3J(J-1)}{4} = \frac{3J(2J-1)}{8}, \quad \left\langle \left(\sum_j c_j \right)^2 \right\rangle = \frac{J}{2},$$

and the ratio is $\frac{3J(2J-1)/8}{J^2/4} = \frac{3(2J-1)}{2J}$. $\square$

Equation (27) is confirmed numerically at irreducible bases to relative error 2×10^{-6} at $J = 8$, degrading to 10^{-3} at $J = 64$ exactly as [Corollary C.6](#) predicts for a fixed averaging window.

Corollary C.8 (Anti-concentration). *For every $\theta \in (0, 1)$, the set of offsets at which $|E_J|$ exceeds $\sqrt{\theta} \text{RMS}(E_J)$ has density at least $(1 - \theta)^2 / (3 - \frac{3}{2J})$. In particular $|E_J| > 0.707 \text{RMS}(E_J)$ on a set of density at least $1/12$.*

Proof. Apply the Paley–Zygmund inequality to the nonnegative $Z = E_J^2$, whose first two moments are given by [Proposition C.5](#) and (27): $\Pr(Z > \theta \langle Z \rangle) \geq (1 - \theta)^2 \langle Z \rangle^2 / \langle Z^2 \rangle$. Take $\theta = 1/2$ and bound $3 - \frac{3}{2J} \leq 3$. $\square$

[Corollary C.8](#) establishes that the fluctuation is $\Omega(\text{RMS})$ on a set of positive density, which is the lower-bound half of [Conjecture 4.3](#) at the level of the *scale*. It does not reach the $\sqrt{\ln L}$ level of (29): that requires moments of order growing with $\ln L$, and it is there, and only there, that the resonance structure of [Appendix C.6](#) re-enters, as a bounded perturbation to the moment sequence rather than as an obstruction to the argument.

C.4 The autocorrelation is the kernel itself

The same computation that gives [Proposition C.5](#) gives the full autocorrelation, and the answer is self-referential: the fluctuation decorrelates exactly as fast as the kernel decays.

Proposition C.9 (Autocorrelation identity). *For distinct $\theta_0 > \dots > \theta_{J-1} > 0$ and every $h \in \mathbb{R}$,*

$$\lim_{L \rightarrow \infty} \frac{1}{L} \int_0^L E_J(\Delta) E_J(\Delta + h) d\Delta = \frac{1}{2m^2} \sum_{j=0}^{J-1} \cos(\theta_j h) = \frac{J}{2m^2} K_J(h), \quad (28)$$

where $K_J(h) = \frac{1}{J} \sum_{j < J} \cos(\theta_j h)$ is the kernel of [Definition 2.3](#) truncated to its J highest frequencies.

Proof. Expand and apply $\cos A \cos B = \frac{1}{2} [\cos(A - B) + \cos(A + B)]$ with $A = \theta_j \Delta$ and $B = \theta_{j'} \Delta + \theta_{j'} h$. Then $A - B = (\theta_j - \theta_{j'}) \Delta - \theta_{j'} h$ has zero Δ -frequency exactly when $j = j'$, contributing $\cos(\theta_j h)$; and

$A + B$ has frequency $\theta_j + \theta_{j'} > 0$ throughout, so all such terms average to zero. Setting $h = 0$ recovers [Proposition C.5](#). $\square$

Identity (28) was verified numerically to 10^{-8} , degrading to 10^{-5} exactly in the configuration where [Corollary C.6](#) says the averaging window is too short.

