REPRESENTABILITY DECIDES WHETHER GROKKING CAN OCCUR

Anonymous authors
Paper under double-blind review

ABSTRACT

A flat test curve has several possible causes. The target may lie outside the architecture's function class, or the training sample may permit memorisation without constraining the population solution, or the optimiser may need a long time to find a generalising representation. Learning curves do not tell these apart, and existing grokking analyses begin only once memorisation is feasible. We separate them before training. Our instrument is an exactly solvable two-layer network with holomorphic activation $\sigma(z) = z^k$ on roots-of-unity-encoded modular inputs. Its expressible class is an algebraically fixed $(k{+}1)$-dimensional subspace of characters of $(\mathbb{Z}_p)^2$, independent of width, and a task is representable exactly when its discrete Fourier support lies in $S_k = \{(\ell, k-\ell) : 0 \le \ell \le k\}$. For linear-phase targets $ma{+}nb \bmod p$ this becomes the arithmetic criterion $m{+}n = k$, an identity in $\mathbb{Z}$ on the canonical residues rather than a congruence modulo p. From the class we extract two quantities computable in advance, the spectral representability defect $\delta_k^2(f)$ and the conditioning of the class restricted to the training sample, and prove two consequences that do not follow from the definitions. The global obstruction reaches the finite sample as a width-independent training-loss floor with an explicit rate, so a non-representable target cannot be memorised even though a flexible model would interpolate it. For representable targets under a well-conditioned restriction, training and population losses are sandwiched together at every iterate, so the prolonged separation that defines grokking is structurally excluded. Across 585 runs the pre-training prediction matches the observed regime with 99.8% accuracy. Two experiments carry the diagnostic beyond the solvable model. A bottlenecked one-hot ReLU network, given no Fourier basis, passes through four regimes as the bottleneck widens. At $d = 4$ test accuracy stops at a ceiling that a fourfold step budget does not lift, and at $d = 8$ the outcome stops being a function of width: six seeds of eight escape after delays spanning two orders of magnitude, and two never reach the chance rate at all. Swapping AdamW for vanilla gradient descent at fixed model and target reverses the degree-dependence of the acquisition delay in all sixteen cells of a four-target grid crossing vanilla gradient descent with three AdamW weight decays, so an observed scaling of grokking time can be optimiser-induced and not a property of the landscape. Which learning regime exists at all is set by the geometry of the effective function class, not by parameter count.

1 INTRODUCTION

Three mechanisms produce the same disappointing learning curve. A network whose test accuracy sits at chance may be unable to represent the target, or able to memorise a training sample that does not constrain the population solution, or on its way to generalising and held up only by optimisation, the regime known as grokking. The three call for different responses: a different architecture, different data, or more training. Training and test curves do not say which is operating (Figure 1), and parameter count does not either, since many parameters can redundantly parameterise a small function class. Song and Ye 2026 derive grokking from a memorisation timescale and a generalisation timescale, both functions of parameter count, with the delay narrowing as capacity grows. That account, like the rest of the grokking literature, presupposes a network able to fit its training data, and leaves open the attribution problem that precedes it. When the target lies outside the architecture's

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

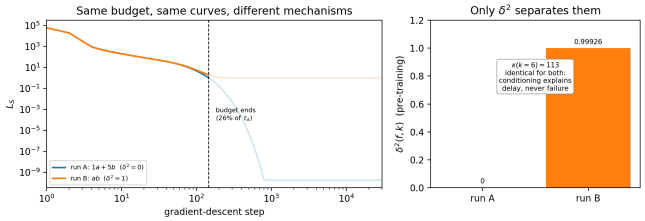

Figure 1: Two mechanisms, one learning curve. Two runs of the same network on different targets follow loss curves that stay within a factor of two under the training budget, yet one is fitting and the other sits on a structural floor. Only the pre-training defect δ^2 separates them. The panels show two of the three mechanisms, and both runs are cut off before either fits the training set, so neither curve is the post-interpolation delay that defines grokking. Protocol in Appendix E.4, general settings in Appendix F.

function class the race between the two timescales never starts, and no memorised solution exists for weight decay to dislodge.

The diagnostic problem. Before asking when a network will grok, we ask whether the absence or delay of generalisation can be attributed, before training, to representational impossibility, to the geometry of the empirical restriction, or to optimisation. Two quantities answer this. The representability defect $\delta_k^2(f) = \|\omega^f - \mathrm{Proj}_{\mathcal{F}_k}\,\omega^f\|_G^2$, written δ^2 below, is the target energy the architecture cannot express. The conditioning (c_S, C_S) of $\mathcal{F}_k$ restricted to the training set measures how much the sample distorts the class it is meant to represent.

The defect $\delta_k^2(f)$ bounds the training loss from below, the conditioning of the empirical restriction bounds the gap between training and population loss, and neither bound follows from the definitions. A universal approximator interpolates finitely many observations without representing the target globally, so for such a model non-representability leaves memorisation available. In this architecture the obstruction reaches the sample. With high probability the training loss is bounded below by $\delta_k^2(f) - \Delta_N$, where $\Delta_N \to 0$ at an explicit rate, and the bound is strictly positive and independent of width (Theorem 6). A universal approximator also memorises before it generalises, so representability leaves room for a train–test gap. Here every iterate satisfies $c_S\, L_G(\theta) \le L_S(\theta) \le C_S\, L_G(\theta)$ whenever the empirical restriction is well conditioned. Training and population loss move together and the separation grokking requires never opens (Theorem 4). Width enters neither statement, adding only redundancy within $\mathcal{F}_k$, and a 256-fold sweep of parameter count leaves the fit and the floor unchanged (Appendix E.3). The two theorems give a dichotomy. Under a well-conditioned empirical restriction a representable target keeps the two losses coupled along the whole trajectory, so the window grokking needs never opens, while a non-representable target sits on a positive floor with neither memorisation nor grokking, so under the fixed encoding the model organism never groks at all and the two extensions below are where grokking enters. Across 585 runs the pre-training prediction agrees with the observed regime in 99.8% of cases.

Contributions.

• **A diagnostic decomposition of learning regimes.** We separate representational failure, memorisation without generalisation, grokking, and gapless generalisation, and locate the boundaries between them in the geometry of the effective function class and not in parameter count (Sections 3–4).

• **Two pre-training diagnostics with theorem-level consequences.** The representability defect yields a width-independent loss floor with an explicit finite-sample rate (Theorem 6). The conditioning of the empirical restriction yields a trajectory-level sandwich that excludes grokking for representable targets (Theorem 4).

• **A standard-network bridge.** A bottlenecked one-hot ReLU network reaches all four regimes under architectural compression alone, so the regime structure belongs to expressivity and not to the holomorphic construction. Quadrupling the budget separates a ceiling from a delay: the $d = 4$ plateau holds, while at $d = 8$ six of eight seeds escape and two stay at chance (Section 5.3).

• **Optimiser versus landscape.** Swapping AdamW for vanilla gradient descent reverses the k-dependence of the acquisition delay at fixed model and target, from $2600 \to 5975$ steps under AdamW to $17475 \to 4650$ under gradient descent, a case where failing to separate the mechanisms gives the wrong causal explanation (Section 5.4).

• **Exact characterisations supporting the diagnostic.** The Fourier support criterion under the fixed encoding, with the linear-phase reduction to $m + n = k$, an encoding-independent rank obstruction under learned embeddings, and a Kuramoto-type reduction with closed-form acquisition time in the phase-only regime (Section 3, with Appendices B.1 and D).

The architecture is a two-layer complex-valued network (Trabelsi et al., 2018) with holomorphic monomial activation $\sigma(z) = z^k$, trained on modular arithmetic tasks under a roots-of-unity encoding (ω^a, ω^b) with $\omega = e^{2\pi i/p}$ and p prime. Since σ is a degree-k polynomial, a binomial expansion gives the output at any hidden width H as a linear combination of the $k + 1$ characters $(a, b) \mapsto \omega^{\ell a+(k-\ell)b}$ for $\ell = 0, \ldots, k$. The expressible class $\mathcal{F}_k$ is a $(k+1)$-dimensional subspace of the p^2-dimensional space of functions on $(\mathbb{Z}_p)^2$, and H adds redundancy within it. We call the architecture *exactly solvable* throughout, meaning that this class has a closed-form characterisation, not that the optimisation dynamics are solved. The architecture is a model organism and not a proposed network. Representational failure never occurs in a universal approximator, so its signature cannot be measured against the alternatives there, and the diagnostic needs a setting where the class is algebraically fixed.

In the Fourier domain $\mathcal{F}_k$ is spanned by the characters with frequency label $(\ell, k - \ell)$, a finite set S_k on the diagonal $u + v \equiv k \pmod p$, so representability becomes a support condition. For the linear-phase targets $ma + nb \bmod p$ that form the core of the grokking literature this reduces to $m + n = k$, an arithmetic test on the target's coefficients. For ab and $a^2 + b^2$ the support spreads across a line or the whole plane through quadratic Gauss sums, so no finite k represents them. The spectral distance to $\mathcal{F}_k$ is then near its maximum, and training accuracy stays at chance throughout.

A ReLU network on the same tasks memorises every target. On the nonlinear-phase ones its test accuracy rises thousands of steps after its training accuracy, the textbook grokking pattern. The targets that make the holomorphic network fail outright make ReLU succeed late. That baseline differs from the holomorphic network in more than the activation (Appendix E.4), which is why Section 5.3 compresses one architecture along a single dimension instead.

2 Setup

The network is two-layer and complex-valued, with forward pass

$$\hat{y}(x;\, W, v) \;=\; \sum_{j=1}^{H} v_j \left(W_{1j}\, x_1 + W_{2j}\, x_2\right)^k, \tag{1}$$

where $x = (x_1, x_2) \in \mathbb{C}^2$, the weights are $W \in \mathbb{C}^{2\times H}$ and $v \in \mathbb{C}^H$, and k is a fixed positive integer. The activation $\sigma(z) = z^k$ is holomorphic, so $\hat{y}$ is a degree-k polynomial for any weights and any hidden width H. Widening adds terms to the sum without raising the degree, which fixes the expressible class algebraically instead of merely keeping it small.

Tasks and encoding. Fix an odd prime p, let $\omega = e^{2\pi i/p}$, and consider tasks $f : \mathbb{Z}_p \times \mathbb{Z}_p \to \mathbb{Z}_p$. The roots-of-unity encoding sends a to ω^a, so the input for (a, b) is $x = (\omega^a, \omega^b)$ and the target is $\omega^{f(a,b)}$. The encoding turns addition modulo p into multiplication of complex numbers, which is what lets the network's polynomial structure meet the task's arithmetic structure. Training minimises $L(W, v) = |S|^{-1} \sum_{(a,b)\in S} \left|\hat{y}(\omega^a, \omega^b; W, v) - \omega^{f(a,b)}\right|^2$ over a training set $S \subseteq \mathbb{Z}_p \times \mathbb{Z}_p$. Output and target both lie on the unit circle, so the loss is a squared chord distance. The target is unimodular while $\hat{y}$ ranges over $\mathbb{C}$ with no normalisation imposed, so this is squared Euclidean error in $\mathbb{C} \cong \mathbb{R}^2$, a chord distance only where $|\hat{y}| = 1$. The architecture *represents* f if some (W, v) gives $\hat{y}(\omega^a, \omega^b;\, W, v) = \omega^{f(a,b)}$ for every (a, b), that is, if the global minimum of the population loss is zero.

Harmonic analysis on $\mathbb{Z}_p \times \mathbb{Z}_p$. The group $G = (\mathbb{Z}_p)^2$ is finite and abelian, so its irreducible characters are $\chi_{u,v}(a, b) = \omega^{ua+vb}$, indexed by $(u, v) \in (\mathbb{Z}_p)^2$. Under the normalised inner

product $\langle f, g\rangle = |G|^{-1} \sum_{(a,b)\in G} f(a,b)\,\overline{g(a,b)}$ they form an orthonormal basis of $\mathbb{C}^G$. We write $\hat{f}(u,v) = \langle f, \chi_{u,v}\rangle$ and $\mathrm{supp}(\hat{f}) = \{(u,v) : \hat{f}(u,v) \neq 0\}$, and $\|f\|^2 = \sum_{u,v} |\hat{f}(u,v)|^2$.

We assume $k < p$ throughout, so that $0, 1, \ldots, k$ are distinct modulo p and the $k+1$ characters appearing in the network's output are distinct basis elements, a condition the representability criterion needs to be sharp. All experiments under the fixed encoding use $p = 97$ and $k \leq 6$, and the p-sweep of Appendix D.7 varies p.

3 The Expressible Class and the First Diagnostic

The construction reduces to the following statement, proved by the binomial theorem, character orthogonality, and a Vandermonde construction in Appendix A.1.

Proposition 1. *Define $\phi_l(a,b) := \chi_{l,\,k-l}(a,b) = \omega^{la+(k-l)b}$ for $l = 0, \ldots, k$, and let $\mathcal{F}_k := \mathrm{span}_{\mathbb{C}}\{\phi_0, \ldots, \phi_k\} \subseteq \mathbb{C}^G$. Then, for every hidden width H and all weights (W, v):*

- *(i) $\hat{y}(x; W, v) = \sum_{l=0}^{k} c_l(W, v)\, x_1^l x_2^{k-l}$ with $c_l(W,v) = \binom{k}{l} \sum_{j=1}^{H} v_j W_{1j}^l W_{2j}^{k-l}$. Under the roots-of-unity encoding every network output therefore lies in $\mathcal{F}_k$.*
- *(ii) For $k < p$ the functions $\phi_0, \ldots, \phi_k$ are distinct irreducible characters of G, hence orthogonal, and $\dim_{\mathbb{C}} \mathcal{F}_k = k+1$, independent of H.*
- *(iii) For $H \geq k+1$ the coefficient map $(W, v) \mapsto (c_0, \ldots, c_k)$ is surjective, so every element of $\mathcal{F}_k$ is the output of some network.*

Parts (i) and (iii) together make $\mathcal{F}_k$ exactly the expressible class, since nothing outside it is produced and all of it is reached as the weights vary. Geometrically this is a bandpass filter on the p^2-dimensional space $\mathbb{C}^G$. The subspace $\mathcal{F}_k$ transmits the $k+1$ frequencies of $S_k = \{(\ell, k-\ell) : 0 \leq \ell \leq k\}$, a finite set of points sitting on the diagonal $u + v \equiv k \pmod p$ without filling it, and blocks everything else. Increasing H neither moves nor widens the pass-band, adding only amplitude redundancy within the transmitted channels. A task f is therefore representable if and only if $\omega^f \in \mathcal{F}_k$, a Fourier-analytic condition decidable before training and independent of optimisation, initialisation, and width.

Theorem 2. *Under the assumption $k < p$ and with $H \geq k+1$, the architecture represents the task $f(a,b) = ma + nb \bmod p$ if and only if there exists $l \in \{0, 1, \ldots, k\}$ such that $m \equiv l \pmod p$ and $n \equiv k - l \pmod p$. For nonnegative integers m, n with $m + n \leq k$, this reduces to the arithmetic condition $m + n = k$, an identity in $\mathbb{Z}$ on the canonical residues and not a congruence modulo p (Remark A.4). (Proof: Appendix A.4.)*

The criterion is sharp in both directions with no continuous gradation, so a linear-phase target either sits in $\mathcal{F}_k$ as a single character or is orthogonal to the whole subspace. Subtraction $f(a,b) = a - b$ shows how strong the condition is. It corresponds to $(m, n) = (1, p-1)$ after reduction modulo p, and $1 + (p-1) \equiv 0 \pmod p$, so representability requires $k \equiv 0 \pmod p$ and the smallest usable degree is $k = p$. At $p = 97$ the task is out of reach for any practical architecture of this type.

For nonlinear-phase tasks the situation is more drastic: under $k < p$, none of ab, $a^2 + b$, $a + b^2$, $a^2 + b^2$ is representable at any degree k (Theorem A.5, stated and proved in Appendix A.5). What blocks them is the shape of the Fourier support, not its size. Representable targets are single points in the frequency plane, while nonlinear-phase targets occupy a line or the whole plane, and no single-line pass-band captures extended support at any degree.

4 The Second Diagnostic and the Two Principal Theorems

The second diagnostic is the conditioning (c_S, C_S) of the expressible class restricted to the training sample. Both principal results follow from the single fact that every network output lies in $\mathcal{F}_k$ (Proposition 1), so both hold pointwise in the parameters and therefore along the entire gradient-descent trajectory, with no optimisation dynamics solved. The linearised fit-time analysis of Appendix C is the one exception. Throughout, $G = (\mathbb{Z}_p)^2$ and the characters $\phi_l(a,b) = \omega^{la+(k-l)b}$, $l = 0, \ldots, k$, form an orthonormal basis of $\mathcal{F}_k$.

4.1 The empirical restriction and its conditioning

For $u, w : G \to \mathbb{C}$ write $\langle u, w\rangle_G = |G|^{-1} \sum_{(a,b)\in G} u\overline{w}$ and $\langle u, w\rangle_S = N^{-1} \sum_{(a,b)\in S} u\overline{w}$, with $\|u\|_G^2 = \langle u, u\rangle_G$ and $\|u\|_S^2 = \langle u, u\rangle_S$, where $S \subseteq G$ is a training set of size N drawn uniformly without replacement. Let $T = \omega^f$ be the target, and for $\theta = (W, v)$ write $g_\theta = \hat{y}(\cdot; \theta) \in \mathcal{F}_k$, $L_S(\theta) = \|g_\theta - T\|_S^2$ and $L_G(\theta) = \|g_\theta - T\|_G^2$. Decompose $T = P + P^\perp$ with $P = \mathrm{Proj}_{\mathcal{F}_k} T$ and $\|P^\perp\|_G^2 = \delta^2$ as in Section 1. The central object is the restriction map $R : \mathcal{F}_k \to \mathbb{C}^S$, $R(g) = (g(a, b))_{(a,b)\in S}$. In the orthonormal basis $\{\phi_l\}$ the empirical form is the $(k+1) \times (k+1)$ Hermitian Gram matrix

$$(M_S)_{lm} = \langle \phi_l, \phi_m\rangle_S = \frac{1}{N} \sum_{(a,b)\in S} \omega^{(l-m)(a-b)}, \tag{2}$$

so that $\|g\|_S^2 = c^\dagger M_S c$ and $\|g\|_G^2 = \|c\|_2^2$ for $g = \sum_l c_l \phi_l$, and we set $c_S := \lambda_{\min}(M_S)$, $C_S := \lambda_{\max}(M_S)$.

The restriction is non-degenerate exactly when the sample realises at least $k + 1$ distinct differences, that is when the difference set $D_S := \{a - b \bmod p : (a, b) \in S\}$ satisfies $|D_S| \geq k + 1$. The precise statement, with the deterministic regime $N > kp$ and the failure probability for $N \leq kp$, is Lemma A.6 in Appendix A.6. The criterion is combinatorial: what matters is the number of *distinct differences* the sample realises, not the number of samples. In the experimental regime it is deterministic and not merely high-probability, since $p = 97$, $k \leq 6$ and $N = 3763 > kp \leq 582$ give injectivity for every draw.

Lemma 3. *The diagonal of M_S is 1, so $E := M_S - I$ has zero diagonal, and for $l \neq m$ (so $0 < |l - m| \leq k < p$) the population mean of $(M_S)_{lm}$ is $\langle \phi_l, \phi_m\rangle_G = 0$. Writing $M_S = \frac{1}{N} \sum_{(a,b)\in S} x_{ab} x_{ab}^\dagger$ with $x_{ab} = (\omega^{l(a-b)})_{l=0}^k$ ($\mathbb{E}[x_{ab} x_{ab}^\dagger] = I$, $\|x_{ab}\|^2 = k + 1$), the matrix Bernstein inequality (Tropp, 2015) (Appendix A.7, transferring to sampling without replacement by Gross & Nesme (2010)) gives, for $N \gtrsim k \log(k/\eta)$, with probability $1 - \eta$, $\|E\|_{\mathrm{op}} \leq \varepsilon_N := O(\sqrt{k \log(k/\eta)/N})$, hence $1 - \varepsilon_N \leq c_S \leq C_S \leq 1 + \varepsilon_N$. In particular $C_S/c_S \to 1$ once $N \gg k \log k$. (An elementary Hoeffding–Serfling argument gives the slightly weaker $\varepsilon_N = O(k\sqrt{\log(k/\eta)/N})$.) (Proof: Appendix A.7.)*

4.2 Gaplessness for representable targets

Theorem 4. *Suppose $T \in \mathcal{F}_k$ (i.e. $\delta^2 = 0$; e.g. a representable linear-phase target $m+n = k$). Then for* every *θ—hence for every gradient-descent iterate $\theta^{(t)}$, regardless of learning rate, momentum, or the nonlinearity of the parametrisation—$c_S L_G(\theta) \leq L_S(\theta) \leq C_S L_G(\theta)$. Equivalently the training and generalisation losses are pinned to a fixed ratio band at every step, $L_G(\theta^{(t)})/L_S(\theta^{(t)}) \in [C_S^{-1}, c_S^{-1}]$.*

Proof. $T \in \mathcal{F}_k$ and $g_\theta \in \mathcal{F}_k$ (Proposition 1) give $g_\theta - T \in \mathcal{F}_k$. Applying the basis identities $\|g\|_S^2 = c^\dagger M_S c$ and $\|g\|_G^2 = \|c\|_2^2$ to $g = g_\theta - T$ yields $c_S \|c\|_2^2 \leq c^\dagger M_S c \leq C_S \|c\|_2^2$. □

The proof uses only that $g_\theta \in \mathcal{F}_k$, a purely expressivity fact that never touches the optimisation dynamics. The statement is therefore robust to the nonlinearity of the coefficient map $c(W, v)$ (Proposition 1(i)), which we could not diagonalise directly.

Non-degeneracy alone excludes an unbounded train–test discrepancy but not a large one. Combined with the concentration of M_S it becomes quantitative, and the two losses coincide up to a factor $1 + O(\sqrt{k \log(k/\eta)/N})$ uniformly over the trajectory, so no memorise-then-generalise window can open at any point (Corollary A.7, Appendix A.8).

Remark 5. In the linear ridge model of Xu et al. (2026), grokking arises from a parameter component in the null space of the feature matrix that the training loss cannot see. Here the analogue, $\ker R|_{\mathcal{F}_k}$, is trivial as soon as the sample realises $k + 1$ distinct differences $a - b$, automatic under $N > kp$ (Lemma A.6), so confinement to $\mathcal{F}_k$ removes the channel that produces grokking there.

4.3 THE LOSS FLOOR FOR NON-REPRESENTABLE TARGETS

Section 3 stated the loss floor informally. We now prove it as an exact finite-dimensional least-squares identity plus concentration, with no uniform convergence over the class and an explicit finite-sample rate that sharpens to $O(1/N)$ for the flat-spectrum targets.

Theorem 6. *Assume $|D_S| \geq k+1$, so that $M_S \succ 0$ (Lemma A.6), which holds deterministically under $N > kp$. Let $(v_S)_l = \langle T, \phi_l \rangle_S$ and v_G be the empirical and population Fourier coefficients of the target, and put $\tau_N := \|v_S - v_G\|_2 = O(\sqrt{k \log(k/\eta)/N})$, with ε_N as in Lemma 3. The minimal empirical loss over the class has the closed form $\Phi(S) := \min_{g \in \mathcal{F}_k} \|g - T\|_S^2 = 1 - v_S^\dagger M_S^{-1} v_S$, with population value $\Phi(G) = 1 - \|v_G\|_2^2 = \delta^2$. Assume $\varepsilon_N \leq \frac{1}{2}$. With probability $1 - \eta$, allocating $\eta/2$ to each of the two concentration events and retaining all cross terms,*

$$\left|\Phi(S) - \delta^2\right| \;\leq\; \left(\|v_G\|_2 + \tau_N\right)^2 \tfrac{\varepsilon_N}{1-\varepsilon_N} \;+\; 2\,\|v_G\|_2\,\tau_N \;+\; \tau_N^2, \tag{3}$$

which is $O(\sqrt{k \log(k/\eta)/N})$ in general and, in the joint proportional-sampling limit $p, N \to \infty$ with $N = \rho p^2$ and ρ fixed, sharpens to $O(1/N)$ for the flat-spectrum targets $\omega^{ab}, \omega^{a^2+b^2}$ and to $O(N^{-3/4})$ for the line-support targets $\omega^{a^2+b}, \omega^{a+b^2}$ (Appendix A.10); at fixed p they do not apply. All bounds are independent of the hidden width H.

Proof sketch. Minimising $\|g - T\|_S^2 = c^\dagger M_S c - 2\Re(c^\dagger v_S) + 1$ over c gives the closed form, using $\|T\|_S^2 = 1$ exactly for a unit-modulus target. The population version sets $M_S = I$, $v_S = v_G$. Splitting $\Phi(S) - \delta^2 = -v_S^\dagger (M_S^{-1} - I) v_S - (\|v_S\|_2^2 - \|v_G\|_2^2)$ and bounding the two terms by Lemma 3 and the Hoeffding–Serfling fluctuation τ_N, each on its own event of failure probability $\eta/2$, yields (3) on their intersection. The cross term $2\,\|v_G\|_2\,\tau_N$ survives a small in-band energy, which is what fixes the rates. Full accounting in Appendix A.10. □

For a given S the floor $1 - v_S^\dagger M_S^{-1} v_S$ can be computed directly. On the training mask used here ($N = 3763$) and for the target ω^{ab} it is 0.9991 at $k = 2$ and 0.9989 at $k = 6$, against a loss of 1 for the zero output. The zero function lies in $\mathcal{F}_k$, so $\Phi(S) \leq 1$ always. The best degree-k fit on the training set is therefore indistinguishable in practice from predicting nothing, improving on it by under 0.1%. The concentration bound of (3) turns this into a statement holding for a random draw at every width, and matches the flat chance-level training curves of every non-representable target.

5 EXPERIMENTS

5.1 PROTOCOL

We use $p = 97$. The task grid is thirty-nine modular functions on $(\mathbb{Z}_{97})^2$: thirty-five linear-phase targets $ma + nb \bmod 97$ with $m, n \geq 0$ and $m + n \in \{1, \ldots, 7\}$, covering representable and non-representable cases at every $k \in \{2, \ldots, 6\}$, and four nonlinear-phase targets ab, $a^2 + b$, $a + b^2$, $a^2 + b^2$. The network has width $H = 128$ and trains full-batch with AdamW on 40% of the 9409 input pairs, three seeds per (k, task) cell, giving 585 runs. Remaining settings are in Appendix F.

All five labels use a 99% accuracy threshold. A run is INST if train and test cross it within 200 steps of each other, GROK if the gap exceeds 500 steps, PART if train crosses it and test settles above chance but below it, MEM if train crosses it and test stays at chance, and FAIL if train never crosses it. The holomorphic network occupies only INST and FAIL. The ReLU baseline adds PART, and MEM appears once the bottleneck sweep of Section 5.3 compresses the feature map.

The theorems concern the regression loss, while these labels use decoded accuracy, the residue c minimising $|\hat{y}(\omega^a, \omega^b) - \omega^c|$. A positive MSE does not by itself force chance-level decoding, since an output that is a small positive multiple of the correct phase decodes perfectly at order-one MSE. The width-independent MSE floor is the theorem, and the chance-level accuracy of the FAIL cells is its empirical counterpart and not a corollary.

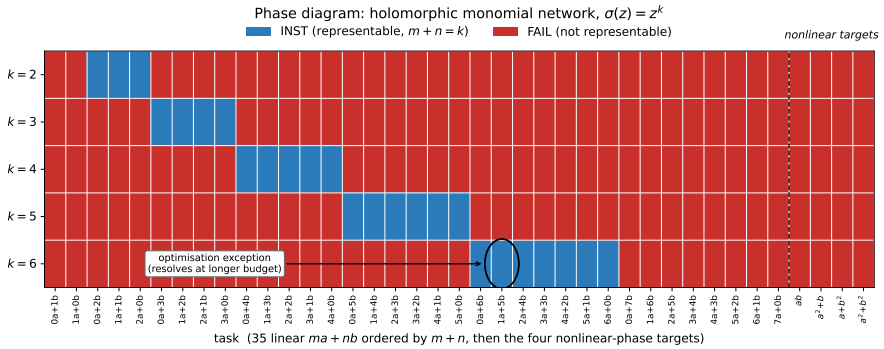

Figure 2: Phase diagram for the holomorphic monomial network across activation degrees k (rows) and the 39 tasks (columns, linear tasks left of the dashed line ordered by $m + n$, nonlinear tasks to the right). Blue cells reach INST and red cells FAIL. The INST cells form the staircase $m + n = k$ predicted by Theorem 2, and all nonlinear tasks fail at every degree, as predicted by Theorem A.5. The circled cell ($k = 6$, $a + 5b$) is the single optimisation exception, which resolves to INST under longer training.

5.2 The algebraic prediction matches the outcome

We apply the pre-training membership test $\omega^f \in \mathcal{F}_k$ to every cell of the grid before any gradient step and compare it against the observed outcome. Figure 2 shows the staircase Theorem 2 predicts: a cell is INST precisely on the diagonal $m + n = k$, FAIL elsewhere in the linear-phase region, and FAIL for every nonlinear-phase target at every degree. Of the 585 runs, 584 match the prediction, a 99.8% agreement rate, with a single discrepancy.

Two features of that tally matter. None of the 510 runs predicted to fail generalises, so the optimiser finds no loophole around the width-independent MSE floor of Theorem 6, and the chance-level decoded accuracy is its empirical counterpart (Section 5.1). What is absent matters as much: under the fixed encoding there are no MEM and no GROK outcomes. The network never memorises a non-representable task, and every run lands in one of the two extreme categories. The small GROK minority that appears once the encoding is learned is treated in Appendix E.6.

The INST rate varies with degree for numerical reasons, not representational ones. For $k \leq 5$ every representable task is learned at all seeds. At $k = 6$ the rate is $20/21$ (Table 4), the exception being $a + 5b$ at one seed, which reaches zero training loss under a longer budget of 8000 steps. The activation z^k has gradients scaling as kz^{k-1}, so higher degrees give more poorly conditioned surfaces and sometimes need more steps to reach the zero-loss solution the theory guarantees.

5.3 Bridging representational failure and grokking with a ReLU bottleneck

The holomorphic network sits at the end of a spectrum where the expressible class is too small for memorisation to occur at all. The capacity literature studies the other end, where the network can always memorise and capacity governs only timing. To test whether the two ends connect, we insert a linear bottleneck of width d after the input layer of a standard one-hot ReLU network on the task $f(a, b) = a + b \bmod 97$, and sweep $d \in \{1, 2, 4, 8, 16, 32, 64, 128\}$ at eight seeds each. The bottleneck caps the rank of the input-to-feature map at d, and every other setting matches the ReLU baseline (Appendix F). Compressing one architecture along a single dimension avoids the confound of activation, encoding and loss that a holomorphic-versus-ReLU comparison carries.

The sweep reaches four regimes. For $d \leq 2$ the network never crosses the 99% training threshold inside the 50000-step budget (FAIL), and its two accuracies track each other: at $d = 2$ the medians are 0.87 on train against 0.76 on test. These runs underfit and generalise what they fit. For $d = 4$ the network crosses the training threshold at a median of $12{,}500$ steps and its test accuracy plateaus at a median of 0.75, between 0.72 and 0.88 across seeds (PART, unanimous). For $d = 8$ it crosses the training threshold at a median of 2000 steps and the seeds split: four sit at chance (0.006 to 0.009 against a chance rate of 0.0103, MEM) and four reach 0.06 to 0.86 (PART). This is the only cell whose

seeds disagree, and the only place in the paper where MEM occurs. For $d \geq 16$ every seed groks, with the median gap falling from 16,600 steps at $d = 16$ to 7500, 5900 and 5400 at $d = 32, 64, 128$, and interquartile ranges that separate $d = 16$ from $d \geq 32$ (Figure 16). Per-cell numbers are in Table 7 and representative curves in Figure 17.

Under this budget test accuracy is not monotone in d: it runs 0.19, 0.76, 0.75, 0.04, then 1.00 from $d = 16$ on (Figure 18). Since the grokking gap grows steeply as d falls toward 16, the dip could be a delay the budget cuts off, so we rerun $d \in \{4, 8\}$ at 200,000 steps (Appendix E.8). The two cells separate. At $d = 4$ the median moves from 0.747 to 0.760 with no seed still climbing, so the plateau is a ceiling, and every seed ends below its own peak. At $d = 8$ the median rises to 0.838: two of the four seeds at chance climb to 0.57 and 0.80, two do not move, and the rest reach 0.88 to 0.95, so most of the dip is an artefact of the budget and what survives is the spread across seeds. The step at which a $d = 8$ seed leaves chance runs from 1250 to 116,750. Neither width reaches the 99% threshold even at 200,000 steps. Fitting quickly therefore does not destroy generalisation by itself, and why it makes the outcome seed-dependent two widths cannot say.

Reaching all four regimes under architectural compression alone places the holomorphic result and the capacity literature on one axis. The FAIL regime is not peculiar to an algebraically fixed class. It appears in a standard architecture once the rank of the feature map falls below what fitting the training set requires. What the sweep adds to the holomorphic picture is that capacity alone does not order the outcomes. Below the grokking widths the sweep gives a ceiling at $d = 4$ that a longer budget does not lift and a seed-dependent outcome at $d = 8$, while the grokking gap grows as capacity falls toward those widths from above.

5.4 THE OPTIMISER SETS THE SIGN OF THE k-DEPENDENCE

Section 5.3 separated representability from optimisation at the level of whether a target is fit. A sharper version of the same confusion concerns when. In the learned-embedding model of Appendix E.6 the acquisition delay grows with the activation degree k, a trend one would naturally read as a property of the loss landscape. We run that model on the same target and change nothing but the optimiser. Medians over eight seeds run $2600, 4125, 4100, 5750, 5975$ steps under AdamW for $k = 2, \ldots, 6$, a $2.3\times$ rise, against $17475, 10400, 7675, 5625, 4650$ under vanilla gradient descent, a monotone fall with no seed censored whose normalised shape moves toward the gradient-flow prediction $\tau \propto 1/k^2$ of Corollary D.16. The curves cross near $k = 5$ (Figure 3), so with every other factor held fixed the choice of optimiser sets the sign of the k-dependence. Adaptive scaling alone does not do it: on the phase-only landscape Adam also decreases with k (Appendix D.6). The reversal needs AdamW together with the parameters that model discards, the output layer with its growing condition number, the free moduli and the weight decay (Conjecture D.23).

Four factorable targets and three weight-decay values leave the sign unchanged in all sixteen cells (Table 8). Writing $\rho := \tau(k{=}6)/\tau(k{=}2)$, every AdamW cell grows ($\rho \in [2.20, 2.61]$) and every gradient-descent cell shrinks ($\rho = 0.27$). The AdamW ratio is flat in the weight decay across three orders of magnitude, which rules out a regularisation effect, and the gradient-descent ratio is identical across targets, as Corollary D.16 requires for factorable targets. The grid is a separate batched run evaluated every 100 steps against every 50 above, which is why the gradient-descent medians share bins and ρ repeats across targets. For this target and weight decay it gives $\rho = 2.54$ against the 2.3 of the run above; the sign, which is what the comparison turns on, is the same in both. Against the predicted $\rho = 0.11$ the measured 0.27 is directional agreement only: gradient descent falls with k as the landscape requires, less steeply than the idealised flow, the gap sitting in the finite width, joint training and free modulus that the phase-only reduction discards.

Had we observed only the AdamW arm, we would have reported a degree-dependent acquisition law and attributed it to the geometry of the loss surface. Nothing in the learning curves supplies the distinction that prevents that misattribution.

Acquisition dynamics under a learned embedding. When the encoding is learned instead of fixed, the phase-only regime admits a slow–fast reduction to a Kuramoto system with a closed-form acquisition time. The reduction, its decoupling regime, and the growth law are developed in Appendix D.

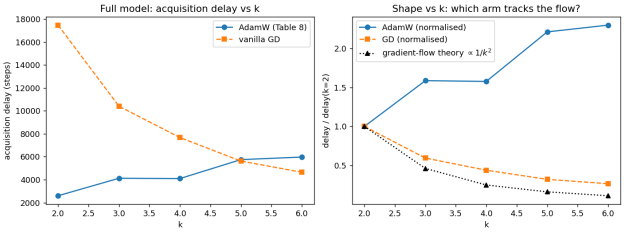

Figure 3: Optimiser versus landscape (learned-embedding model of Appendix E.6, median over 8 seeds, target $1a + 1b$). *Left:* with only the optimiser changed, the AdamW delay grows with k while vanilla gradient descent falls, the curves crossing near $k = 5$. *Right:* normalised to $k = 2$, the gradient-descent arm follows $\tau \propto 1/k^2$ (Corollary D.16) and the AdamW arm moves the other way.

6 Related Work

Grokking and capacity. Grokking was introduced by Power et al. 2022 on small algorithmic datasets, with the delay sensitive to weight decay and training fraction. Subsequent accounts locate the transition in the internal computation (Nanda et al., 2023), in representation-learning phases set by model size and weight decay (Liu et al., 2022), in the move from lazy to rich dynamics (Kumar et al., 2024), in a first-order phase transition between competing basins (Rubin et al., 2023; Cullen et al., 2026), in a competition between memorising and generalising circuits (Varma et al., 2023), and in optimisation alone (Boursier et al., 2025). Žunkovič & Ilievski (2024) derive closed-form grokking times in solvable models, and Humayun et al. 2024 report near-universal grokking of the decision boundary in deep networks. Closest to our setting, Mohamadi et al. 2024 prove sample-complexity separations for quadratic networks on modular addition, and Lyu et al. 2024 induce grokking through early- and late-phase implicit biases. Song and Ye 2026 give the most direct capacity account, separating the delay into two timescales that intersect near the scale where grokking appears. All of these assume a network large enough to memorise the training data. Our work sits in the regime where memorisation is structurally impossible, and Section 5.3 traces the continuous path between the two.

Doshi et al. 2024 hypothesise a learnable/non-learnable classification of modular polynomials of the same form as Theorem 2, supported empirically but left open. Their one-hot inputs with a real power activation form a universal approximator, so failure there means memorisation without generalisation, while here it means the network cannot fit the training data at all. Universal approximation (Cybenko, 1989; Hornik, 1991) and depth separation (Telgarsky, 2016) say nothing about what a fixed architecture expresses exactly, which is what we compute here and use as a pre-training diagnostic. Appendix G discusses each of these lines of work individually.

7 Discussion

Three causes produce the same flat test curve, and the two diagnostics tell them apart before any gradient step. A non-representable target cannot drive the training loss to zero, so the network fails before grokking becomes possible. A representable target under a well-conditioned restriction keeps training and population loss together, so the delay grokking needs never opens. In neither case does the question of when the network will generalise arise.

Both quantities are cheap to compute: (c_S, C_S) is a $(k{+}1) \times (k{+}1)$ eigenvalue computation on the training set, and δ^2 needs the target's Fourier coefficients outside the pass-band. Both presuppose the effective function class, which only an exactly solvable architecture supplies, so estimating them where the class has no closed form is the natural next step.

The theory covers one two-layer architecture on modular arithmetic and relies on the group structure of $(\mathbb{Z}_p)^2$, so deeper networks, other activations and tasks without this symmetry lie outside it. The bottleneck experiments place the regime boundaries in a standard architecture, but locate them empirically and leave open why $d = 8$ depends on the seed. Roots-of-unity inputs also sidestep representation learning, which Appendix E.6 separates out.

Reproducibility statement

All theoretical claims are stated with their assumptions in the main text and proved in full in Appendix A. Assumption D.1 and Remark D.22 state where the acquisition-dynamics results hold and where they do not. The experimental protocol, task grid and run labels are in Section 5.1, and Appendix F gives the initialisation, decoding rule, optimiser settings, step budgets and compute for the phase-diagram grid, the ReLU baseline, the bottleneck sweep, the encoding ablation and the generic-rank experiment. The remaining experiments state their settings where they are reported. Seed counts are given with each experiment, and censoring is recorded where it occurs.

AI use statement

The authors used a large language model as a writing and editing assistant for this manuscript, primarily to condense and refine sections of the prose. All passages and derivations that the model interacted with were carefully reviewed by the authors, who take full responsibility for the final content.

References

James Alexander and André Hirschowitz. Polynomial interpolation in several variables. *Journal of Algebraic Geometry*, 4(2):201–222, 1995.

Etienne Boursier, Scott Pesme, and Radu-Alexandru Dragomir. A theoretical framework for grokking: Interpolation followed by riemannian norm minimisation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.20172.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Pierre Comon, Gene Golub, Lek-Heng Lim, and Bernard Mourrain. Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1254–1279, 2008.

Ben Cullen, Sergio Estan-Ruiz, Riya Danait, and Jiayi Li. A basin-selection perspective on grokking via singular learning theory. *arXiv preprint arXiv:2603.01192*, 2026.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Darshil Doshi, Tianyu He, Aritra Das, and Andrey Gromov. Grokking modular polynomials. *arXiv preprint arXiv:2406.03495*, 2024.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023.

David Gross and Vincent Nesme. Note on sampling without replacing from a finite collection of matrices. *arXiv preprint arXiv:1001.2738*, 2010.

Joe Harris. *Algebraic Geometry: A First Course*, volume 133 of *Graduate Texts in Mathematics*. Springer, New York, 1992.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, 1991.

Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Deep networks always grok and here is why. *arXiv preprint arXiv:2402.15555*, 2024.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. arXiv:1806.07572.

Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 10310–10319, 2019. arXiv:1905.12207.

Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.06110.