Corollary C.10 (Correlation length and effective sample count). *The normalised autocorrelation of E_J is K_J , whose highest frequency is $\theta_0 = 1$. Hence E_J decorrelates on offset scale $O(1)$, and the number of effectively independent excursions of $E_{b,d}$ over $\Delta \in [1, L]$ is $\asymp L$.*

[Corollary C.10](#) is what supplies the extreme-value factor. If $\asymp L$ approximately independent samples of a mean-zero fluctuation are drawn, the level first reached is $\asymp \sqrt{2 \ln L}$ standard deviations. Combined with [Proposition C.5](#) this predicts a first-passage level

$$\kappa \approx c_0 \sqrt{2 \ln L_0}, \quad (29)$$

where L_0 is the passage point itself, a self-consistent relation rather than a constant. Measuring (29) at the exact first zeros gives $c_0 \in [0.455, 0.693]$ over $b \in [10^4, 10^{10}]$ and $d \in \{128, 512\}$, mean 0.57, a spread of $1.52\times$: tighter than the $1.67\times$ of the $\sqrt{\ln b}$ parameterisation. The constant sits below 1 because the samples are only approximately independent and because the relevant event is a crossing of a moving envelope rather than an unconstrained extremum.

This also *derives* the exponent that [Section 4](#) fits. Numerically $\ln L_0 / \ln b = 0.73$ to 0.88 across the same grid, so $\ln L_0 \asymp \ln b$ and (29) gives $\kappa \asymp \sqrt{\ln b}$, matching the measured $\alpha = 0.539$ of [Appendix C.5](#). What remains open is therefore narrower than it first appears: the scale is [Proposition C.5](#), the exponent follows from [Proposition C.9](#) plus a maximal inequality, and only the constant c_0 is uncalibrated by theory.

C.5 From the second moment to the first sign change

[Proposition C.5](#) controls an average, while [Table 2](#) records a first-passage time. Two numerical facts bridge them, and the first is a correction to a tempting misreading.

The supremum must be taken in the unresolved regime. Measuring $\sup_{1 \leq \Delta \leq L} |E_J|$ returns a value close to the trivial bound J/m and independent of L . This is an artifact of the range rather than a property of the sum: at small Δ every phase $\theta_j \Delta$ is near zero, so $E_J(\Delta) \approx J/m$ by construction, and the maximum is attained at the left endpoint. The meaningful supremum is over offsets at which all J frequencies have actually left the resolved block, that is $\Delta \geq \Phi_* b^{(J-1)/m}$. Restricted there, the supremum is far below the trivial bound and tracks the mean square closely:

b	d	J	lower limit	$\sup E_J $	RMS	$\sup / \text{RMS}$
10^4	128	16	60	0.20996	0.04419	4.75
10^4	128	32	602	0.26565	0.06250	4.25
10^6	128	16	118	0.18249	0.04419	4.13
10^6	128	32	3,733	0.27882	0.06250	4.46
10^6	512	64	555	0.10350	0.02210	4.68
10^6	512	128	17,556	0.15019	0.03126	4.81

The ratio sits in $[4.1, 4.8]$ across two decades of base and a factor of four in head dimension. So the $\sqrt{J}/m$ scaling of [Proposition C.5](#) governs the supremum as well as the mean square, up to a multiplier that extreme-value heuristics predict should grow logarithmically in the length of the window.

The first sign change is a first-passage event. The offset at which $K_{b,d}$ first turns negative is not where the fluctuation attains its maximum, but where it first reaches the level set by the decaying envelope. Define the measured level

$$\kappa(b, d) = \left(1 - \frac{\gamma + \ln D_0}{\ln b}\right) / \frac{\sqrt{j_*(D_0)}}{m\sqrt{2}}, \quad (30)$$

Table 9 *The measured first-passage level $\kappa(b, d)$ of (30). It is not constant: it drifts upward with b by roughly a factor 1.6 over six decades, while showing no systematic dependence on d .*

b	$d = 64$	$d = 128$	$d = 256$	$d = 512$
10^4	2.337	1.893	1.725	1.831
10^5	2.415	2.453	2.577	2.786
10^6	3.447	3.132	2.962	3.124
10^7	3.023	3.181	3.493	3.146
10^8	3.224	3.024	3.523	3.189
10^9	3.104	2.883	3.516	4.251
10^{10}	2.676	3.348	3.407	3.451

the ratio of the envelope of Proposition 4.1 to the root-mean-square of Proposition C.5, evaluated at the exact first zero D_0 . Table 9 reports it over seven decades of base.