Joseph M. Landsberg. *Tensors: Geometry and Applications*, volume 128 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 34651–34663, 2022. arXiv:2205.10343.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon S. Du, Jason D. Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2311.18817.

Seth A. Marvel, Renato E. Mirollo, and Steven H. Strogatz. Identical phase oscillators with global sinusoidal coupling evolve by Möbius group action. *Chaos*, 19(4):043104, 2009.

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica J. Sutherland. Why do you grok? a theoretical analysis of grokking modular addition. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2407.12332.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2301.05217.

Edward Ott and Thomas M. Antonsen. Low dimensional behavior of large systems of globally coupled oscillators. *Chaos*, 18(3):037113, 2008.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Noa Rubin, Inbar Seroussi, and Zohar Ringel. Droplets of good representations: Grokking as a first order phase transition in two layer networks. *arXiv preprint arXiv:2310.03789*, 2023.

Robert J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48, 1974.

Yiding Song and Hanming Ye. Model capacity determines grokking through competing memorisation and generalisation speeds. *arXiv preprint arXiv:2605.09724*, 2026.

Matus Telgarsky. Benefits of depth in neural networks. In *29th Annual Conference on Learning Theory (COLT)*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1517–1539. PMLR, 2016. arXiv:1602.04485.

Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J. Pal. Deep complex networks. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1705.09792.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023.

Shinya Watanabe and Steven H. Strogatz. Constants of motion for superconducting Josephson arrays. *Physica D: Nonlinear Phenomena*, 74(3–4):197–253, 1994.

Mingyue Xu, Gal Vardi, and Itay Safran. To grok grokking: Provable grokking in ridge regression. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2601.19791.

Bojan Žunkovič and Enej Ilievski. Grokking phase transitions in learning local rules with gradient descent. *Journal of Machine Learning Research*, 25(199):1–52, 2024. arXiv:2210.15435.

Appendix contents

A Proofs **13**

A.1 Expanding the forward pass . . . 13
A.2 The expressible class as a character subspace . . . 13
A.3 Every function in $\mathcal{F}_k$ is reachable . . . 14
A.4 Theorem 2 . . . 14
A.5 Theorem A.5 . . . 15
A.6 Lemma A.6 . . . 15
A.7 Lemma 3 . . . 15
A.8 Corollary A.7 . . . 16
A.9 Why non-representability blocks memorisation . . . 17
A.10 Theorem 6 . . . 17

B Beyond the fixed encoding **19**

B.1 The rank barrier beyond the fixed encoding . . . 19
B.2 The frequency-embedding hierarchy . . . 20

C Fit time and the conditioning of the monomial NTK **21**

D Acquisition dynamics under a learned embedding **23**

D.1 The phase-only model . . . 23
D.2 Slow–fast reduction to a Kuramoto system . . . 24
D.3 The single-channel flow is exactly solvable . . . 25
D.4 The escape-time closed form . . . 27
D.5 Decoupling, and where it stops holding . . . 29
D.6 The growth law lives outside the exact regime . . . 30
D.7 Numerical tests of the acquisition dynamics . . . 31

E Additional experiments **33**

E.1 Numerical validation of the dichotomy . . . 33
E.2 Per-degree outcome of the fixed-encoding grid . . . 34
E.3 Parameter count carries no information about the outcome . . . 34
E.4 The expressible class is the mechanism, not the task . . . 35
E.5 Isolating optimisation from representability . . . 37
E.6 Encoding ablation . . . 38

E.7 The rank condition is necessary but not sufficient 41
E.8 Bottleneck sweep 42
E.9 Optimiser reversal 45

F Reproducibility details **45**

G Extended related work **46**

A Proofs

This appendix derives the expressible class and proves the results stated in Sections 3 and 4, restating each main-text result before its proof. We use the notation of Section 4 throughout: the population and empirical Hermitian forms $\langle\cdot,\cdot\rangle_G, \langle\cdot,\cdot\rangle_S$, the orthonormal characters $\{\phi_l\}_{l=0}^{k}$ of $\mathcal{F}_k$, the empirical Gram matrix M_S of (2) with $c_S = \lambda_{\min}(M_S)$, $C_S = \lambda_{\max}(M_S)$, and the target $T = \omega^f$ with $|T| \equiv 1$.

A.1 Expanding the forward pass

The starting point is a direct application of the binomial theorem to the hidden units.

Lemma A.1. *For all $x = (x_1, x_2) \in \mathbb{C}^2$,*

$$\hat{y}(x;\, W, v) \;=\; \sum_{l=0}^{k} c_l(W, v)\, x_1^{l}\, x_2^{k-l}, \qquad c_l(W, v) \;=\; \binom{k}{l} \sum_{j=1}^{H} v_j\, W_{1j}^{l}\, W_{2j}^{k-l}. \tag{4}$$

Proof. For each hidden unit j, the binomial theorem gives

$$(W_{1j}x_1 + W_{2j}x_2)^k = \sum_{l=0}^{k} \binom{k}{l} W_{1j}^{l} W_{2j}^{k-l} x_1^{l} x_2^{k-l}.$$

Summing over j in (1) and interchanging the two finite sums yields (4), with c_l collecting the terms that do not depend on the monomials $x_1^l x_2^{k-l}$. □

The expansion (4) shows that the network output, for any weights and any hidden width, is a linear combination of exactly $k+1$ monomials: $x_1^k, x_1^{k-1}x_2, \ldots, x_2^k$. The coefficients $c_l(W, v)$ are determined by the weights, but the monomials themselves are fixed by the degree. The expressible class is algebraically fixed in the strong sense that no choice of H, W or v produces a monomial of degree other than k, or a function outside the span of these $k+1$ terms.

What the expansion does not yet tell us is what these monomials look like as functions on G under the roots-of-unity encoding. Substituting $x_1 = \omega^a$ and $x_2 = \omega^b$ gives $x_1^l x_2^{k-l} = \omega^{la+(k-l)b}$, which is the character $\chi_{l,\,k-l}$ evaluated at (a, b). The network output thus becomes a linear combination of the $k+1$ functions $\phi_l(a, b) := \chi_{l,\,k-l}(a, b) = \omega^{la+(k-l)b}$, and the expressible class on G is the subspace they span.

A.2 The expressible class as a character subspace

Lemma A.2. *Define $\phi_l(a, b) = \omega^{la+(k-l)b}$ for $l = 0, \ldots, k$, and let*

$$\mathcal{F}_k := \operatorname{span}_{\mathbb{C}}\{\phi_0, \ldots, \phi_k\} \subseteq \mathbb{C}^G.$$

Every network output $\hat{y}(\omega^a, \omega^b; W, v)$ lies in $\mathcal{F}_k$, for any H, W, and v. Under the assumption $k < p$, the functions $\phi_0, \ldots, \phi_k$ are distinct irreducible characters of G, hence mutually orthogonal with respect to the inner product on $\mathbb{C}^G$, hence linearly independent, and $\dim \mathcal{F}_k = k + 1$.

Proof. The first claim follows from the identification $x_1^l x_2^{k-l} = \phi_l$ under the encoding, which gives $\hat{y}(\omega^a, \omega^b; W, v) = \sum_l c_l(W, v)\phi_l(a, b) \in \mathcal{F}_k$.

For the second claim, note that $\phi_l = \chi_{l,\,k-l}$, so two functions ϕ_l and ϕ_m are identical as characters if and only if $l \equiv m \pmod p$ and $k - l \equiv k - m \pmod p$, both of which reduce to $l \equiv m \pmod p$. Since $0 \le l, m \le k < p$, the residues are their own integer representatives, so $l \equiv m \pmod p$ forces $l = m$. The $k + 1$ characters are therefore distinct, and by Schur orthogonality distinct irreducible characters are mutually orthogonal. Orthogonal non-zero vectors are linearly independent, so $\dim \mathcal{F}_k = k + 1$. □

A.3 Every function in $\mathcal{F}_k$ is reachable

The previous lemma establishes an upper bound on the expressible class: network outputs are confined to $\mathcal{F}_k$. The converse is also true, and for essentially trivial reasons once the right construction is identified.

Lemma A.3. *For any hidden width $H \ge k + 1$, the coefficient map $(W, v) \mapsto (c_0, \ldots, c_k) \in \mathbb{C}^{k+1}$ defined by (4) is surjective. Consequently, every element of $\mathcal{F}_k$ is the output of some network.*

Proof. Since the binomial coefficients $\binom{k}{l}$ are nonzero, it suffices to show that the rescaled coefficients $\tilde{c}_l = c_l / \binom{k}{l} = \sum_j v_j W_{1j}^l W_{2j}^{k-l}$ can be chosen freely. Restrict to the first $k + 1$ hidden units, set $W_{2j} = 1$ for $j = 1, \ldots, k + 1$, and write $x_j := W_{1j}$. The system $\sum_{j=1}^{k+1} v_j x_j^l = \tilde{c}_l$ for $l = 0, \ldots, k$ has the matrix form $Vv = \tilde{c}$, where $V_{l,j} = x_j^l$ is a $(k + 1) \times (k + 1)$ Vandermonde matrix. Choosing the x_j to be distinct—for example $x_j = j$—makes $\det V = \prod_{i<j}(x_j - x_i) \ne 0$, so the system has the unique solution $v = V^{-1}\tilde{c}$. This gives an explicit construction for any target coefficient vector, and the conclusion follows. □

Lemmas A.1–A.3 together say that $\mathcal{F}_k$ is *exactly* the expressible class. The network traces out all of $\mathcal{F}_k$ as the weights vary, and nothing outside $\mathcal{F}_k$ is ever produced. This makes the representability question precise: a task f is representable if and only if $\omega^f \in \mathcal{F}_k$, which is a purely Fourier-analytic condition with nothing to do with optimisation or initialisation.

A.4 Theorem 2

Theorem 2. *Under the assumption $k < p$ and with $H \ge k + 1$, the architecture represents the task $f(a, b) = ma + nb \bmod p$ if and only if there exists $l \in \{0, 1, \ldots, k\}$ such that $m \equiv l \pmod p$ and $n \equiv k - l \pmod p$. For nonnegative integers m, n with $m + n \le k$, this reduces to the arithmetic condition $m + n = k$.*

Remark A.4. The reduction is an integer identity on the canonical residues $m, n \in \{0, \ldots, p - 1\}$. The congruence $m + n \equiv k \pmod p$ alone does not suffice: at $p = 97$, $k = 2$ the target $10a + 89b$ satisfies $10 + 89 \equiv 2$, yet $m = 10$ lies outside $\{0, 1, 2\}$, so the Fourier support misses S_k and the target is unrepresentable. Conversely $m = l \le k$ and $n = k - l \le k$ force $m + n \le 2k < k + p$, which excludes the wrapped solution $m + n = k + p$ and makes the integer form equivalent to the congruence pair.

Proof. The target function is the single character $\psi = \chi_{m,n}$.

For the *if* direction: if $(m \bmod p, n \bmod p) = (l, k - l)$ for some $l \le k$, then $\psi = \phi_l \in \mathcal{F}_k$, and Proposition 1(iii) gives weights achieving $c_l = 1$, $c_j = 0$ for $j \ne l$.

For the *only if* direction: suppose $\psi \in \mathcal{F}_k$, so $\psi = \sum_{j=0}^k c_j \phi_j$. Taking the inner product of both sides with ψ and applying Schur orthogonality, $1 = \langle \psi, \psi \rangle = \sum_j c_j \langle \phi_j, \psi \rangle$. If $(m \bmod p, n \bmod p) \notin S_k$, then $\psi \ne \phi_j$ for every j, so each inner product $\langle \phi_j, \psi \rangle = 0$ and the right-hand side is zero, giving $1 = 0$, a contradiction. Hence $(m \bmod p, n \bmod p) \in S_k$, where $S_k = \{(l, k - l) \bmod p : l = 0, \ldots, k\}$ is the support of $\mathcal{F}_k$ in the frequency plane. The reduction to $m + n = k$ for small non-negative integers follows since then the residues coincide with the integers. □

A.5 Theorem A.5

Theorem A.5. *Under the assumption $k < p$, none of the tasks $f(a,b) \in \{ab,\ a^2+b,\ a+b^2,\ a^2+b^2\}$ is representable for any k.*

Proof. By Proposition 1, representability requires $\operatorname{supp}(\widehat{\omega^f}) \subseteq S_k$. We compute the Fourier support of each target directly, using the identity $\sum_{b\in\mathbb{Z}_p} \omega^{cb} = p \cdot [c \equiv 0]$ and the fact that for an odd prime p the quadratic Gauss sum $g = \sum_{x\in\mathbb{Z}_p} \omega^{x^2}$ satisfies $|g| = \sqrt{p}$.

For $f = ab$: $\widehat{\omega^{ab}}(u,v) = \frac{1}{p^2}\sum_{a,b} \omega^{ab-ua-vb} = \frac{\omega^{-uv}}{p}$, which is nonzero for every (u,v). The support is all of $(\mathbb{Z}_p)^2$, which has p^2 elements, while $|S_k| = k+1 \le p$, so containment is impossible.

For $f = a^2 + b$: the Fourier transform factors as a product of a Gauss sum in a and a geometric sum in b; the latter forces $v = 1$, and the former is nonzero for all u. The support is the line $\{(u,1) : u \in \mathbb{Z}_p\}$, which has p elements. This line intersects S_k at the single point $(k-1,1)$, so the support is not contained in S_k for $k < p$.

The cases $f = a + b^2$ and $f = a^2 + b^2$ follow by symmetry and by the same Gauss sum argument, with supports $\{(1,v)\}$ and $(\mathbb{Z}_p)^2$ respectively. □

A.6 Lemma A.6

Lemma A.6. *Let $D_S := \{\, a - b \bmod p : (a,b) \in S \,\}$ be the set of sampled differences. Then*

$$R|_{\mathcal{F}_k} \text{ is injective} \quad\Longleftrightarrow\quad |D_S| \ \ge\ k+1,$$

equivalently $M_S \succ 0$ and $c_S > 0$. In particular: (i) *deterministically, if $N > kp$ then $|D_S| \ge k+1$ and $R|_{\mathcal{F}_k}$ is injective;* (ii) *for $N \le kp$, a uniform draw without replacement satisfies*

$$\Pr\big(|D_S| \le k\big) \ \le\ \binom{p}{k}\Big(\frac{k}{p}\Big)^N.$$

Proof. Every $g = \sum_{l=0}^{k} c_l \phi_l \in \mathcal{F}_k$ factors as $g(a,b) = \omega^{kb}\, Q(\omega^{a-b})$ with $Q(z) = \sum_{l=0}^{k} c_l z^l$, and the map $g \mapsto Q$ is a linear bijection between $\mathcal{F}_k$ and the polynomials of degree $\le k$. Since $|\omega^{kb}| = 1$, g vanishes on S iff Q vanishes on $\{\omega^d : d \in D_S\}$. Note the zero set of a nonzero g is therefore a union of at most k *full diagonals* $\{(a,b) : a - b \equiv d\}$ of the $p \times p$ grid—up to kp cells, not k cells.

($\Leftarrow$) If $|D_S| \ge k+1$, a nonzero polynomial of degree $\le k$ cannot vanish at $k+1$ distinct points of $\mathbb{C}$, so $g|_S = 0$ forces $Q \equiv 0$, i.e. $c = 0$: $R|_{\mathcal{F}_k}$ is injective. ($\Rightarrow$) If $|D_S| \le k$, then $Q(z) = \prod_{d\in D_S}(z - \omega^d)$ is a nonzero polynomial of degree $|D_S| \le k$ whose associated $g \in \mathcal{F}_k$ vanishes on all of S: $R|_{\mathcal{F}_k}$ is not injective. The criterion is therefore exact.

(i) If $|D_S| \le k$, then S is contained in the union of at most k diagonals, which holds at most kp cells; sampling without replacement, $N > kp$ makes this impossible. *(ii)* By a union bound over the $\binom{p}{k}$ candidate difference sets D of size k, $\Pr(|D_S| \le k) \le \binom{p}{k}\Pr\big(S \subseteq \bigcup_{d\in D} \operatorname{diag}(d)\big) = \binom{p}{k}\binom{kp}{N}\big/\binom{p^2}{N} \le \binom{p}{k}(k/p)^N$, using $(kp-i)/(p^2-i) \le k/p$ for all $i \ge 0$. □

A.7 Lemma 3

Lemma 3. *The diagonal of M_S is 1, so $E := M_S - I$ has zero diagonal, and for $l \ne m$ (so $0 < |l-m| \le k < p$) the population mean of $(M_S)_{lm}$ is $\langle \phi_l, \phi_m \rangle_G = 0$. Writing $M_S = \frac{1}{N}\sum_{(a,b)\in S} x_{ab}x_{ab}^\dagger$ with $x_{ab} = (\omega^{l(a-b)})_{l=0}^{k}$ ($\mathbb{E}[x_{ab}x_{ab}^\dagger] = I$, $\|x_{ab}\|^2 = k+1$), the matrix Bernstein inequality (Tropp, 2015) (transferring to sampling without replacement by Gross & Nesme (2010)) gives, for $N \gtrsim k\log(k/\eta)$, with probability $1-\eta$, $\|E\|_{\mathrm{op}} \le \varepsilon_N := O(\sqrt{k\log(k/\eta)/N})$, hence $1 - \varepsilon_N \le c_S \le C_S \le 1 + \varepsilon_N$. In particular $C_S/c_S \to 1$ once $N \gg k\log k$. (An elementary Hoeffding–Serfling argument gives the slightly weaker $\varepsilon_N = O(k\sqrt{\log(k/\eta)/N})$.)*

Proof. Diagonal. $(M_S)_{ll} = \frac{1}{N}\sum_{(a,b)\in S}\omega^0 = 1$.

Off-diagonal mean. For $l \neq m$ put $c = (l-m) \bmod p$, so $c \neq 0$ because $0 < |l-m| \leq k < p$. The population value is

$$\langle \phi_l, \phi_m \rangle_G = \frac{1}{p^2}\sum_{a,b\in\mathbb{Z}_p}\omega^{c(a-b)} = \frac{1}{p^2}\Big(\sum_a \omega^{ca}\Big)\Big(\sum_b \omega^{-cb}\Big) = 0,$$

since each factor is a sum of a nontrivial additive character over $\mathbb{Z}_p$ and vanishes.

Concentration. Fix $l \neq m$. Each summand $\omega^{c(a-b)}$ has unit modulus, so its real and imaginary parts lie in $[-1,1]$, and $(M_S)_{lm}$ is the average of N such terms drawn *without replacement* from the p^2 cells, with population mean 0. By Serfling's inequality for sampling without replacement (Serfling, 1974) (which is at least as concentrated as the with-replacement Hoeffding bound), for $s > 0$,

$$\Pr\big(|\Re(M_S)_{lm}| \geq s\big) \leq 2e^{-Ns^2/2}, \qquad \Pr\big(|\Im(M_S)_{lm}| \geq s\big) \leq 2e^{-Ns^2/2}.$$

Hence $\Pr(|(M_S)_{lm}| \geq \sqrt{2}\,s) \leq 4e^{-Ns^2/2}$. A union bound over the at most $(k+1)^2$ off-diagonal entries gives, with probability at least $1-\eta$,

$$\max_{l\neq m}|(M_S)_{lm}| \leq \sqrt{\tfrac{4}{N}\log\tfrac{4(k+1)^2}{\eta}} =: \rho_N.$$

On this event $E = M_S - I$ has zero diagonal, so

$$\|E\|_{\mathrm{op}} \leq \|E\|_F \leq (k+1)\,\rho_N = O\big(k\sqrt{\log(k/\eta)/N}\big).$$

By Weyl's inequality every eigenvalue of $M_S = I + E$ lies within $\|E\|_{\mathrm{op}}$ of 1, so $1 - \|E\|_{\mathrm{op}} \leq c_S \leq C_S \leq 1 + \|E\|_{\mathrm{op}}$.

Sharpening via matrix Bernstein. The Frobenius factor $(k+1)$ is removable. Writing $x_{ab} = (\omega^{l(a-b)})_{l=0}^{k} \in \mathbb{C}^{k+1}$, one has $(x_{ab}x_{ab}^\dagger)_{lm} = \omega^{(l-m)(a-b)}$, so $M_S = \frac{1}{N}\sum_{(a,b)\in S} x_{ab}x_{ab}^\dagger$ is an average of rank-one Hermitian matrices with $\mathbb{E}[x_{ab}x_{ab}^\dagger] = I$. Every coordinate of x_{ab} is a root of unity, so $\|x_{ab}\|^2 = k+1$ and $\|x_{ab}x_{ab}^\dagger - I\|_{\mathrm{op}} \leq k$ hold deterministically, with no random variable of unbounded or polynomial tail anywhere in M_S and no truncation needed. The matrix Bernstein inequality (Tropp, 2015) for bounded Hermitian summands, which transfers to sampling without replacement by the reduction of Gross & Nesme (2010), gives, provided $N \gtrsim k\log(k/\eta)$, with probability $1-\eta$,

$$\|M_S - I\|_{\mathrm{op}} \;\leq\; \varepsilon_N = O\Big(\sqrt{\tfrac{k\log(k/\eta)}{N}}\Big),$$

a factor $\sqrt{k}$ tighter than the elementary bound. Hence $1-\varepsilon_N \leq c_S \leq C_S \leq 1+\varepsilon_N$ and $C_S/c_S \to 1$ once $N \gg k\log k$. □

Remark. The elementary route above gives $\varepsilon_N = O(k\sqrt{\log(k/\eta)/N})$. The matrix-Bernstein sharpening gives $O(\sqrt{k\log(k/\eta)/N})$ under $N \gtrsim k\log(k/\eta)$. Either suffices for the conclusions of this paper. We quote the sharper constant throughout.