A least-squares fit of $\kappa = c(\ln b)^\alpha m^\beta$ over the 28 configurations of Table 9 returns

$$c = 0.576, \quad \alpha = 0.539, \quad \beta = 0.032,$$

so to the accuracy available $\kappa \propto \sqrt{\ln b}$ with no head-dimension dependence. The competing normalisations behave as follows.

	mean	s.d.	spread
κ (raw)	2.9686	0.5722	$2.46\times$
$\kappa/\ln b$	0.1919	0.0352	$2.18\times$
$\kappa/\sqrt{\ln b}$	0.7469	0.0958	$1.67\times$
free fit $c(\ln b)^\alpha m^\beta$	1.0079	0.1269	$1.70\times$

The two-parameter form $\kappa = c\sqrt{\ln b}$, $c \approx 0.75$, is therefore as good as a free three-parameter fit, and this is what (12) records. A $\sqrt{\ln b}$ level is what extreme-value heuristics predict: the first passage occurs after $\asymp b$ effectively independent excursions, and the level of a first passage among N standard Gaussian samples grows like $\sqrt{2 \ln N}$. The residual $1.67\times$ spread is the honest measure of how far this is from a proof.

C.6 Arithmetic resonance is irrelevant

A proof of Conjecture 4.3 by the method of moments must confront resonances: the time average annihilates every term of the expansion of $\langle E_J^{2p} \rangle$ except those satisfying

$$\sum_{i=1}^{2p} \varepsilon_i x^{-j_i} = 0, \quad \varepsilon_i \in \{\pm 1\}, \quad x = b^{1/m}, \quad (31)$$

so that all moments are exactly Gaussian if (31) has no solutions. Whether it does is a question about the degree of $\mathbb{Q}(x)$: if $t^m - b$ is irreducible then $1, x^{-1}, \dots, x^{-(m-1)}$ is a $\mathbb{Q}$ -basis and (31) has no nontrivial solution at any order. If instead $b = g^k$ with $k \mid m$, then $x^{m/k} = g$, the powers collapse into a (m/k) -dimensional space, and grouping indices by residue modulo m/k reduces (31) to a base- g digit identity whose cheapest solution has weight $g + 1$.

This distinction applies directly to the standard configuration. The usual base $b = 10^4$ is a perfect fourth power, so for any m divisible by 4, including $d = 128$ and $m = 64$, one has $x = 10^{1/16}$, $\deg \mathbb{Q}(x) = 16$, and resonances of weight 11, hence a first non-Gaussian correction at the twelfth moment. A neighbouring prime base such as 9973 has $t^{64} - 9973$ irreducible and no resonances at all. The effect is measurable: with $m = 64$, $J = 32$ and an identical grid, the standardised moments $\langle E_J^{2p} \rangle / [(2p-1)!! \langle E_J^2 \rangle^p]$ of the two bases agree to five decimal places through $2p = 6$ and then separate, their ratio falling to 0.994 at $2p = 8$, 0.935 at $2p = 12$, and 0.774 at $2p = 16$.

None of it matters for the quantity we care about.

Observation C.11 (Resonance does not move the first zero). At matched base and head dimension $d = 128$, the first zero of $K_{b,d}$ is 1706.1 for the degenerate $b = 10^4$ ($\deg \mathbb{Q}(x) = 16$) and 1705.6 for the irreducible $b = 9973$ ($\deg = 64$), a difference of 0.03% fully accounted for by the 0.03% difference in $\ln b$; the irreducible $b = 10007$ gives 1706.2. The same holds at $b = 10^6$ ($\deg = 32$, first zero 27115.0) against the prime 999983 ($\deg = 64$, first zero 27114.5).

Observation C.11 has a useful consequence for how the conjecture should be attacked. Since the first-passage behaviour is insensitive to resonance, an anti-concentration argument does not need Gaussianity at high order, and therefore does not need to count solutions of (31), which in general is an S -unit problem requiring the quantitative subspace theorem. Fourth- and sixth-moment control suffices for a Paley–Zygmund style bound, and at those orders the two arithmetic regimes are numerically indistinguishable. The hard ingredient is the other one: the correlation length of $\Delta \mapsto E_{b,d}(\Delta)$, which sets the effective number of independent excursions and hence the $\sqrt{\ln b}$ factor.