A.8 Corollary A.7

Corollary A.7. *Under Theorem 4, at every step $L_G(\theta^{(t)}) \leq L_S(\theta^{(t)})/c_S$. With Lemma 3, with probability $1-\eta$ and for $\varepsilon_N \leq \frac{1}{2}$,*

$$\sup_t \frac{L_G(\theta^{(t)})}{L_S(\theta^{(t)})} \;\leq\; \frac{1}{c_S} \;\leq\; 1 + 2\varepsilon_N \;=\; 1 + O\Big(\sqrt{\tfrac{k\log(k/\eta)}{N}}\Big),$$

so the two losses coincide up to a factor tending to 1.

Proof. By Theorem 4, for every iterate $\theta^{(t)}$, $c_S\,L_G(\theta^{(t)}) \leq L_S(\theta^{(t)}) \leq C_S\,L_G(\theta^{(t)})$, so whenever $c_S > 0$ the ratio is pinned to $L_G/L_S \in [C_S^{-1}, c_S^{-1}]$ and in particular $\sup_t L_G(\theta^{(t)})/L_S(\theta^{(t)}) \leq 1/c_S$.

This alone excludes only an *unbounded* discrepancy: if c_S is merely positive but small, the bound $1/c_S$ is large, and a training loss of order c_S is compatible with a population loss of order one. The quantitative statement uses Lemma 3: with probability $1-\eta$, $c_S \geq 1-\varepsilon_N$, and for $\varepsilon_N \leq \frac{1}{2}$, $1/c_S \leq (1-\varepsilon_N)^{-1} \leq 1+2\varepsilon_N = 1+O(\sqrt{k\log(k/\eta)/N})$. Thus at any step where $L_S \leq \epsilon$ one has $L_G \leq \epsilon\,(1+2\varepsilon_N)$: there is no window in which the training loss is small while the generalisation loss stays large, at any point of the trajectory and for any optimiser whose iterates remain in the parameter space (the argument uses only $g_\theta \in \mathcal{F}_k$). Finally, a large persistent ratio would require c_S close to 0; in our sampling regime this is excluded deterministically ($N > kp$, Lemma A.6) and quantitatively ($c_S \geq 1-\varepsilon_N$, Lemma 3). □

A.9 Why non-representability blocks memorisation

The representability theorems above address whether a target can be expressed in principle. A subtler question is what happens during training when the target is not representable. Standard intuition from overparameterised networks suggests that a wide network should be able to memorise any finite training set, regardless of whether it can express the full target function. This intuition fails here, and the reason is worth making precise because it accounts for the most distinctive feature of the experimental results: the complete absence of a memorisation regime. It is also what makes δ^2 a *diagnostic* and not a definition, since the non-trivial content is that a population-level quantity controls a finite-sample one.

The key is that the width-independence of $\mathcal{F}_k$ is not just a statement about the function class in the limit of infinite parameters. It holds for every finite H. Because $\hat{y}(\omega^a, \omega^b; W, v) \in \mathcal{F}_k$ for all (W, v) by Proposition 1, the minimum training loss is at least the squared distance from the target values to the nearest element of $\mathcal{F}_k$ restricted to the training set. When the global target function is not in $\mathcal{F}_k$, this distance is bounded away from zero by a term that reflects the spectral energy of the target outside the pass-band.

To make this precise, define the *representability defect* (the spectral gap of the target relative to the pass-band)

$$\delta^2 \;:=\; \|\,\omega^f - \mathrm{Proj}_{\mathcal{F}_k}\,\omega^f\,\|_G^2 \;=\; \sum_{(u,v)\notin S_k} |\widehat{\omega^f}(u,v)|^2 \;>\; 0, \tag{5}$$

where the inequality holds whenever f is not representable. Theorem 6 turns this population quantity into a floor on the empirical loss, proved in Appendix A.10.

For the nonlinear-phase tasks we study, the defect δ^2 is close to its maximum. The target ω^{ab} has $|\widehat{\omega^{ab}}(u,v)| = 1/p$ for all (u,v), so the fraction of spectral energy in S_k is $(k+1)/p^2$. For $p = 97$ and $k \leq 6$ this is at most $7/9409 < 0.1\%$. The training MSE is therefore bounded away from zero by a margin indistinguishable from the maximum loss of an untrained network, and the floor does not depend on H: a network with $H = 10{,}000$ hidden units is subject to the same bound as one with $H = 10$. Empirically, this manifests as flat loss curves and decoded training accuracy pinned at chance for every non-representable task (on the distinction between the MSE floor, which is the theorem, and chance-level decoded accuracy, which is its empirical counterpart, see Section 5.1).

A.10 Theorem 6

Theorem 6. *Assume $|D_S| \geq k+1$, so that $M_S \succ 0$ (Lemma A.6), which holds deterministically under $N > kp$. Let $(v_S)_l = \langle T, \phi_l\rangle_S$ and v_G be the empirical and population Fourier coefficients of the target, and put $\tau_N := \|v_S - v_G\|_2 = O(\sqrt{k\log(k/\eta)/N})$, with ε_N as in Lemma 3. The minimal empirical loss over the class has the closed form $\Phi(S) := \min_{g\in\mathcal{F}_k} \|g-T\|_S^2 = 1 - v_S^\dagger M_S^{-1} v_S$, with population value $\Phi(G) = 1 - \|v_G\|_2^2 = \delta^2$. Assume $\varepsilon_N \leq \frac{1}{2}$. With probability $1-\eta$, allocating $\eta/2$ to each of the two concentration events and retaining all cross terms,*

$$\left|\Phi(S) - \delta^2\right| \;\leq\; \left(\|v_G\|_2 + \tau_N\right)^2 \tfrac{\varepsilon_N}{1-\varepsilon_N} \;+\; 2\,\|v_G\|_2\,\tau_N \;+\; \tau_N^2,$$

which is $O(\sqrt{k\log(k/\eta)/N})$ in general and, in the proportional-sampling regime $N = \rho p^2$, sharpens to $O(1/N)$ for the flat-spectrum targets $\omega^{ab}, \omega^{a^2+b^2}$ and to $O(N^{-3/4})$ for the line-support targets $\omega^{a^2+b}, \omega^{a+b^2}$. All bounds are independent of the hidden width H.

Proof. For $g = \sum_l c_l \phi_l \in \mathcal{F}_k$ with coefficient vector $c \in \mathbb{C}^{k+1}$,

$$\|g - T\|_S^2 = c^\dagger M_S c - 2\,\Re\big(c^\dagger v_S\big) + \|T\|_S^2, \qquad (v_S)_l := \langle T, \phi_l \rangle_S.$$

Because $|T| \equiv 1$, $\|T\|_S^2 = \frac{1}{N}\sum_S |T|^2 = 1$ exactly, with no sampling fluctuation. As $M_S \succ 0$ (Lemma A.6) the quadratic is minimised at $c^\star = M_S^{-1} v_S$, giving

$$\Phi(S) := \min_{g \in \mathcal{F}_k} \|g - T\|_S^2 = 1 - v_S^\dagger M_S^{-1} v_S.$$

Setting $S \to G$ (so $M_S \to I$, $v_S \to v_G$ with $P = \mathrm{Proj}_{\mathcal{F}_k} T = \sum_l (v_G)_l \phi_l$ and $\|v_G\|_2^2 = \|P\|_G^2$) gives $\Phi(G) = 1 - \|v_G\|_2^2 = \|P^\perp\|_G^2 = \delta^2$.

Write $d := v_S - v_G$. Each component $(v_S)_l - (v_G)_l$ is the deviation of the empirical mean of the unit-modulus quantity $T\overline{\phi_l}$ from its population mean; by the same Serfling bound (Serfling, 1974) as in Appendix A.7 and a union bound over the $k+1$ components, with probability $1-\eta$, $\|d\|_2 \le \tau_N = O\big(\sqrt{k \log(k/\eta)/N}\big)$. Decompose

$$\Phi(S) - \delta^2 = -\underbrace{v_S^\dagger \big(M_S^{-1} - I\big) v_S}_{A} - \underbrace{\big(\|v_S\|_2^2 - \|v_G\|_2^2\big)}_{B}.$$

For A, since $M_S^{-1} - I = M_S^{-1}(I - M_S)$ and $\|M_S^{-1}\|_{\mathrm{op}} \le (1-\varepsilon_N)^{-1}$,

$$|A| \le \|v_S\|_2^2 \,\frac{\varepsilon_N}{1-\varepsilon_N} \le \big(\|v_G\|_2 + \tau_N\big)^2 \,\frac{\varepsilon_N}{1-\varepsilon_N}.$$

For B, $\|v_S\|_2^2 - \|v_G\|_2^2 = 2\Re\langle v_G, d\rangle + \|d\|_2^2$, so $|B| \le 2\|v_G\|_2 \tau_N + \tau_N^2$. Using $\|v_G\|_2 \le \|T\|_G = 1$ gives the *general bound*

$$|\Phi(S) - \delta^2| \le (1+\tau_N)^2 \frac{\varepsilon_N}{1-\varepsilon_N} + 2\tau_N + \tau_N^2 = O\Big(\sqrt{\tfrac{k\log(k/\eta)}{N}}\Big).$$

Sharpening. We retain every term; nothing is dropped on the grounds that the in-band energy is small, since the cross term $2\|v_G\|_2\tau_N$ in $|B|$ is first order in τ_N and does not vanish with $\|v_G\|_2^2$. Collecting the bounds above,

$$|\Phi(S) - \delta^2| \;\le\; \underbrace{\big(\|v_G\|_2 + \tau_N\big)^2 \frac{\varepsilon_N}{1-\varepsilon_N}}_{|A|} \;+\; \underbrace{2\|v_G\|_2\,\tau_N + \tau_N^2}_{|B|},$$

which is (3). Now impose the proportional-sampling regime $N = \rho p^2$ with $\rho \in (0,1]$ and k fixed, so that $1/p = \sqrt{\rho/N}$, and recall $\tau_N, \varepsilon_N = O(\sqrt{k\log(k/\eta)/N})$.

Flat spectrum ($\|v_G\|_2^2 \le (k+1)/p^2 = (k+1)\rho/N$; the targets ω^{ab} and $\omega^{a^2+b^2}$, for which $|\widehat{T}(u,v)| = 1/p$ at every frequency): then $\|v_G\|_2 = O(\sqrt{k\rho/N})$ is of the same order as τ_N, and

$$|A| = O\Big(\tfrac{k(\rho+\log(k/\eta))}{N}\Big) \cdot O\Big(\sqrt{\tfrac{k\log(k/\eta)}{N}}\Big) = O\big(N^{-3/2}\big),$$

$$2\|v_G\|_2\tau_N = O\Big(\tfrac{k\sqrt{\rho\log(k/\eta)}}{N}\Big), \qquad \tau_N^2 = O\Big(\tfrac{k\log(k/\eta)}{N}\Big),$$

so every term is $O(k\log(k/\eta)/N)$ and $|\Phi(S) - \delta^2| = O(1/N)$ at fixed (k, ρ, η).

Line support ($\|v_G\|_2^2 = 1/p = \sqrt{\rho/N}$; the targets ω^{a^2+b} and ω^{a+b^2}, whose spectrum lives on a single line meeting S_k in one point of squared mass $1/p$): here $\|v_G\|_2 = \rho^{1/4} N^{-1/4}$, the cross term dominates,

$$2\|v_G\|_2\tau_N = O\big(\rho^{1/4}\sqrt{k\log(k/\eta)}\; N^{-3/4}\big),$$

$$|A| = O\big(\sqrt{k\rho\log(k/\eta)}\; N^{-1}\big), \qquad \tau_N^2 = O\big(k\log(k/\eta)\, N^{-1}\big),$$

giving the intermediate rate $|\Phi(S) - \delta^2| = O\big(\sqrt{k\log(k/\eta)}\, N^{-3/4}\big)$. All bounds are independent of the hidden width H. □

B BEYOND THE FIXED ENCODING

The results above fix the roots-of-unity encoding. This appendix asks what survives when the encoding is learned, and what happens when the input is replaced by a higher-dimensional frequency embedding.

B.1 THE RANK BARRIER BEYOND THE FIXED ENCODING

Theorems 2 and A.5 characterise representability for the specific roots-of-unity encoding, in which the expressible class is the degree-limited set S_k of $k+1$ frequencies on the diagonal $u+v \equiv k \pmod p$. It is natural to ask how much of this structure survives when the encoding itself is allowed to vary, in particular, whether the failure of a target such as ab is an artefact of the fixed basis or an intrinsic property of the architecture. The next observation isolates the part of the obstruction that is encoding-independent.

Let $e_1, e_2 : \mathbb{Z}_p \to \mathbb{C}$ be *arbitrary* embeddings and write the forward pass as $\hat{y}(a,b) = \sum_{j=1}^{H} v_j \big(W_{1j}\, e_1(a) + W_{2j}\, e_2(b)\big)^k$. Collecting the outputs into a matrix $Y \in \mathbb{C}^{p\times p}$ with $Y_{ab} = \hat{y}(a,b)$, and writing ω^f for the target matrix $(\omega^{f(a,b)})_{a,b}$, we have the following.

Proposition B.1. *For every choice of embeddings e_1, e_2, weights W, and v, the output matrix satisfies $\operatorname{rank} Y \le k+1$. Consequently, a task f is representable under some choice of (e_1, e_2, W, v) only if $\operatorname{rank}(\omega^f) \le k+1$.*

Proof. The binomial expansion of Proposition 1(i) does not use the encoding and gives

$$\hat{y}(a,b) = \sum_{l=0}^{k} c_l\, e_1(a)^l\, e_2(b)^{k-l}, \qquad c_l = \binom{k}{l} \sum_j v_j W_{1j}^l W_{2j}^{k-l}.$$

Each summand is the outer product of the column vector $(e_1(a)^l)_a$ with the row vector $(e_2(b)^{k-l})_b$, hence a rank-one $p \times p$ matrix. A sum of $k+1$ rank-one matrices has rank at most $k+1$, so $\operatorname{rank} Y \le k+1$. If $Y = \omega^f$ for some parameters, then $\operatorname{rank}(\omega^f) \le k+1$. □

This condition is strictly weaker than the fixed-encoding criterion: fixing $e_1(a) = e_2(a) = \omega^a$ forces the support onto a single Fourier character on the line $u+v \equiv k$ (Theorem 2), whereas a learned encoding constrains only the *rank*. The gap between the two conditions is exactly the set of targets that a learned embedding makes reachable but the fixed basis does not.

Corollary B.2. *The matrix $\omega^{ab} = (\omega^{ab})_{a,b}$ equals $\sqrt{p}$ times the unitary DFT matrix of $\mathbb{Z}_p$; it therefore has full rank p, with all singular values equal to $\sqrt{p}$. For $k < p-1$ we have $\operatorname{rank}(\omega^{ab}) = p > k+1$, so by Proposition B.1 the task ab is* not representable under any encoding*, and the normalised residual of the best rank-$(k+1)$ output is $1-(k+1)/p$. In contrast, every linear-phase target ω^{ma+nb} and every separable target in $\{a^2+b,\ a+b^2,\ a^2+b^2\}$ yields a rank-one matrix and is representable under a suitable learned embedding for every $k \ge 2$.*

Proof. For ab, the matrix $(\omega^{ab})_{a,b}$ is the character table of $\mathbb{Z}_p$, i.e. the DFT matrix scaled by $\sqrt{p}$; its singular values are all $\sqrt{p}$ and its rank is p. The residual is the Eckart–Young–Mirsky tail of the singular spectrum beyond the top $k+1$ values (Eckart & Young, 1936), normalised by the total energy: $(p-k-1)\,p/p^2 = 1-(k+1)/p$. For the sufficiency direction, a rank-one target factors as $T_{ab} = g(a)h(b)$ with g, h unit-modulus; taking $l=1$, $c_1 = 1$, $c_{l\neq 1} = 0$, $e_1(a) = g(a)$, and $e_2(b) = h(b)^{1/(k-1)}$ gives $e_1(a)^1 e_2(b)^{k-1} = g(a)h(b) = T_{ab}$, where the $(k-1)$-th root is a clean ω-power because $k-1$ is invertible modulo p for $2 \le k \le 6 < p$; the coefficient vector is realised by the Vandermonde construction of Proposition 1(iii). □

Proposition B.1 gives the representability defect (5) a basis-free form. Under a learned encoding the energy of ω^f outside the pass-band S_k becomes the singular-value tail of ω^f beyond the top $k+1$. The reachable family $\mathcal{R}_k = \{\sum_{l=0}^{k} c_l\, e_1(\cdot)^l e_2(\cdot)^{k-l}\}$ is a Veronese/secant-type variety (Harris, 1992; Landsberg, 2012) strictly inside the matrices of rank $\le k+1$, because a generic rank-r matrix has left and right factors that are not powers of a single embedding. The rank condition is therefore

necessary for every target and sufficient only for the rank-one ones proved above. $\mathcal{R}_k$ is also not a $(k{+}1)$-dimensional linear subspace, so the floor argument of Theorem 6 does not transfer. Fitting a random subset of entries becomes a structured low-rank matrix completion problem (Candès & Recht, 2009). We therefore report the learned-encoding floor as the Eckart–Young rank-$(k{+}1)$ lower bound (Eckart & Young, 1936), which the experiments saturate to numerical precision, and we do not claim it as a proved minimum. Within these limits ab is the only task in our grid on the wrong side of the boundary. Appendix E.6 tests the prediction directly, and Appendix E.7 proves strict containment for generic targets by a dimension count.

B.2 The frequency-embedding hierarchy

A d-dimensional embedding of a *single* argument does not define a class of functions on the two-argument domain $(\mathbb{Z}_p)^2$. The well-defined object is a family of feature maps whose coordinates are characters of $(\mathbb{Z}_p)^2$.

Setup. Fix pairwise-distinct *base frequencies* $f_1, \dots, f_d \in (\mathbb{Z}_p)^2$, $f_i = (u_i, v_i)$, and define the frequency embedding $E_d : (\mathbb{Z}_p)^2 \to \mathbb{C}^d$ by $E_d(a,b)_i = \chi_{f_i}(a,b) = \omega^{u_i a + v_i b}$. The network is the same holomorphic architecture with input $E_d(a,b)$: $\hat{y}(a,b) = \sum_{j=1}^{H} v_j \langle w_j, E_d(a,b)\rangle^k$ with $w_j \in \mathbb{C}^d$. The multinomial theorem expands each unit into monomials $\prod_i E_d(a,b)_i^{\alpha_i} = \chi_{\sum_i \alpha_i f_i}(a,b)$ with $\alpha \in \mathbb{N}^d$, $\sum_i \alpha_i = k$, so every output lies in the character subspace

$$\mathcal{F}_{d,k} := \operatorname{span}\{\chi_g : g \in \operatorname{Freq}(d,k)\}, \qquad \operatorname{Freq}(d,k) := \Big\{\textstyle\sum_i \alpha_i f_i \bmod p : \alpha \in \mathbb{N}^d,\ \sum_i \alpha_i = k\Big\},$$

of dimension $|\operatorname{Freq}(d,k)| \le \binom{d+k-1}{k}$. Abstractly, $\mathcal{F}_{d,k}$ is the image of the symmetric power $\operatorname{Sym}^k(\mathbb{C}^d)$ under the linear map sending the monomial basis to characters. This map is injective, and the dimension bound an equality, exactly when the frequency sums $\sum_i \alpha_i f_i$ are pairwise distinct in $(\mathbb{Z}_p)^2$, a genericity condition on the base frequencies. The $d = 2$ case with $f_1 = (1,0)$, $f_2 = (0,1)$ has $\operatorname{Freq}(2,k) = \{(l, k-l) : 0 \le l \le k\} = S_k$ and recovers the fixed model of Section 3. To speak of a hierarchy we fix one infinite sequence $f_1, f_2, \dots$ and let the d-th class use the first d frequencies, so that the classes are nested by construction. Because $\mathcal{F}_{d,k}$ is again a fixed character subspace, the entire dichotomy of Section 4, the sandwich of Theorem 4 and the floor of Theorem 6, applies verbatim with $\mathcal{F}_k$ replaced by $\mathcal{F}_{d,k}$ and $k+1$ by $\dim \mathcal{F}_{d,k}$. Along a nested family the floor collapses monotonically in the discrete parameter d, which is Proposition B.3.