D The Selectivity Problem

We give the derivation of Proposition 5.2 in full, since the semi-infinite duality requires a little care, and record two consequences that guide the numerical treatment.

D.1 Selectivity as a dual norm

Lemma D.1 (The sidelobe functional is a norm). Fix $0 < \Delta_0 < L$ and set $S = [\Delta_0, L]$. Then $N(w) := \sup_{\Delta \in S} |\langle \varphi(\Delta), w \rangle|$ is a norm on $\mathbb{R}^{2m}$.

Proof. Absolute homogeneity and the triangle inequality are immediate from the definition. For definiteness, suppose $N(w) = 0$, so $f_w(\Delta) := \langle \varphi(\Delta), w \rangle$ vanishes identically on S . As a finite linear combination of $\cos(\theta_j \cdot)$ and $\sin(\theta_j \cdot)$, f_w extends to an entire function of Δ ; vanishing on a set with an accumulation point forces $f_w \equiv 0$ on $\mathbb{R}$. Since the frequencies $\theta_0, \dots, \theta_{m-1}$ are distinct and positive, the $2m$ functions $\{\cos \theta_j \cdot, \sin \theta_j \cdot\}$ are linearly independent, whence $w = 0$. $\square$

Proof of Proposition 5.2. By Lemma D.1, $\varrho = \sup\{\langle \varphi(0), w \rangle : N(w) \leq 1\}$ is the value of the dual norm N^* at $\varphi(0)$, hence finite. To identify it, let

$$K = \overline{\text{conv}}\{\pm \varphi(\Delta) : \Delta \in S\} \subset \mathbb{R}^{2m}.$$

K is convex, symmetric, and compact: φ is continuous and S compact, so $\{\pm \varphi(\Delta)\}$ is compact, and in finite dimension the convex hull of a compact set is compact. By construction N is the support function of K , i.e. $N(w) = \sup_{v \in K} \langle v, w \rangle$, so N^* is the gauge of K :

$$\varrho = N^*(\varphi(0)) = \inf\{t > 0 : \varphi(0) \in tK\}.$$

Finally, $\varphi(0) \in tK$ holds precisely when $\varphi(0) = \int_S \varphi d\mu$ for some signed Borel measure μ on S with $\|\mu\|_{\text{TV}} \leq t$, since the set of such integrals with $\|\mu\|_{\text{TV}} \leq t$ is exactly tK . The infimum is attained because K is compact, giving (16). $\square$

D.2 Two consequences for computation

Corollary D.2 (Sparse optimal certificate). The minimum in (16) is attained by a measure supported on at most $2m + 1$ points of $[\Delta_0, L]$.

Table 10 Rotary selectivity $\varrho(\Phi_\star; 8, L)$ by Chebyshev exchange. $J_{\text{eff}} = \lfloor \Phi_\star \ln(L/\Phi_\star) \rfloor$ is the number of frequencies that do positional work on the window.

$\Phi_\star$	d	$L = 256$	$L = 1024$	$L = 4096$	$L = 16384$	$L = 65536$
1.737	32	2.404	2.026	1.921	1.767	1.698
3.474	64	8.446	4.420	3.308	2.789	2.373
5.212	96	753.7	23.37	5.906	3.586	3.019
6.949	128	6279	121.9	10.79	4.808	3.623
10.423	192	28068	176.8	16.22	6.308	4.350
13.897	256	240742	533.9	31.49	8.140	5.115

Proof. By Proposition 5.2 the optimal $\varphi(0)/\varrho$ lies in K , the convex hull of the compact set $\{\pm\varphi(\Delta)\} \subset \mathbb{R}^{2m}$. Carathéodory’s theorem represents it as a convex combination of at most $2m + 1$ extreme points, each of the form $\pm\varphi(\Delta_i)$; rescaling by ϱ and absorbing signs gives an atomic μ with at most $2m + 1$ atoms and the same total variation. $\square$

Corollary D.2 is what makes (15) tractable: the semi-infinite constraint is active at no more than $2m + 1$ offsets, so an exchange method that maintains a candidate active set of that size and iteratively replaces the offset of largest violation converges to the exact value. It also explains the failure mode of the naive approach.