Proposition B.3. *Fix k and a nested family as above, and let*

$$\delta^2(d,k) := \Big\| T - \operatorname{Proj}_{\mathcal{F}_{d,k}} T \Big\|_G^2 .$$

*Then $\delta^2(d,k)$ is non-increasing in d, and with $d^\star := \min\{d : \delta^2(d,k) = 0\}$ (possibly ∞): for $d < d^\star$ the floor is positive and Theorem 6 applies (*FAIL*), while for $d \ge d^\star$ the target is representable and Theorem 4 applies (gapless). The collapse is monotone in the discrete parameter d; we claim no continuity.*

Proof. Adjoining a base frequency f_{d+1} can only enlarge the frequency set, $\operatorname{Freq}(d,k) \subseteq \operatorname{Freq}(d+1,k)$ (take $\alpha_{d+1} = 0$), hence $\mathcal{F}_{d,k} \subseteq \mathcal{F}_{d+1,k}$. Since the characters are orthonormal under $\langle \cdot, \cdot \rangle_G$,

$$\delta^2(d,k) = \big\| T - \operatorname{Proj}_{\mathcal{F}_{d,k}} T \big\|_G^2 = 1 - \sum_{g \in \operatorname{Freq}(d,k)} |\widehat{T}(g)|^2,$$

and enlarging the index set can only increase the in-band energy, so $\delta^2(d+1,k) \le \delta^2(d,k)$. The sequence is bounded below by 0, so it converges; if $T \in \bigcup_d \mathcal{F}_{d,k}$ the threshold $d^\star$ is finite and the stated dichotomy holds on either side of it, and otherwise $\delta^2(d,k)$ decreases to a positive limit with $d^\star = \infty$. □

We define the threshold $d^\star$ by the projection because no rank or dimension formula gives it. For $d \ge 3$, $k \ge 3$ the correspondence between $\dim \operatorname{Sym}^k(\mathbb{C}^d)$ and generic symmetric (Waring) rank has the defective exceptions of Alexander–Hirschowitz (Alexander & Hirschowitz, 1995; Landsberg, 2012; Comon et al., 2008), and $d^\star$ depends on the choice of embedding frequencies.

C FIT TIME AND THE CONDITIONING OF THE MONOMIAL NTK

For a representable target the two-sided sandwich (Theorem 4) already forces train and test to move together. What remains is the *fit time* itself. Because the coefficient map $c(W, v)$ is holomorphic (Proposition 1(i)), the gradient flow on θ pushes forward to the output coefficients through the output-space neural tangent kernel $\Theta(\theta) = JJ^\dagger$, $J = \partial c/\partial\theta$, a $(k{+}1) \times (k{+}1)$ complex Hermitian matrix whose real representation governs the real dynamics. In the lazy regime $\Theta \approx \Theta_0$ and, with $u = c - c_T$, $\dot{u} = -2\eta\,\Theta_0 M_S\,u$, exactly solvable in the eigenbasis.

Theorem C.1. *Let $T \in \mathcal{F}_k$ and suppose the trajectory stays ε_Θ-close to its linearisation (lazy regime; e.g. large enough init scale (Jacot et al., 2018)). Let $0 < \mu_1 \le \cdots \le \mu_{k+1}$ be the eigenvalues of $\Theta_0 M_S$. Then the* training *loss is an exact modal sum, $L_S(t) = \sum_i \mu_i\, e^{-4\eta\mu_i t}\,|\tilde{u}_i(0)|^2$, so $t_{\mathrm{fit}} = \Theta\big((\eta\mu_1)^{-1}\log(1/\epsilon)\big)$. Once the slowest mode dominates L_S, the steps t_1, t_2 at which L_S, L_G first cross ϵ satisfy*

$$|t_1 - t_2| \;\le\; \frac{\ln(C_S/c_S)}{4\eta\,\mu_1}\,\big(1 + o(1)\big) \;\le\; \frac{1}{4\eta\,\mu_1}\cdot\frac{2\varepsilon_N}{1-\varepsilon_N}\,\big(1 + o(1)\big) \;\xrightarrow[N\to\infty]{}\; 0,$$

independently of the threshold ϵ and of $L_S(0)$. Under the stated convention the explicit fit-time constant is $t_{\mathrm{fit}} = (4\eta\mu_1)^{-1}\ln(A_1/\epsilon)(1 + o(1))$ with $A_1 = \mu_1|\tilde{u}_1(0)|^2$ the slow-mode amplitude. Substituting $L_S(0)$ for A_1 adds $(4\eta\mu_1)^{-1}\ln(L_S(0)/A_1)$, which is not negligible at moderate ϵ when $\kappa \gg 1$.

Proof. Let $T \in \mathcal{F}_k$ with coefficient vector c_T, and write $u = c - c_T$. In the lazy regime the output-space gradient flow is linear, $\dot{u} = -\eta\,\Theta_0\,\nabla_u L_S$, and since $L_S = u^\dagger M_S u$ we have $\nabla_u L_S = 2M_S u$ (real gradient in the $(\Re u, \Im u)$ coordinates), hence

$$\dot{u} = -2\eta\,\Theta_0 M_S\,u.$$

$\Theta_0 M_S$ is the product of two Hermitian positive-definite matrices, so it is similar to the Hermitian positive-definite matrix $H := \Theta_0^{1/2} M_S \Theta_0^{1/2}$ via $\tilde{u} := \Theta_0^{-1/2} u$, which satisfies $\dot{\tilde{u}} = -2\eta H\tilde{u}$, i.e. $\tilde{u}(t) = e^{-2\eta H t}\tilde{u}(0)$. Let $0 < \mu_1 \le \cdots \le \mu_{k+1}$ be the eigenvalues of H (equivalently of $\Theta_0 M_S$), with H-eigenbasis coordinates $\tilde{u}_i$. Then

$$L_S(t) = u^\dagger M_S u = \tilde{u}^\dagger H\tilde{u} = \sum_i \mu_i\, e^{-4\eta\mu_i t}\,|\tilde{u}_i(0)|^2,$$

an *exact* modal sum. The slowest mode μ_1 dominates for large t, so with $A_1 = \mu_1|\tilde{u}_1(0)|^2$ the step at which L_S first falls below ϵ satisfies

$$t_{\mathrm{fit}} = \frac{1}{4\eta\,\mu_1}\,\ln\frac{A_1}{\epsilon}\,(1 + o(1)).$$

The population loss, by contrast, is $L_G(t) = \|u\|_2^2 = \tilde{u}^\dagger\Theta_0\,\tilde{u}$, a *different* quadratic form in the same evolving coordinates: since Θ_0 and H do not commute in general, L_G contains cross terms between the H-eigenmodes and is *not* a modal sum. We therefore do not expand it; we control it through the dynamics-free sandwich. The linearised trajectory produces output coefficients $c(t) \in \mathbb{C}^{k+1}$, i.e. outputs in $\mathcal{F}_k$, so Theorem 4 applies verbatim: $C_S^{-1}L_S(t) \le L_G(t) \le c_S^{-1}L_S(t)$ for every t.

Crossing times. Let $t_S(x) := \inf\{t : L_S(t) \le x\}$ (well defined and non-increasing in x since L_S is strictly decreasing), and let $t_1 = t_S(\epsilon)$, $t_2 = \inf\{t : L_G(t) \le \epsilon\}$. The sandwich gives the two-sided localisation

$$t_S(C_S\,\epsilon) \;\le\; t_2 \;\le\; t_S(c_S\,\epsilon),$$

because $L_S \le c_S\epsilon$ forces $L_G \le \epsilon$, while $L_G \le \epsilon$ forces $L_S \le C_S\epsilon$. Since $\operatorname{tr} M_S = k + 1$, the eigenvalue average is 1, so $c_S \le 1 \le C_S$ and t_1 lies in the *same* bracket. Hence

$$|t_1 - t_2| \;\le\; t_S(c_S\epsilon) - t_S(C_S\epsilon).$$

The bracket length is bounded without any single-mode assumption. Writing the modal sum $L_S(t) = \sum_i \mu_i e^{-4\eta\mu_i t}|\tilde{u}_i(0)|^2$ and differentiating,

$$|\dot{L}_S(t)| = 4\eta\sum_i \mu_i^2 e^{-4\eta\mu_i t}|\tilde{u}_i(0)|^2 \;\ge\; 4\eta\mu_1 L_S(t)$$

for every $t \geq 0$, since $\mu_i \geq \mu_1$ termwise. Hence $|t_S'(x)| \leq (4\eta\mu_1 x)^{-1}$ and, integrating over $x \in [c_S\epsilon,\, C_S\epsilon]$,

$$t_S(c_S\epsilon) - t_S(C_S\epsilon) = \int_{c_S\epsilon}^{C_S\epsilon} |t_S'(x)|\, dx \;\leq\; \frac{\ln(C_S/c_S)}{4\eta\,\mu_1},$$

for every $\epsilon \in (0, L_S(0))$, with no $(1+o(1))$ factor. By Lemma 3, $C_S/c_S \leq (1+\varepsilon_N)/(1-\varepsilon_N)$, so $\ln(C_S/c_S) \leq 2\varepsilon_N/(1-\varepsilon_N)$, giving the bound quoted in the theorem; it depends on neither ϵ nor $L_S(0)$, and tends to 0 as $N \to \infty$ at fixed spectrum.

□

Remark (convention). The loss-decay rate $4\eta\mu_i$ and hence the explicit constant $1/(4\eta\mu_1)$ hold under the real-parameter gradient flow $\dot{c} = -\eta\Theta_0\nabla_c L_S$ with $\nabla_u(u^\dagger M_S u) = 2M_S u$. Explicitly, for $L = u^\dagger M_S u$ and real coordinates $(x, y) = (\Re c, \Im c)$, $\partial L/\partial x_j = 2\Re(M_S u)_j$ and $\partial L/\partial y_j = 2\Im(M_S u)_j$, i.e. $\dot{c} = -2\eta\Theta_0 M_S u$. A Wirtinger flow $\dot{u} = -\eta M_S u$ would replace 4 by 2. Since the NTK linearisation descends from gradient descent on the real network parameters, the real-coordinate convention is the physical one, and the theorem statement quotes the convention-independent $\Theta\big((\eta\mu_1)^{-1}\log(1/\epsilon)\big)$.

A fixed learning rate is inadmissible across k: $\lambda_{\max}(\Theta_0)$ grows steeply with k, so stability forces $\eta \sim 1/\lambda_{\max}$, whence $t_{\text{fit}} \sim \kappa(k)\,\ln(1/\epsilon)$ with $\kappa(k) = \lambda_{\max}(\Theta_0)/\lambda_{\min}(\Theta_0)$ the condition number of the output-space NTK at initialisation, which coincides with $\lambda_{\max}/\mu_1$ up to the $O(\varepsilon_N)$ departure of M_S from the identity. The following identifies the mechanism.

Corollary C.2. *Under the initialisation of Appendix F the diagonal of $\mathbb{E}\,\Theta_0$ is $\propto (k!)^2/(l!(k-l)!)$, so the condition number of the mean kernel is*

$$\kappa\big(\mathbb{E}\,\Theta_0\big) = \frac{\max_l\, l!\,(k-l)!}{\min_l\, l!\,(k-l)!} = C(k, \lfloor k/2 \rfloor) \quad (2, 3, 6, 10, 20 \text{ for } k = 2, \ldots, 6).$$

Proof. With the initialisation of Appendix F the output weights have variance $\sigma_v^2 = E|v_j|^2 = 1/H$ and the input weights $\sigma_W^2 = E|W_{ij}|^2 = 1$, so the three Jacobian blocks contribute to $\Theta_0 = \sum_\theta d_\theta d_\theta^\dagger$ with weights $\sigma_W^{2k} = 1$ (d_v) and $\sigma_v^2\sigma_W^{2(k-1)} = 1/H$ (d_{W_1}, d_{W_2}). Summing the input-weight blocks over the H units cancels the $1/H$, so they contribute at order 1 against the d_v block's order H, and the d_v block dominates by a factor H/k. As shown below the two orders carry the *same* dependence on l, so the condition number is the same at every finite width. The d_v entries are $d_{v_j}[l] = \binom{k}{l} W_{1j}^l W_{2j}^{k-l}$, so

$$\big(\mathbb{E}\,\Theta_0\big)_{lm} = \sum_j \binom{k}{l}\binom{k}{m}\, \mathbb{E}\big[W_{1j}^l \overline{W_{1j}}^{\,m}\big]\, \mathbb{E}\big[W_{2j}^{k-l} \overline{W_{2j}}^{\,k-m}\big].$$

For a circularly symmetric complex Gaussian W with $E|W|^2 = \sigma_W^2$, $\mathbb{E}[W^a \overline{W}^{\,b}] = \delta_{ab}\, a!\, \sigma_W^{2a}$, so the expectation is diagonal:

$$\big(\mathbb{E}\,\Theta_0\big)_{ll}^{(v)} = H\binom{k}{l}^2 l!\,(k-l)!\,\sigma_W^{2k} = H\sigma_W^{2k}\,\frac{(k!)^2}{l!\,(k-l)!}, \qquad (\mathbb{E}\,\Theta_0)_{lm} = 0\ (l \neq m).$$

The input-weight blocks are diagonal in l for the same reason. With $\partial c_l/\partial W_{1j} = v_j\binom{k}{l}\, l\, W_{1j}^{l-1} W_{2j}^{k-l}$ and $\mathbb{E}|v_j|^2 = 1/H$, summing over the H units gives

$$\sum_{j=1}^{H} \mathbb{E}\Big[\tfrac{\partial c_l}{\partial W_{1j}}\,\overline{\tfrac{\partial c_l}{\partial W_{1j}}}\Big] = \binom{k}{l}^2 l^2\,(l-1)!\,(k-l)! = l\,\frac{(k!)^2}{l!\,(k-l)!},$$

and symmetrically $(k-l)\,(k!)^2/(l!\,(k-l)!)$ for W_2. Their sum is $k\,(k!)^2/(l!\,(k-l)!)$, because $l + (k-l) = k$ carries no l-dependence. Adding the three blocks,

$$\big(\mathbb{E}\,\Theta_0\big)_{ll} = (H+k)\,\frac{(k!)^2}{l!\,(k-l)!},$$

so the l-dependence is a single common factor and the condition number below holds exactly at every finite width $H \geq 1$, not only as $H \to \infty$. Thus the condition number of the mean kernel is

$$\kappa\big(\mathbb{E}\,\Theta_0\big) = \frac{\max_l (l!\,(k-l)!)^{-1}}{\min_l (l!\,(k-l)!)^{-1}} = \frac{\max_l l!\,(k-l)!}{\min_l l!\,(k-l)!} = \frac{0!\,k!}{\lfloor k/2 \rfloor!\,\lceil k/2 \rceil!} = \binom{k}{\lfloor k/2 \rfloor},$$

the central binomial coefficient. Under the stability-limited learning rate $\eta \sim 1/\lambda_{\max}$, Theorem C.1 gives $t_{\text{fit}} \propto \kappa(k) \log(1/\epsilon)$.

□

Remark (why mean-field is the right claim). The identity $\kappa(\mathbb{E}\Theta_0) = \binom{k}{\lfloor k/2 \rfloor}$ is a statement about the *mean* kernel. Upgrading it to a high-probability lower bound $\kappa(\Theta_0) \geq \binom{k}{\lfloor k/2 \rfloor}(1 - o(1))$ would require resolving the smallest diagonal entry $D_{\min} = D_{\max}/\binom{k}{\lfloor k/2 \rfloor}$ in operator norm, i.e. relative accuracy $o(2^{-k}\sqrt{k})$. Because the entries $W_1^l W_2^{k-l}$ have polynomial tails with $\sim 4^k$ fourth-moment growth, matrix Bernstein delivers this only for width $H \gtrsim c^k$. At fixed k with $H \to \infty$, $\kappa(\Theta_0) \to \binom{k}{\lfloor k/2 \rfloor}$ is a genuine limit. In the practically relevant regime $H = \text{poly}(k)$ no high-probability lower bound is available and the mean-field value is all that is rigorously claimed, which is what the corollary states.

This is a mean-field value, and the measured finite-width $\kappa(\Theta_0)$ lies above it (Table 2). Under the stability-limited learning rate $\eta \sim 1/\lambda_{\max}$, $t_{\text{fit}} \propto \kappa(k)$ (Theorem C.1) gives quantitative form to the poorly conditioned loss surface that $\sigma(z) = z^k$ produces. Its steep growth into $k = 6$ is consistent with the optimisation difficulty observed there, though the experiments use AdamW while the theorem assumes lazy gradient flow with a stability-limited step, so we do not claim the linearised theory predicts the single failed seed.

Remark C.3. Theorem C.1 assumes lazy training. A weaker sufficient condition follows from the Polyak–Łojasiewicz inequality: since L_S is convex in the output c and the parametrisation is surjective (Proposition 1(iii)), whenever the trajectory keeps $\sigma_{\min}(J) \geq \sigma_\star > 0$ one has $\|\nabla_\theta L_S\|^2 \geq 2\sigma_\star^2 c_S\, L_S$, giving $t_{\text{fit}} \leq (2\eta\sigma_\star^2 c_S)^{-1} \ln(L_S(0)/\epsilon)$ with no lazy assumption. The degeneration of $\sigma_\star$ at high k and under learned embeddings is precisely the rich-regime boundary discussed below.

The rank barrier of Appendix B.1 is one face of a representation-theoretic structure, developed next.

D Acquisition dynamics under a learned embedding

This section is a self-contained theoretical extension. Its role in the diagnostic argument is to supply the null model behind the optimiser-versus-landscape separation of Section 5.4: an exact account of what the gradient-flow landscape alone predicts for the acquisition delay, against which the optimiser's contribution can be measured. Readers interested mainly in the diagnostic results can note its two conclusions, a closed-form acquisition time that *decreases* with k, and the localisation of the observed growth outside the landscape, and return to the derivations later.

The delay in question is the one Section 4 rules out under the fixed encoding: there, the linearised fit time of Theorem C.1 runs from a few steps to ∼130 at $k = 6$ (Table 2), and the network itself fits in a few hundred. Once the embedding is *learned*, a far larger delay appears and grows with the activation degree, from ≈ 3800 steps at $k = 2$ to ≈ 7900 at $k = 6$ (Table 6). We reduce the phase-only learned-embedding flow, in the mean-field limit, to an exactly solvable Kuramoto-type system and obtain its escape time in closed form. The conclusion runs against the data. The exact theory predicts an acquisition time that *decreases* with k, so the growth in Table 6 does not live in the gradient-flow landscape at all. Simulating the idealised model confirms the decreasing delay under both vanilla gradient descent and Adam (Appendix D.7), pinning the growth outside the theorem's object. The decisive full-model test, swapping the optimiser with everything else fixed, was reported in Section 5.4, where it resolves the attribution in favour of the optimiser.

D.1 The phase-only model

We isolate the mechanism by which gradient descent locates embedding phases, under assumptions that make the analysis exact while retaining the feature the experiments vary.

Assumption D.1. *(A1)* The embeddings are pure phases, $e_1(a) = e^{i\theta_a}$, $e_2(b) = e^{i\psi_b}$, $\theta, \psi \in \mathbb{R}^p$. *(A2)* Two timescales: the output coefficients $c = (c_l)_{l=0}^k$ are solved to optimality at each step (fast), and the $2p$ phases evolve by gradient flow (slow). *(A3)* The loss is the population loss on the full grid $G = (\mathbb{Z}_p)^2$; restriction to the 40% training subset is handled by the standard concentration of Section 4. *(A4)* Initialisation: θ_a, ψ_b i.i.d. uniform on $[0, 2\pi)$. *(A5)* The target is factorable and unimodular, $T(a, b) = e^{i\gamma_a}\, e^{i\delta_b}$, with both factors non-constant—covering every linear-phase target $ma + nb$ with $m, n \neq 0$ and the separable quadratics $a^2 + b$, $a + b^2$, $a^2 + b^2$.

Remark D.2. Assumptions (A1)–(A2) drop two things present in the experiments of Appendix E.6: the moduli of the complex embeddings with their weight decay, and the joint training of (W, v, e) in place of a two-timescale split. We take the phase-only flow as the object of analysis and the full model as its perturbation, and Appendix D.7 tests whether that division of labour holds. Experiment (i) there runs vanilla gradient descent with c solved in closed form, matching (A2) exactly.

The degree-k activation makes every output a combination of the $k + 1$ features

$$\phi_l(a, b) = e^{i(l\theta_a + (k-l)\psi_b)}, \qquad l = 0, \dots, k.$$

Definition D.3. Let $r_n(\theta) = p^{-1} \sum_a e^{in\theta_a}$ and $s_n(\psi) = p^{-1} \sum_b e^{in\psi_b}$ be the Daido order parameters, and define the *channel coherences*

$$A_l = \frac{1}{p} \sum_a e^{i(\gamma_a - l\theta_a)}, \qquad B_l = \frac{1}{p} \sum_b e^{i(\delta_b - (k-l)\psi_b)}, \qquad R_l = |A_l|\,, \;\; S_l = |B_l|\,.$$

The *available channel set* is $\mathcal{C}(T) = \{l : 1 \leq l \leq k - 1\}$: the channels $l = 0$ and $l = k$ carry a constant factor that cannot align a non-constant target and are excluded.

Note $|\mathcal{C}| = k - 1$, so the number of competing channels itself grows with k. This is the origin of every k-dependence below.

D.2 Slow–fast reduction to a Kuramoto system

Lemma D.4. *Fix the phases. The inner problem* $\min_c \|\sum_l c_l \phi_l - T\|_G^2$ *is convex with optimum*

$$L^\star(\theta, \psi) = 1 - v^\dagger M^{-1} v, \qquad M_{lm} = \langle \phi_l, \phi_m \rangle_G, \quad v_l = \langle T, \phi_l \rangle_G,$$

and, by the envelope theorem, the gradient of $L^\star$ in the phases equals the partial gradient of the joint loss at $c = c^\star(\theta, \psi)$. Here $M \succ 0$ whenever the features are independent on G.

Proof. Write $g = \sum_l c_l \phi_l$ and $J(c, \theta, \psi) = \|g - T\|_G^2$. Expanding, $J = \|T\|_G^2 - 2\Re(c^\dagger v) + c^\dagger M c$ with $M_{lm} = \langle \phi_l, \phi_m \rangle_G$ and $v_l = \langle T, \phi_l \rangle_G$, and $\|T\|_G^2 = 1$ because $|T| \equiv 1$. The matrix M is a Gram matrix, hence Hermitian positive semi-definite, so J is convex in c. When the ϕ_l are independent on G the Gram matrix is nonsingular, $M \succ 0$, and the unique stationary point is $c^\star = M^{-1} v$ with value $J(c^\star) = 1 - v^\dagger M^{-1} v = L^\star$. Since $c^\star$ is interior and $\partial_c J(c^\star, \theta, \psi) = 0$, the envelope theorem gives $\nabla_{(\theta,\psi)} L^\star = \partial_{(\theta,\psi)} J$ evaluated at $c = c^\star(\theta, \psi)$. □

Lemma D.5.

$$M_{lm} = r_{l-m}(\theta)\, \overline{s_{l-m}(\psi)}, \qquad M_{ll} = 1, \qquad v_l = A_l\, B_l.$$

Proof. $\phi_l \overline{\phi_m} = e^{i(l-m)\theta_a} e^{-i(l-m)\psi_b}$; averaging over $(a, b) \in G$ factorises into $r_{l-m}\overline{s_{l-m}}$, and unimodularity of the phases gives $M_{ll} = 1$. For v_l, factorability (A5) splits $\langle T, \phi_l \rangle_G$ into the product $A_l B_l$. □

Lemmas D.4–D.5 recast the learning-to-embed problem exactly as a generalised Kuramoto–Daido system: M is the Gram matrix of two families of order parameters, and v is the vector of channel coherences. The slow flow is gradient ascent of the in-band energy $v^\dagger M^{-1} v$ in these collective coordinates.

Lemma D.6. *Under (A4), $\mathbb{E}\, R_l^2 = \mathbb{E}\, S_l^2 = 1/p$ exactly, the channels are asymptotically independent, and the off-diagonal $|M_{lm}| = O_P(1/p)$, so $M = I + O_P(k/p)$. Moreover, for a* linear-phase *target factor, $\gamma_a = 2\pi ma/p$ with $m \not\equiv 0$, the two end channels carry no target overlap at all: $v_0 = A_0 B_0 = 0$ and $v_k = A_k B_k = 0$ exactly, since $A_0 = p^{-1} \sum_a e^{i\gamma_a}$ and $B_k = p^{-1} \sum_b e^{i\delta_b}$ are then*

full cycles of pth roots of unity and vanish. Only the $k-1$ available channels $l \in \mathcal{C} = \{1, \dots, k-1\}$ contribute, and the initial loss sits at

$$L^\star(0) = 1 - \frac{k-1}{p^2}\,\xi, \qquad \xi = O_P(1),$$

i.e. the plateau lies a distance $(k-1)/p^2$ below the maximal loss 1.

Proof. $\mathbb{E}\, R_l^2 = p^{-2} \sum_{a,a'} \mathbb{E}\, e^{-il(\theta_a - \theta_{a'})} = p^{-1}$ by (A4). With $M \approx I$, $L^\star(0) = 1 - \|v\|_2^2 + o = 1 - \sum_l R_l^2 S_l^2$; the $l = 0, k$ terms vanish identically ($v_0 = v_k = 0$ above), and $\mathbb{E}[R_l^2 S_l^2] = 1/p^2$ by independence for $1 \le l \le k-1$, so the sum is $(k-1)/p^2$. Concentration is the standard sub-Gaussian argument of Section 4. □

Remark D.7. The exact vanishing uses that the factor phase is linear. For the separable quadratics of (A5) it fails: with $\gamma_a = 2\pi a^2/p$ the sum $\sum_a e^{i\gamma_a}$ is a quadratic Gauss sum of modulus $\sqrt{p}$ (Appendix A.5), so $|A_0| = p^{-1/2}$ rather than 0, which at $p = 97$ is 0.102 against the $O(1/p) = 0.010$ of the cross-channel terms. For those targets the end channels carry an $O(p^{-1})$ overlap in $|v|^2$, the same order as a generic channel, and $|\mathcal{C}|$ is $k+1$ rather than $k-1$. The analysis below, and the experiments of Section 5.4 and Table 8, use linear-phase targets throughout, so we state Lemma D.6 and everything that follows from it for those.

At $p = 97$ this plateau is $1 - (k-1)/p^2 \ge 0.99947$, visually indistinguishable from 1, and it is a falsifiable microscopic prediction. Experiment (iv) of Appendix D.7 confirms it to two significant figures.

The reduced single-channel flow admits an exact solution, derived in the next subsection. Its output is Proposition D.13 and the reduced equations (6).

D.3 The single-channel flow is exactly solvable

Fix $l \in \mathcal{C}$ and analyse the flow of channel l in the non-coherent regime, where the other channels decouple (quantified in Appendix D.5). Write $w_a = \gamma_a - l\theta_a - \arg A_l$.

Lemma D.8. *Gradient flow $\dot\theta_a = -\partial L^\star/\partial\theta_a$ gives*

$$\dot w_a = -K \sin w_a, \quad K = \frac{2l^2 S_l^2 R_l}{p}, \qquad \dot R_l = \frac{2l^2 S_l^2 R_l}{p^2} \sum_a \sin^2 w_a \;\ge\; 0,$$

on the invariant manifold of Lemma D.12, and symmetrically for S_l. Both coherences are non-decreasing, so the escape has no turning back.

Proof. Write $z_a = e^{i(\gamma_a - l\theta_a)}$ and $\alpha = \arg A_l$, so $w_a = \arg z_a - \alpha$ and $\sum_a \sin w_a = 0$ by the definition of α. As in Lemma D.9, $\partial R_l^2/\partial\theta_a = (2l/p) R_l \sin w_a$, so the gradient flow of $L^\star = 1 - R_l^2 S_l^2 + O(k/p)$ reads $\dot\theta_a = (2l S_l^2 R_l/p)\sin w_a$ and

$$\dot w_a = -l\dot\theta_a - \dot\alpha = -K \sin w_a - \dot\alpha, \qquad K = \frac{2l^2 S_l^2 R_l}{p}.$$

Differentiating $\sum_a \sin w_a = 0$ along the flow gives $\sum_a \cos w_a\, \dot w_a = 0$, hence $\dot\alpha = -\frac{K}{2} \sum_a \sin 2w_a / \sum_a \cos w_a$. Now $\sum_a \sin 2w_a = p\, \Im m_2$, which vanishes on the invariant manifold of Lemma D.12, where $m_2 = m_1^2$ with $m_1 = R_l$ real. There $\dot\alpha = 0$ and $\dot w_a = -K \sin w_a$. Finally

$$\frac{d}{dt} R_l^2 = \sum_a \frac{\partial R_l^2}{\partial \theta_a}\, \dot\theta_a = \frac{4l^2 S_l^2 R_l^2}{p^2} \sum_a \sin^2 w_a,$$

so $\dot R_l = (2l^2 S_l^2 R_l/p^2) \sum_a \sin^2 w_a \ge 0$. □

Lemma D.9. *The critical set of $-R_l^2$ on the torus is: (i) the incoherent stratum $R = 0$ (non-isolated, the loss-maximal set); (ii) partially anti-aligned configurations (saddles); (iii) the fully aligned orbit $w \equiv 0$, the unique local—hence global—minimum. Branch multiplicity is invisible: shifting $\theta_a \mapsto \theta_a + 2\pi j_a/l$ leaves $e^{il\theta_a}$ fixed, so all l^p aligned branches are zero-loss global minima and the dynamics never sees a spurious barrier from them.*

Proof. Write $z_a = e^{i(\gamma_a - l\theta_a)}$, so $A_l = p^{-1}\sum_a z_a$ and $\partial A_l/\partial\theta_a = -(il/p)z_a$. Hence

$$\frac{\partial R_l^2}{\partial \theta_a} = 2\Re\Big(\overline{A_l}\,\frac{\partial A_l}{\partial \theta_a}\Big) = \frac{2l}{p}\,\Im\big(\overline{A_l} z_a\big),$$

which vanishes for every a exactly when $R_l = 0$ or every z_a is a real multiple of A_l, that is $w_a \in \{0, \pi\}$. This gives the three strata. Differentiating once more,

$$\frac{\partial^2 R_l^2}{\partial \theta_a^2} = \frac{2l^2}{p}\Big(\frac{1}{p} - \Re\big(\overline{A_l} z_a\big)\Big) = \frac{2l^2}{p}\Big(\frac{1}{p} \mp R_l\Big)$$

with the upper sign for $w_a = 0$ and the lower for $w_a = \pi$. For $R_l > 1/p$ an aligned coordinate is a local maximum of R_l^2 and an anti-aligned one a local minimum, so a configuration with at least one anti-aligned oscillator has both ascent and descent directions for $-R_l^2$ and is a saddle, while $w \equiv 0$ is a local minimum of $-R_l^2$. It attains $R_l = 1$, the largest possible value, so it is global. Finally $\theta_a \mapsto \theta_a + 2\pi j_a/l$ shifts $l\theta_a$ by a multiple of 2π and leaves every z_a, hence A_l and the loss, unchanged, and each of the p coordinates admits l such shifts. □

Remark D.10. The incoherent stratum $R = 0$ is non-isolated and loss-maximal, but it is not attracting: by Lemma D.8 the coherence is non-decreasing, $\dot R \geq 0$, with $\dot R > 0$ as soon as $R > 0$ and the drive $\sum_a \sin^2 w_a$ is nonzero, so any initial fluctuation grows monotonically and the flow leaves the stratum instead of stalling on it. This is what makes the escape start from the $O_P(p^{-1/2})$ initial fluctuation. The full Morse–Bott normal form transverse to the stratum is standard and we omit it; only the non-attracting property, which the monotonicity already gives, is used below.

The decisive structural fact is that this flow lies in the *exactly solvable* Watanabe–Strogatz class (Watanabe & Strogatz, 1994), which is stronger than being Kuramoto-like.

Lemma D.11. *In the co-rotating frame the decoupled single-channel flow is $\dot\zeta_a = -K(t)\sin(\zeta_a - \alpha(t))$, i.e. a single-harmonic drive $\dot w = f(t) + G(t)e^{iw} + \overline{G(t)}e^{-iw}$ with $f = -\dot\alpha \in \mathbb{R}$ and $G = iK/2$, identical across all p oscillators (the γ_a are absorbed into initial conditions; there are no heterogeneous frequencies). By the Watanabe–Strogatz theorem (Watanabe & Strogatz, 1994; Marvel et al., 2009) the entire orbit is a single time-dependent Möbius map acting on the initial phases, $e^{iw_a(t)} = \mathcal{M}_t\big(e^{iw_a(0)}\big)$, determined by three real parameters, with the $p - 3$ cross-ratios conserved.*

Proof. The drive $f(t) + G(t)e^{iw} + \overline{G(t)}e^{-iw}$ has coefficients that do not depend on a, so the p oscillators are identical and globally coupled through a single harmonic. This is the hypothesis of the Watanabe–Strogatz theorem (Watanabe & Strogatz, 1994), in the Möbius form of Marvel et al. (2009), which for N identical oscillators under such a drive gives a time-dependent Möbius map of the unit circle carrying the initial phases to the current ones, parametrised by three real degrees of freedom, with the remaining $N - 3$ cross-ratios constant. Taking $N = p$ gives the statement. □

Lemma D.12. *For the wrapped-Cauchy measure with parameter b, pushed forward by $\mathcal{M}(z) = (z + b)/(1 + \bar b z)$, the moments are*

$$m_n = \oint \mathcal{M}(z)^n\,\frac{dz}{2\pi i z} = \mathcal{M}(0)^n = b^{\,n},$$

so the Ott–Antonsen relation (Ott & Antonsen, 2008) $m_2 = m_1^2$ holds exactly on the invariant manifold, with $R = |m_1| = |b|$.

Proof. The integrand has a unique pole inside the contour, at $z = 0$, and the residue there is $\mathcal{M}(0)^n = b^{\,n}$. Setting $n = 1, 2$ gives $m_2 = m_1^2$. □

Proposition D.13. *Under (A4), in the continuum mean-field limit $p \to \infty$ the flow stays on the Ott–Antonsen manifold, and*

$$\sum_a \sin^2 w_a = \frac{p}{2}\big(1 - \operatorname{Re} m_2\big) = \frac{p}{2}\,(1 - R^2).$$

At finite p the first identity is exact, while the closure $m_2 = m_1^2$ holds up to the empirical fluctuation of a discrete population: under (A4) the initial moments are $m_1, m_2 = O_P(p^{-1/2})$, so $m_2 - m_1^2 = O_P(p^{-1/2})$ and the second equality carries a relative correction of that order.

Proof. $\sin^2 w = \frac{1}{2}(1-\cos 2w)$, so $\sum_a \sin^2 w_a = \frac{p}{2}(1-\Re m_2)$, and Lemma D.12 gives $m_2 = m_1^2$ with $|m_1| = R$. □

Throughout the escape phase $R \le \frac{1}{2}$ this quantity lies in $[0.375\,p,\ 0.5\,p]$. What would otherwise be an uncontrolled quasi-uniformity *assumption* on $\sum \sin^2 w_a$ is thus an identity, exact in the mean-field limit and accurate to $O_P(p^{-1/2})$ at finite p, the same order as the deviation measure ν recorded in Appendix D.7. The saturation factor $(1-R^2)$ corrects the late-stage dynamics below.

Remark D.14. A direct attempt to control $\sum \sin^2 w_a$ by a Grönwall pair on $(R, |m_2|)$ fails: the moment equations do not close ($\dot m_2$ involves m_3, etc.), and the crude closure $|m_3| \le 1$ only bounds $|m_2|$ by $O(1)$, too weak to exclude $\mathrm{Re}\, m_2 \to 1$. The Watanabe–Strogatz structure closes the hierarchy exactly instead, upgrading the quasi-uniformity from a hypothesis to Proposition D.13.

D.4 The escape-time closed form

On the Ott–Antonsen manifold the two coupled channels obey, in the variables $X = R^2$, $Y = S^2$,

$$\dot X = \frac{2l^2}{p}\,XY(1-X), \qquad \dot Y = \frac{2(k-l)^2}{p}\,XY(1-Y), \tag{6}$$

with the exact conserved quantity

$$Q = (k-l)^2 \ln(1-X) - l^2 \ln(1-Y) \tag{7}$$

(the coupling factor XY cancels in the orbit equation, in the small-X, Y regime (7) reduces to the linear invariant $(k-l)^2X - l^2Y = c_0$). The saturation factors $(1-X), (1-Y)$ make the escape approach $X = 1$ instead of diverging. The finite-time blow-up of the unsaturated Riccati is the (correct) signal that the trajectory leaves the non-coherent regime.

Theorem D.15. *Under Assumption D.1 and Proposition D.13, with $X_0 = \xi_A/p$, $Y_0 = \xi_B/p$ ($\xi_A, \xi_B = O_P(1)$ by Lemma D.6), the time for channel l to escape the non-coherent regime is, to leading order in p,*

$$\tau_l = (1+o_P(1))\; p^2\, \frac{\ln\!\big((k-l)^2\xi_A/(l^2\xi_B)\big)}{2\big[(k-l)^2\xi_A - l^2\xi_B\big]}. \tag{8}$$

For the typical value $\xi_A = \xi_B = 1$ this is $\tau_l = p^2\,\ln\big|(k-l)/l\big|/\big(k\,|k-2l|\big)$ for every $l \ne k/2$, and $\tau_{k/2} = 2p^2/k^2$ at the symmetric channel. The saturation stage adds a subleading $O\big((p/l^2)\log(1/\epsilon)\big)$.

Proof. Work on the Ott–Antonsen manifold, where (6) is exact. Divide the two equations:

$$\frac{dX}{dY} = \frac{l^2(1-X)}{(k-l)^2(1-Y)} \implies (k-l)^2\frac{dX}{1-X} = l^2\frac{dY}{1-Y},$$

integrating to the conserved Q of (7); expanding $\ln(1-x) = -x + O(x^2)$ recovers the linear invariant $(k-l)^2X - l^2Y = c_0$ with $c_0 = (k-l)^2X_0 - l^2Y_0$ in the small-X, Y region. There $\dot X = \frac{2}{p}X\big((k-l)^2X - c_0\big)$, and

$$\int \frac{dX}{X\big((k-l)^2X - c_0\big)} = -\frac{1}{c_0}\ln\frac{X}{(k-l)^2X - c_0}$$

(check by differentiation). Evaluating from X_0 to the blow-up ($X \to \infty$, where the ratio tends to $(k-l)^{-2}$), and using $(k-l)^2X_0 - c_0 = l^2Y_0$,

$$\tau_l = \frac{p}{2}\cdot\frac{1}{c_0}\,\ln\frac{(k-l)^2X_0}{l^2Y_0} = \frac{p^2\,\ln\!\big((k-l)^2\xi_A/(l^2\xi_B)\big)}{2\big[(k-l)^2\xi_A - l^2\xi_B\big]},$$

which is (8). Setting $\xi_A = \xi_B = 1$ and using $(k-l)^2 - l^2 = k(k-2l)$ gives $p^2\ln((k-l)/l)/(k(k-2l))$; the symmetric case $c_0 = 0$ integrates $\dot X = \frac{2}{p}(k/2)^2X^2$ to $\tau = 2p^2/(k^2\xi)$. The saturation stage (X from a fixed threshold to $1-\epsilon$) contributes $O\big((p/l^2)\log(1/\epsilon)\big)$, subleading in p. □

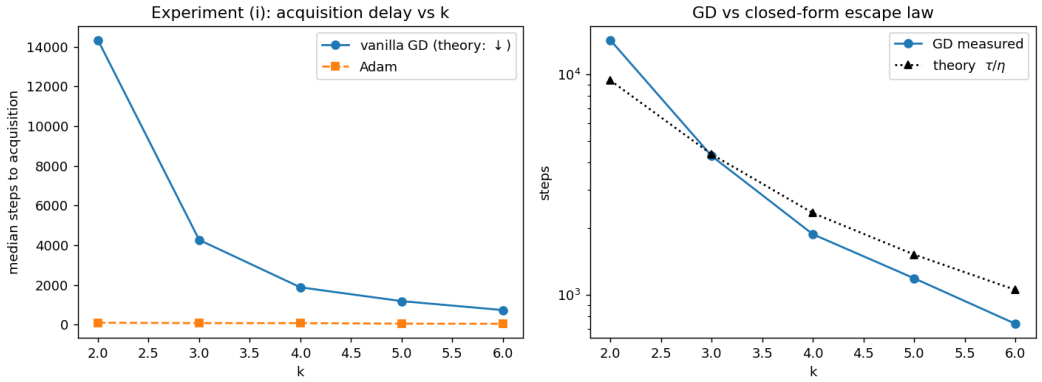

Figure 4: Test (i): idealised two-timescale model, median over 24 seeds ($p = 97$, $\eta = 0.5$). *Left:* under vanilla gradient descent the acquisition delay *decreases* monotonically with k, the opposite of the AdamW growth of Table 6. Adam in the same model also decreases (it only rescales the clock). *Right:* on a log scale the measured GD delay tracks the closed-form escape law τ_l/η of Theorem D.15 across k.

Corollary D.16. *The acquisition time is $\tau_{\mathrm{acq}} = \min_{l\in\mathcal{C}} \tau_l$. In units of p^2 (typical ξ):*

k	2	3	4	5	6
τ_{acq}/p^2	0.500	0.231	0.125	0.081	0.056
winner $l^\star$	1	$\{1,2\}$	2	$\{2,3\}$	3

Moreover τ_{acq} decreases with k as $\sim 2p^2/k^2$, the gap between the winner and the nearest distinct channel is 9.9% at $k = 4$ and 4.0% at $k = 6$ with exact mirror ties at $k = 3, 5$, and each single-seed escape time is heavy-tailed with tail index 2, so it has a finite mean and an infinite variance.