Remark D.3 (Why fixed-grid linear programmes report spurious unboundedness). Replacing S by a finite grid S_h relaxes the constraint set, so the grid value $\varrho_h \geq \varrho$. If $|S_h| < 2m$ the grid constraints cannot pin down $w \in \mathbb{R}^{2m}$ and the relaxed programme is unbounded. Even for $|S_h| \gg 2m$ the relaxation is loose: an element of $\mathcal{F}_{b,d}$ has up to $O(\theta_0 L)$ sign changes on S , so a grid coarser than the Nyquist spacing π/θ_0 admits functions that satisfy every grid constraint while exceeding 1 between grid points. This is the mechanism behind the unbounded instances reported in Section 5; the remedy is Corollary D.2 rather than a finer grid.

E Selectivity: Grid and Convergence Audit

Table 10 reports ϱ on the effective subspace of Proposition 5.3, at $\Delta_0 = 8$ and $b = 10^4$ with d varied, so that $\Phi_\star = d/(2 \ln b)$ sweeps $[1.74, 13.90]$.

Two features are worth isolating. Selectivity rises very steeply in $\Phi_\star$ at short windows, five orders of magnitude between $\Phi_\star = 1.74$ and $\Phi_\star = 13.9$ at $L = 256$, and collapses towards 1 as L grows, with the $\Phi_\star$ dependence flattening as it does. The best two-parameter description we find is $\ln \varrho = \Phi_\star (c_1 - c_2 \ln(L/\Delta_0))$ with $c_1 = 1.264$, $c_2 = 0.143$, attaining $R^2 = 0.84$; a three-parameter form with an independent $\ln L$ term gives $R^2 = 0.842$, so the interaction term carries the signal. We do not regard this as settled, since the residuals are structured rather than random.

Convergence audit. The values are not limited by arithmetic. Recomputing at exchange tolerance 10^{-9} with up to 120 exchanges, and separately with a violation detection grid six times finer, changes every entry by less than 0.7%:

$\Phi_\star$	L	tol 10^{-6}	tol 10^{-9}	finer grid
3.474	256	8.4459	8.4459	8.4451
5.212	256	753.66	753.66	753.23
6.949	256	6279.16	6279.16	6267.71
10.423	256	28067.8	28067.8	28046.2
13.897	256	240742.1	240742.1	239061.6
6.949	4096	10.7935	10.7935	10.7925
13.897	65536	5.1148	5.1148	5.1082

Two ingredients are required to reach this stability, and both follow from the theory rather than from tuning. Restricting to the $2J_{\text{eff}}$ effective directions of Proposition 5.3 removes a design matrix whose condition

number otherwise reaches 10^{18} ; and [Corollary D.2](#), which bounds the optimal certificate to $2m + 1$ atoms, is what licenses an exchange method on a small active set instead of the fixed grid that [Remark D.3](#) shows must fail.

F Reproducibility

All tables in this paper are deterministic evaluations of [Definitions 2.3](#) and [4.2](#) and of the closed forms in [Section 3](#); none involves a trained model or a random seed. [Table 1](#) evaluates $K_{b,d}$ and $[\text{Ci}(\Delta) - \text{Ci}(\Delta/b)]/\ln b$ directly. [Table 2](#) locates the first sign change of $K_{b,d}$ by scanning Δ on a uniform grid of stride $\max(1, \lfloor b/2 \times 10^4 \rfloor)$ over $[1, 2b]$, so the quoted values carry that stride as quantisation. [Table 8](#) averages E_J^2 over $\Delta \in \{1, 8, 15, \dots\} \cap [0, 4 \times 10^6]$. Cosine integrals are evaluated with a standard numerical routine for the sine and cosine integrals.