Proof. Acquisition occurs when the first channel escapes, so $\tau_{\mathrm{acq}} = \min_{l\in\mathcal{C}} \tau_l$. Put $u = (k-2l)/k$, so that $(k-l)/l = (1+u)/(1-u)$ and $k - 2l = ku$. Setting $\xi_A = \xi_B = 1$ in (8) gives

$$\tau_l = \frac{p^2}{k^2}\, h(u), \qquad h(u) = \frac{1}{u}\ln\frac{1+u}{1-u} = 2\Big(1 + \frac{u^2}{3} + \frac{u^4}{5} + \cdots\Big),$$

with $h(0) = 2$ by continuity. The series shows h is even and strictly increasing in $|u|$, so τ_l is minimised at the admissible l closest to $k/2$, which is $l = k/2$ for even k and the mirror pair $l = (k \pm 1)/2$ for odd k. This gives the winners and the tabulated values, and $\tau_{\mathrm{acq}} \to 2p^2/k^2$ as the smallest $|u|$ tends to 0. The gap to the nearest distinct channel is $h(u_2)/h(u_1) - 1$ for the two smallest values of $|u|$. For the tail, write $u = (k-l)^2\xi_A$ and $v = l^2\xi_B$, with ξ_A, ξ_B independent and asymptotically $\mathrm{Exp}(1)$ under (A4), so that $\tau_l = \frac{p^2}{2}\,\ln(u/v)/(u-v)$ by (8). The coincidence $u = v$ is *not* a singularity: the integral representation

$$\frac{\ln(u/v)}{u-v} = \int_0^1 \frac{ds}{(1-s)u + sv}$$

is continuous there with value $1/v$, which reproduces the symmetric-channel value $\tau_{k/2} = 2p^2/(k^2\xi)$ computed above. The same representation gives the two-sided bound $\max(u,v)^{-1} \le \ln(u/v)/(u-v) \le \min(u,v)^{-1}$, so a large escape time requires *both* initial coherences to be small, not their difference to vanish. Since $\Pr(\max(u,v) < \epsilon) \asymp \epsilon^2$ for independent exponentials while one-sided smallness costs only a logarithm, $\Pr(\tau_l > t) \asymp C/t^2$, a tail index of 2. The mean is therefore finite and the variance is not. □

Three readings follow. More degree gives faster acquisition under gradient flow, because the fastest available channel sits nearer $l = k/2$ where the product $l(k-l)$ peaks. The near-degenerate channel competition means the winner shifts with the seed. The heavy tail comes from both channel coherences $\xi \sim \mathrm{Exp}(1)$ being small at once, which (8) turns into a large escape time. Taking the minimum over channels leaves the index at 2 and shrinks the scale, so the tail shortens as the channel count $|\mathcal{C}| = k - 1$ grows.

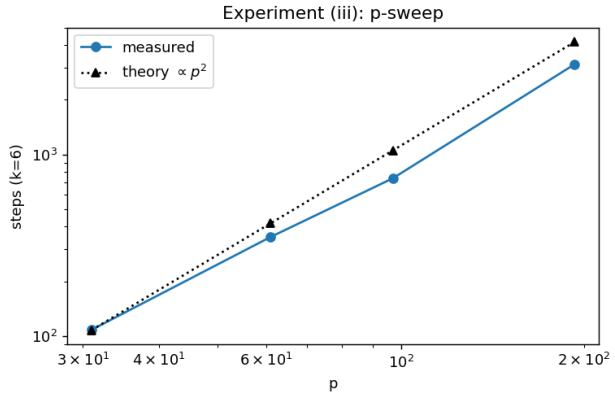

Figure 5: Test (iii): p-sweep at $k = 6$. The median acquisition delay scales as p^2 (Theorem D.15). The mild decrease of τ/p^2 over $p = 31 \to 97$ matches the $O(k/\sqrt{p})$ interference correction shrinking with p.

Corollary D.17. *For fixed-step gradient descent the step count is τ/η. Near the aligned orbit the linear rate of channel l tends to $2l^2/p$, so the largest rate over $\mathcal{C}$ is $2(k-1)^2/p$ and a single step size $\eta < p/(k-1)^2$ is stable for every channel at once. With any such η held fixed across k, as in the experiments, the step count inherits the p^2/k^2 scaling of Corollary D.16 and decreases with k. Stability caps η rather than bounding it below, so over stable step sizes the count obeys $\tau/\eta \gtrsim 2p(1-1/k)^2$, with the minimum approached only at the limit $\eta \to p/(k-1)^2$. That minimum is $O(p)$, but it requires a step size that grows with p; at a step size held fixed as p varies, as in Figure 5, the count keeps the p^2 scaling of τ itself.*

Proof. Near the aligned orbit w_a is small and $\dot{w}_a = -K \sin w_a = -K w_a(1 + O(w_a^2))$ with $K = 2l^2 S_l^2 R_l/p$, so the linear rate of channel l tends to $2l^2/p$ as $R_l, S_l \to 1$. The available channels are $l \in \mathcal{C} = \{1, \dots, k-1\}$, so the largest such rate is $2(k-1)^2/p$, attained at $l = k-1$. Fixed-step gradient descent on a flow with linear rate λ is stable when $\eta\lambda < 2$, giving $\eta < 2/\lambda_{\max} = p/(k-1)^2$. For η fixed across k the count τ_{acq}/η is proportional to τ_{acq}, which Corollary D.16 shows decreases as p^2/k^2. At the stability limit, using $\tau_{\mathrm{acq}} = p^2 h(u^\star)/k^2$ with $h(u^\star) \to 2$,

$$\frac{\tau_{\mathrm{acq}}}{\eta} \gtrsim \frac{2p^2}{k^2} \cdot \frac{(k-1)^2}{p} = 2p\Big(1 - \frac{1}{k}\Big)^2,$$

the inequality running this way because $\eta < p/(k-1)^2$ gives $1/\eta > (k-1)^2/p$. A smaller, more conservative step size only raises the count. □

D.5 Decoupling, and where it stops holding

The single-channel analysis assumed the channels evolve independently in the non-coherent regime. The exact bookkeeping is a lattice of generalised moments $G_{m,n} = p^{-1} \sum_a e^{i(m\gamma_a + n\theta_a)}$ (with an analogous ψ-side lattice H), of which the signal coherences $A_l = G_{1,-l}$ are particular nodes.

Lemma D.18. *Under the multi-channel gradient flow,*

$$\dot{G}_{m,n} = \frac{n}{2} \sum_l \kappa_l \Big(e^{-i\alpha_l} G_{m+1,\,n-l} - e^{i\alpha_l} G_{m-1,\,n+l} \Big), \qquad \kappa_l = \frac{2l}{p} R_l S_l^2,$$

so each node (m, n) couples only to $(m \pm 1,\ n \mp l)$, with scalar, particle-independent coefficients κ_l, α_l.

Proof. Differentiating $G_{m,n} = p^{-1} \sum_a e^{i(m\gamma_a + n\theta_a)}$ along the flow, $\dot{G}_{m,n} = p^{-1} \sum_a i n \dot{\theta}_a\, e^{i(m\gamma_a + n\theta_a)}$. The multi-channel gradient flow gives $\dot{\theta}_a = \sum_l \kappa_l \sin(\gamma_a - l\theta_a - \alpha_l)$

with $\kappa_l = 2lR_lS_l^2/p$ and $\alpha_l = \arg A_l$, so $in\dot{\theta}_a = \frac{n}{2}\sum_l \kappa_l\big(e^{i(\gamma_a - l\theta_a - \alpha_l)} - e^{-i(\gamma_a - l\theta_a - \alpha_l)}\big)$. Substituting and collecting the a-sums,

$$p^{-1}\sum_a e^{i((m+1)\gamma_a + (n-l)\theta_a)} = G_{m+1,\,n-l}, \qquad p^{-1}\sum_a e^{i((m-1)\gamma_a + (n+l)\theta_a)} = G_{m-1,\,n+l},$$

which is the stated recursion. The coefficients κ_l, α_l depend on the channel but not on a. □

Lemma D.19. *The neighbours of the deterministic node $G_{0,0} \equiv 1$ are exactly $(\pm 1, \mp l)$, i.e. the signal coherences $A_l, \overline{A_l}$. Only the signal channels receive a source term.*

Proof. By Lemma D.18 the node (m, n) couples only to $(m \pm 1, n \mp l)$, so the neighbours of $(0, 0)$ are $(\pm 1, \mp l)$. Every other auxiliary moment therefore grows only through products $\kappa \times$ (other moment), with no deterministic source. □

This combinatorial fact is the load-bearing wall of the decoupling estimate, which we can now state. It rests on one quantitative hypothesis, which we make explicit.

Hypothesis D.20 (Equidistribution)**.** Along the flow, up to the acquisition time, the phase differences between distinct channels equidistribute in the sense that

$$\sup_{t \le \tau_{\mathrm{acq}}} \max_{l \ne l' \in \mathcal{C}} \left| \frac{1}{p} \sum_a e^{i(l-l')\theta_a(t)} \right| = O\big(p^{-1/2}\big),$$

and likewise for ψ. Under (A4) this holds at $t = 0$ by the central limit theorem for sums of independent unit phases; the content of the hypothesis is that the flow does not create a persistent correlation between channels.

Theorem D.21. *Under Assumption D.1 and Hypothesis D.20, with probability $1 - \eta$ (excluding an ε-near-coincidence event of the channel clocks, whose measure is $O(\varepsilon)$), the winner's escape time obeys*

$$\tau_{\mathrm{acq}} = \tau_{\mathrm{OA}}\left(1 + O_\eta(k^2 p^{-1/2})\right),$$

improving to $O_\eta(k\, p^{-1/2})$ for a generic winner $l^\star \approx k/2$, where τ_{OA} is the closed form of Corollary D.16.

Proof sketch. Lemma D.19 confines deterministic driving to the signal nodes; Lemma D.11 controls each signal node exactly. The cross-channel term in $\dot{R}_l$ is $p^{-1}\sum_a \sin w_a^{(l')} \sin w_a^{(l)}$; the two phase families decorrelate by Hypothesis D.20, so the sum is $O(\sqrt{p})$ and the cross/self ratio is $O(k\, p^{-1/2})$. The hyperbolic orbit $1/X_0 - 1/X = 2l^2 t/p$ concentrates the time measure in the small-coherence region ($\langle\sqrt{X}\rangle_{\text{time}} = 2\sqrt{X_0} = O(p^{-1/2})$), and the near-degenerate but non-tied clocks yield winner-take-all under gradient flow. □

The mechanism is that interference requires coherence, coherence exists only in the last instant of the hyperbolic escape, and the time-measure concentrates before that instant.

Remark D.22. Theorem D.21 is a genuine asymptotic statement, and at the experimental parameters it says nothing. At $k = 6$, $p = 97$ one has $k^2/\sqrt{p} \approx 3.7$ and $k/\sqrt{p} \approx 0.61$: the bound is not small. To reach $k^2 p^{-1/2} < 0.1$ requires $p \gtrsim 1.3 \times 10^5$; even the generic-winner bound $k\, p^{-1/2}$ requires $p \gtrsim 3600$. The experiments of Appendix E.6 live *outside* the regime where the exact theory controls the interference—and, as we now argue, that is precisely where the observed growth law comes from.

D.6 The growth law lives outside the exact regime

Corollary D.16 predicts $\tau_{\mathrm{acq}} \sim p^2/k^2$, decreasing in k. Corollary D.17 predicts a gradient-descent step count that decreases with k at a step size held fixed across k. The learned-embedding delay of Table 6 does the opposite, rising $3800 \to 7900$ with a concave, saturating profile. The two are not in conflict: the exact theory has drawn its own boundary (Remark D.22), and the data live on the far side of it. This turns "explain the $3800 \to 7900$ growth" into the sharper question of *which* out-of-regime effect produces it. Two candidates are structural and one is routine.

(I) Finite-p channel interference. The cross term of Theorem D.21 is $O(k/\sqrt{p})$, asymptotically negligible but equal to 0.61 at $(k, p) = (6, 97)$, order one. The $k-1$ channels, coupled through the shared phase θ, exert quasi-random pulls on the leader. With the 4% near-degeneracy of Corollary D.16, the winner-selection window lengthens with k. This is landscape-intrinsic but only at finite size.

(II) Adaptive normalisation. AdamW normalises the $O(p^{-3/2})$ non-coherent gradient up to the learning-rate scale, compressing the p^2 law by the ratio between the two: $\tau \sim p^2/2$ natural-time units maps to $O(10^2)$ Adam steps.

We can now test (I) and (II) directly, because the two-timescale model of Assumption D.1 is simulable at the experimental $p = 97$ (Appendix D.7). The result sharpens the picture and rules out the first guess. In the idealised model the acquisition delay *decreases* with k under *both* vanilla gradient descent ($\{14300, 4300, 1900, 1190, 740\}$ steps for $k = 2, \dots, 6$) *and* Adam ($\{102, 83, 82, 56, 50\}$). Adam compresses the clock by a k-dependent factor between $15\times$ and $140\times$ and does not reverse the trend. So neither (I) nor (II), as they act inside the theorem's object, produces growth, the finite-p interference at $(k, p) = (6, 97)$ is real ($k/\sqrt{p} \approx 0.61$) but not strong enough to flip the sign. The growth of Table 6 must therefore originate in what the two-timescale idealisation *omits*: the jointly trained finite-width output layer (W, v) (which conditions the escape instead of solving it in closed form), the free embedding modulus, and weight decay.

Conjecture D.23. *The delay that decreases with k is a property of the gradient-flow landscape of the two-timescale model. The increasing delay under a learned embedding is not, and comes instead from the degrees of freedom the idealisation omits, namely the jointly trained output layer, the free embedding modulus and weight decay, interacting with AdamW. Swapping the optimiser in the full model tests this, and Section 5.4 reports that the growth reverses.*

D.7 Numerical tests of the acquisition dynamics

We simulate the two-timescale model of Assumption D.1 at $p = 97$ with the closed-form envelope loss of Lemma D.4, evaluated through the order parameters of Lemma D.5 (so each step is $O(kp)$, not $O(p^2)$), on the factorable target $\gamma_a = \delta_b = 2\pi a/p$. Five tests isolate five claims. The first four confirm the theory and the fifth is the remaining full-model experiment.

(i) Vanilla gradient descent (the pivotal test). Fixed step size, no adaptivity, no weight decay, c solved in closed form. The median acquisition delay over 24 seeds is 14300, 4300, 1900, 1190, 740 steps for $k = 2, \dots, 6$: **monotonically decreasing**, the opposite of Table 6, and tracking the closed-form escape law τ_l/η of Theorem D.15 across k. The per-seed distribution matches Corollary D.16: $k = 2$ (a single channel, the full $\mathrm{Exp}(1)$ tail exposed) has median $0.76\,p^2$ against the $\xi{=}1$ value $0.50\,p^2$ and the $\mathrm{median}(\xi){=}\ln 2$ value $0.72\,p^2$. For $k \geq 4$ the minimum over channels compresses the tail and pulls the median below the $\xi{=}1$ table. The two-dimensional reduced ODE (6) for the winning channel reproduces the measured escape time across all k (e.g. $k = 6$: ODE 522 vs. GD 370 in matched units), validating the mean-field reduction (Figure 4).

(ii) Per-channel telemetry. Logging $|A_l B_l|\,(t)$ at $k = 6$ shows the predicted braided competition: one channel escapes hyperbolically while the losers are suppressed (winner-take-all, Theorem D.21). The seed shown is won by $l = 2$ and not the marginally fastest $l = 3$, consistent with the 4% near-degeneracy of Corollary D.16, under which the winner shifts with the seed (Figure 6).

(iii) p-sweep. At $k = 6$, $p \in \{31, 61, 97, 193\}$ gives median delays 108, 350, 740, 3143, i.e. τ/p^2 equal to 0.112, 0.094, 0.079, 0.084: the p^2 scaling of Theorem D.15 holds, and the mild decrease $0.112 \to 0.079$ over $p = 31 \to 97$ is consistent with the $O(k/\sqrt{p})$ interference correction shrinking with p (Figure 5).

(iv) Plateau height. Over 2000 seeds the initial loss satisfies $1 - L^\star(0) = (k{-}1)/p^2 \cdot O_P(1)$ to two significant figures (Lemma D.6). E.g. $k = 6$ gives 5.33×10^{-4} against $(k{-}1)/p^2 = 5.31 \times 10^{-4}$, and $|v_0|\,, |v_k|$ are numerically zero, confirming that the end channels carry no target overlap.

(v) Tail. The per-seed escape-time distribution is heavy-tailed and the tail lightens as k (hence $|\mathcal{C}| = k - 1$) grows (Corollary D.16): the $k = 2$ histogram has an outlier beyond 8×10^4 steps, while $k = 6$ is tightly concentrated (Figure 7).

The full-model test of Conjecture D.23. This test was reported in Section 5.4: with every factor held fixed and only the optimiser changed, AdamW reproduces the growth of Table 6 while vanilla

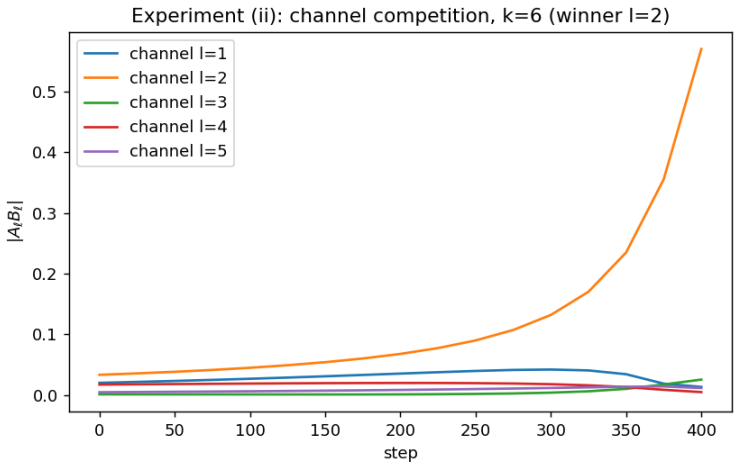

Figure 6: Test (ii): per-channel coherence $|A_l B_l|\,(t)$ at $k = 6$ under vanilla gradient descent. One channel escapes hyperbolically while the losers stay suppressed (winner-take-all, Theorem D.21). The winner is $l = 2$ and not the marginally fastest $l = 3$, within the 4% near-degeneracy of Corollary D.16.

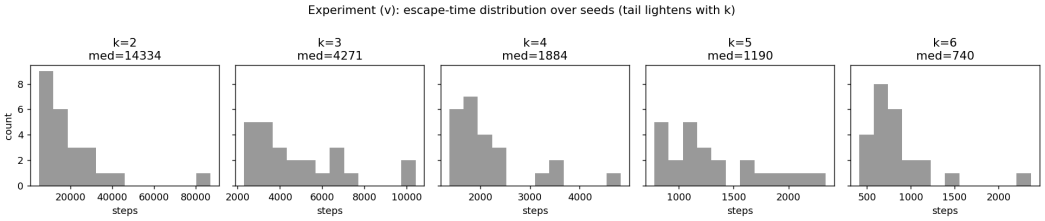

Figure 7: Test (v): escape-time distribution over seeds. Each single-seed time is heavy-tailed (tail index 2, Corollary D.16). The minimum over the $|\mathcal{C}| = k - 1$ channels leaves the index unchanged and shrinks the scale as k grows, so the $k = 2$ histogram has a far outlier while $k = 6$ is tightly concentrated.

Table 1: Empirical Gram conditioning, averaged over 40 random training subsets. M_S stays close to the identity, so representable targets are gapless with the train/test ratio pinned to $[C_S^{-1}, c_S^{-1}]$. Since M_S is Hermitian, $\|M_S - I\|_{\mathrm{op}} = \max(1 - c_S,\, C_S - 1)$, which Lemma 3 bounds by ε_N.

k	2	3	4	5	6
c_S	0.983	0.972	0.969	0.960	0.955
C_S	1.019	1.026	1.033	1.043	1.046
C_S/c_S	1.037	1.055	1.067	1.087	1.095
$\|M_S - I\|_{\mathrm{op}}$	0.019	0.028	0.033	0.043	0.046

gradient descent reverses it, the two curves crossing near $k = 5$ (Figure 3). Conjecture D.23 therefore passes, and tests (i)–(iv) above show the same decreasing trend in the idealised model, where it follows in closed form from Theorem D.15.

Status of the rigorous gaps. We record, for transparency, where the analysis is complete and where it is conditional. (a) The single load-bearing gap is Hypothesis (H) of Hypothesis D.20 of Theorem D.21, the equidistribution of the phase differences $(l - l')\theta_a$ along the flow, from which the $O(k/\sqrt{p})$ interference constant follows. The theorem is stated as conditional on it. The remaining items are subsidiary and do not affect the main chain. (b) A finite-p transport bound pairs the deviation measure ν against $\mathcal{M}_t^n$ in a negative Sobolev norm with constant $\sim (1 - |b|)^{-s}$. This is routine. (c) The early-time signal/noise balance, with b_0 and ν both $O_P(p^{-1/2})$, affects only the $O_P(1)$ constant inside the logarithm of (8), already carried by ξ. (d) The non-attracting character of the $R = 0$ stratum follows from the coherence monotonicity of Lemma D.8 (Remark after Lemma D.9). We omit only the standard transverse normal form. (e) The modulus and weight-decay degrees of freedom sit outside the flow analysed here, under the division of labour set out in Remark D.2 and tested in Appendix D.7. (f) Lemma D.11 applies the Watanabe–Strogatz theorem (Watanabe & Strogatz, 1994; Marvel et al., 2009) and Lemma D.4 the envelope theorem, and neither is claimed as new. Apart from (H), nothing on the main chain is an unstated approximation, and the results are asymptotic with explicit validity boundaries (Remark D.22).

E Additional experiments

E.1 Numerical validation of the dichotomy

The three quantities below are computable from the initialisation statistics or the analytic Fourier structure alone, with no training. We use $p = 97$, $H = 128$, the initialisation of Appendix F, and $N = \lfloor 0.4\, p^2 \rfloor = 3763$.

Gram conditioning (Theorem 4, Lemma 3). Table 1 reports c_S, C_S and the condition number C_S/c_S of the empirical Gram M_S (2), averaged over 40 random training subsets. The deviation $\|M_S - I\|_{\mathrm{op}}$ stays below 0.05 and $C_S/c_S \le 1.10$ throughout, so the sandwich of Theorem 4 is nearly tight and gaplessness is quantitative and not merely observed.

The two rows above the ratio carry information the ratio hides. The smallest eigenvalue c_S falls from 0.983 to 0.955 across $k = 2, \ldots, 6$ while the largest rises from 1.019 to 1.046, so the empirical form drifts away from the identity in both directions at once and at a comparable rate. What Theorem 4 delivers is the one-sided constant $1/c_S$, which stays below 1.05 over the whole range, so a training loss of 10^{-8} certifies a population loss below 1.05×10^{-8} at every degree we run. The drift is also slow in k: the deviation roughly doubles from $k = 2$ to $k = 6$ while the class dimension more than doubles, so the sample preserves the geometry of $\mathcal{F}_k$ no worse as the pass-band widens.

NTK conditioning and fit time (Corollary C.2). Table 2 reports $\mu_1, \lambda_{\max}, \kappa$ of the output-space NTK Θ_0 and the linearised fit time t_{fit} (median over 40 inits, stability-limited step). The condition number grows super-linearly, exceeds its central-binomial mean-field value, and blows up at $k = 6$ ($\kappa \approx 132$, and the t_{fit} inter-quartile range widens to 234), consistent with the $k = 6$ optimisation difficulty observed in Section 5. The two eigenvalue rows separate effects that κ combines. The

Table 2: Output-space NTK conditioning and linearised fit time. $\kappa(k)$ exceeds the mean-field value $C(k, \lfloor k/2 \rfloor) = \kappa(\mathbb{E}\,\Theta_0)$ and blows up at $k = 6$.

k	2	3	4	5	6
μ_1	221	585	1.65×10^3	5.53×10^3	1.45×10^4
$\lambda_{\max}$	549	2.9×10^3	2.2×10^4	2.2×10^5	1.9×10^6
$\kappa = \lambda_{\max}/\mu_1$	2.6	5.1	14.4	40.3	132
$C(k, \lfloor k/2 \rfloor)$	2	3	6	10	20
t_{fit} (steps)	4	8	21	50	136

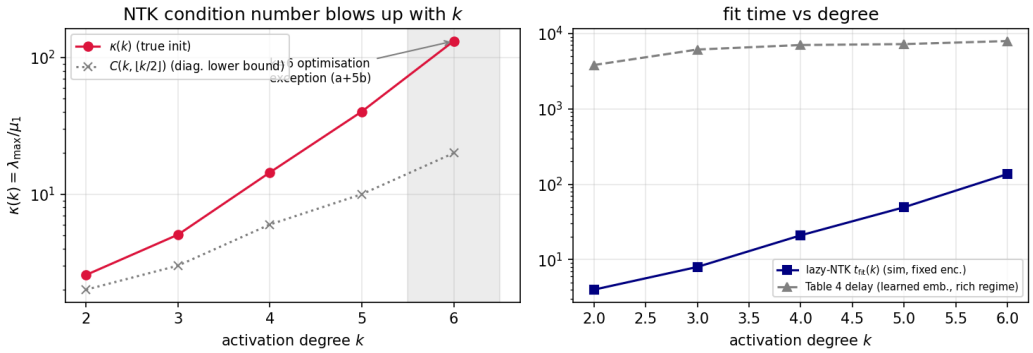

Figure 8: **NTK conditioning and fit time.** *Left:* the output-space NTK condition number $\kappa(k)$ grows super-linearly and exceeds the mean-field value $C(k, \lfloor k/2 \rfloor)$. The shaded band marks the $k = 6$ optimisation exception. *Right:* the linearised fit time $t_{\text{fit}}(k)$ tracks $\kappa(k)$ (navy). Learned-embedding delays of Table 6 (grey) are the rich-regime acquisition phase and are not predicted by the lazy NTK, the two shapes differ by design.

largest eigenvalue grows by three and a half orders of magnitude from $k = 2$ to $k = 6$ while the smallest grows by less than two, so the condition number is driven by the top of the spectrum and not by a collapsing bottom. This is why the stability-limited step size, and not the slowest mode, is the binding constraint under Theorem C.1. $\eta \sim 1/\lambda_{\max}$ shrinks faster than μ_1 grows, and the linearised fit time rises from 4 to 136 steps as the product of the two. Figure 8 plots $\kappa(k)$ and $t_{\text{fit}}(k)$. The learned-embedding delays of the encoding ablation are shown for contrast and are *not* predicted by the lazy NTK (they are the rich-regime acquisition phase).

Spectral floors (Theorem 6). Table 3 and Figure 9 give the analytic floors $\delta^2(f, k)$ for the non-representable targets. They match the closed forms $1-(k+1)/p^2$ (for ab and a^2+b^2) and $1-1/p$ (for $a^2 + b$ and $a + b^2$), so they sit within 10^{-3} of the maximum for the first pair and within $1/p \approx 10^{-2}$ for the second, and, being width-independent, coincide with the observed flat, chance-level training curves. Representable linear targets give $\delta^2 = 0$ exactly at $m + n = k$ and $\delta^2 = 1$ otherwise, reproducing the staircase of Theorem 2.

E.2 Per-degree outcome of the fixed-encoding grid

Every representable cell is solved at every seed for $k \leq 5$. The single shortfall is at $k = 6$, where one seed of $a + 5b$ does not reach the threshold within the classification budget and converges under the longer rerun of Section 5.1. Table 4 gives the rates.

E.3 Parameter count carries no information about the outcome

Proposition 1(ii) makes $\mathcal{F}_k$ independent of the hidden width, so overparameterisation should move neither boundary. Figure 10 tests this over a 256-fold range of parameter count, from $H = 8$ to $H = 2048$ (48 to 12288 real parameters), at $k \in \{4, 6\}$, with the step size selected per cell so that every configuration is given a fair chance to converge, on a 50% split ($N = 4704$). The outcome is

Table 3: Analytic spectral floors $\delta^2(f,k)$. By the $a \leftrightarrow b$ symmetry $ab \equiv a^2{+}b^2$ and $a^2{+}b \equiv a{+}b^2$.

f	$k{=}2$	$k{=}3$	$k{=}4$	$k{=}5$	$k{=}6$
$ab,\ a^2{+}b^2\ (=1-\frac{k+1}{p^2})$	0.99968	0.99957	0.99947	0.99936	0.99926
$a^2{+}b,\ a{+}b^2\ (=1-\frac{1}{p})$	0.98969	0.98969	0.98969	0.98969	0.98969

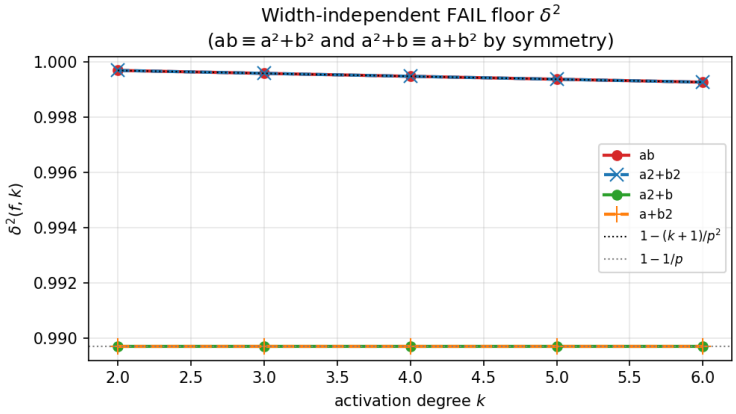

Figure 9: **Width-independent FAIL floor δ^2.** The four non-representable targets collapse onto two symmetry classes: $ab \equiv a^2+b^2$ at $1-(k+1)/p^2$ and $a^2+b \equiv a+b^2$ at $1-1/p$. The floor is analytic and independent of H, matching the flat training curves of the FAIL cells.

flat throughout. Representable targets reach numerical zero at every width, including $H = 8$, which exceeds $k+1$ by one for $k = 6$. Non-representable targets sit on their empirical projection floor at every width, 0.99917 at $k = 4$ and 0.99873 at $k = 6$, in both cases matching δ_S^2 to five decimals, and a 256-fold increase in parameters moves neither value.

Width does change the time to reach the floor, though not the floor itself. At $k = 6$ the representable run at $H = 8$, one unit above the minimum $k+1 = 7$, takes 4450 steps to reach numerical zero, while every wider configuration reaches it in 350 to 450. At $k = 4$ the corresponding figures are 900 against 350. The expressible class is the same at every width, so what widening buys is conditioning of the parametrisation and not reach. The two diagnostics divide the question between them. Representability fixes which targets are attainable, and the conditioning of Appendix C fixes how long attaining them takes.

The contrast with the right panel is the point. Parameter count, the quantity that organises the capacity account of grokking, is uninformative here: the same count occurs in both regimes and predicts neither. The representability defect, computed before training from the target's Fourier support, predicts both exactly. The width sweep therefore separates the two axes the account of Section 6 runs together, since here the parameter count moves over 256-fold with the outcome fixed, while a change of target at fixed count moves the outcome from numerical zero to the floor.

E.4 The expressible class is the mechanism, not the task

The results above establish that representability determines the outcome of training for the holomorphic network. They do not by themselves rule out the possibility that the tasks themselves are simply easy or hard in a way that happens to correlate with the algebraic condition. The ReLU baseline addresses this, with one qualification. The ReLU network differs from the holomorphic one in encoding (one-hot versus roots of unity), loss (cross-entropy versus complex MSE) and width at once, so it is the canonical grokking baseline of the literature (Power et al., 2022; Nanda et al., 2023), evidence that the same tasks can be memorised and grokked in a universal architecture. It is not a matched control isolating the activation as the single varying factor.

Table 4: Among representable cells, the fraction of the three seeds reaching INST in the fast classification pass.

Degree k	2	3	4	5	6
INST rate	9/9	12/12	15/15	18/18	20/21

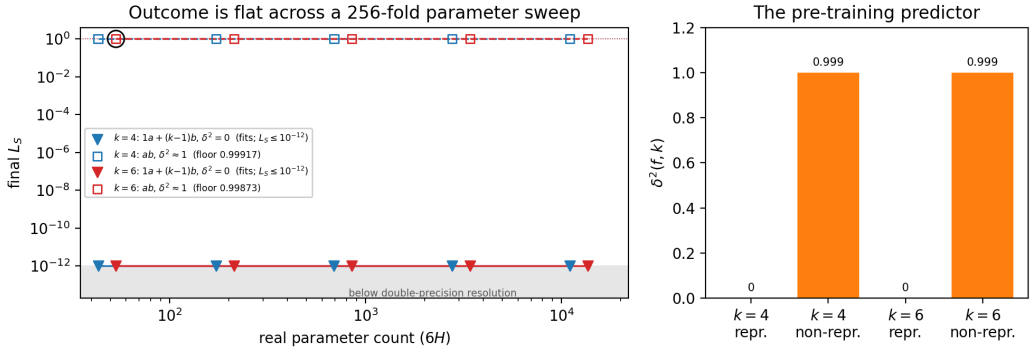

Figure 10: Width sweep, 4 seeds per cell, 50% split, with the step size selected per cell so that every configuration is given a fair chance to converge. *Left:* final training loss against real parameter count. The two degrees are offset horizontally for legibility only. Representable targets (filled triangles) fit at every width, the marker denotes the upper bound $L_S \leq 10^{-12}$ and not a measured value, since the achieved losses (10^{-20} to 10^{-13}) lie below the resolution of double-precision arithmetic at this scale. Non-representable targets (open squares) sit on their empirical projection floors (dotted lines), 0.99917 at $k = 4$ and 0.99873 at $k = 6$. The two regimes are separated by twelve orders of magnitude and neither responds to a 256-fold increase in parameter count. The circled point ($k = 6$, ab, $H = 8$) had not fully settled within its step budget and sits 0.13% *above* its floor. As Theorem 6 bounds L_S from below, this is consistent with the theorem and not a violation of it. *Right:* the representability defect for the same four configurations, computed from the target's Fourier support before any training, exactly 0 for the representable targets, 0.99947 and 0.99926 for the others.

We run the same thirty-nine tasks on a standard two-layer ReLU network with $2p$-dimensional one-hot inputs, hidden width 256, cross-entropy loss, AdamW with weight decay 1.0, and 8000 training steps, the same recipe used in the grokking literature (Power et al., 2022; Nanda et al., 2023). The contrast with the holomorphic results is complete. The ReLU network never fails: it memorises every task without exception, including the nonlinear-phase targets that the holomorphic network cannot represent at all. On the harder targets it exhibits textbook grokking, with test accuracy lagging training accuracy by thousands of steps. On the easier linear-phase targets it generalises more quickly, with the gap shrinking as the task's arithmetic structure becomes easier to exploit.

Figure 11 shows the distribution of outcomes. The holomorphic network occupies only two categories, FAIL and INST, with nothing in between. The ReLU network occupies the full spectrum: INST on the simplest tasks, GROK on the harder ones, and a continuum of PART outcomes in between, but never FAIL. Three of the four targets that produce FAIL for the holomorphic network, $a^2 + b$, $a + b^2$ and $a^2 + b^2$, produce some of the cleanest grokking for ReLU, with gaps between 3700 and 7500 steps. The fourth, ab, is fitted on the training set but its test accuracy climbs only partway inside the budget, so it lands in PART.

Figure 12 makes the same point through the distribution of generalisation gaps. Among the holomorphic runs that succeed, the gap is zero by definition: the network cannot memorise, so there is no delay to measure. Among the ReLU runs that succeed, the gap ranges from zero to 7500 steps, with the grokking cluster sitting at the upper end and corresponding precisely to the tasks that are out of reach for the holomorphic network. The same arithmetic difficulty that makes a task algebraically unrepresentable in the holomorphic setting makes it slow to grok in the ReLU setting, but the two outcomes are qualitatively different. One is a hard floor imposed by the expressible class. The other is a delay imposed by the competition between memorisation and generalisation dynamics.

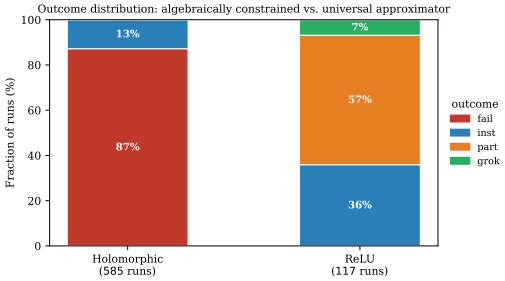

Figure 11: Outcome distribution for the holomorphic and ReLU networks. The holomorphic network occupies only the two extreme categories. The ReLU network spans the full spectrum and never fails.

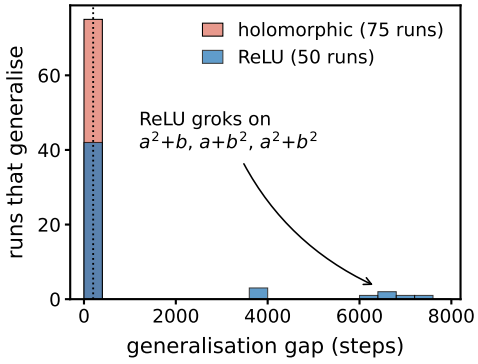

Figure 12: Generalisation gaps among runs that succeed. Holomorphic runs are gapless by construction. ReLU runs span a continuum, with the grokking cluster at 3700–7500 steps corresponding to the tasks that the holomorphic network cannot represent.

Two entries of the tally carry the control. No run of the 117 fails, so the tasks are not intrinsically unfittable at this sample size and the holomorphic failures cannot be attributed to the data. And no run memorises without generalising either. The 67 PART outcomes are runs whose test accuracy is still climbing when the budget ends, above chance and short of the 99% threshold, and not runs that have fitted the training set and stopped. The ReLU baseline therefore occupies one end of the capacity axis throughout, where fitting is never in question and only the rate varies, which is the regime the capacity account of grokking assumes and the holomorphic network leaves.

E.5 ISOLATING OPTIMISATION FROM REPRESENTABILITY

The $k = 6$ optimisation exception raises a question the phase diagram alone cannot settle: is the occasional failure a matter of representability or of the non-convex landscape induced by the factorisation through (W, v)? We separate the two with a convex baseline. Fitting a representable target directly in the coefficient space of $\mathcal{F}_k$, i.e. solving $\min_c \|\sum_l c_l \phi_l - T\|_S^2$, a convex least-squares problem, removes the parametrisation entirely and always attains the global optimum. Comparing it to the network trained through (W, v) on the *same* target isolates the optimisation cost.

Table 5 reports both for the representable targets $f = 1{\cdot}a + (k{-}1)b$. The convex baseline reaches machine-zero training loss at every k: representability holds throughout, as the theory requires. The network reaches the same optimum but takes progressively more steps as k grows, consistent with the NTK condition number $\kappa(k)$ of Corollary C.2. The growth is steepest at $k = 6$, the degree at which the phase-diagram grid of Section 5.1 also shows its one delayed cell. The gap between the two columns is therefore pure optimisation difficulty, the z^k-induced ill-conditioning quantified in Appendix C, and not a failure of representability.

This is the misdiagnosis scenario of the introduction in miniature, and Figure 1 exhibits it directly. Two $k = 6$ runs on a 50% split are trained by vanilla gradient descent from the same initialisation and step size, differing only in the target: a representable one ($1a + 5b$, $\delta^2 = 0$) and a non-representable one (ab, $\delta^2 = 0.9993$). The step size is the largest stable one, $0.9 \times 2/\lambda_{\max}$ of the initial NTK, so the representable run is proceeding as fast as gradient descent permits. It reaches $L_S < 10^{-6}$ after $\tau_A = 549$ steps, against 270 for the same target under AdamW (Table 5) and a linearised prediction of 136 (Table 2). The three differ by optimiser and threshold, and not by target. Truncate training at 26.4% of τ_A and the two runs are indistinguishable at chance decoded accuracy, with losses of order one, $L_S = 1.03$ against 2.02, on an axis the fitted run later crosses down to 1.8×10^{-10}. Continue, and they separate absolutely, the first falls to 1.8×10^{-10}, the second settles at 0.99873, its empirical projection floor δ_S^2 to five decimals, and stays there.

Table 5: Convex $\mathcal{F}_k$ baseline versus the non-convex network on representable targets $f = 1a+(k-1)b$ at $p = 97$. The convex fit attains the global optimum at every k (representability holds). The network's steps-to-fit grow monotonically with k and are largest at $k = 6$, isolating the optimisation cost from representability. Each entry is a single run at a fixed seed rather than an average over seeds, so the column shows the trend in k and not its spread; the seed-to-seed behaviour at $k = 6$ is measured by the phase-diagram grid instead.

k	2	3	4	5	6
convex $\mathcal{F}_k$, train loss	$< 10^{-30}$ at every k				
network (AdamW), steps to $L_S < 10^{-8}$	179	181	222	234	270

The truncation fraction is not an artefact of the step size. Repeating the comparison at 0.5 and 0.05 of the stability limit rescales τ_A as $1/\eta$ ($549 \to 987 \to 9888$ steps, an 18-fold range) while the fraction at which the two runs become distinguishable stays at 26.4% in all three cases. Gradient descent here simply traverses the same trajectory at different rates, and the ambiguity occupies the same quarter of it however fast one runs.

The two situations demand opposite responses (train longer, or change the architecture), and no feature of the truncated curves distinguishes them. The representability defect does, before training: 0 against 0.9993. The conditioning does not, because the two diagnostics are of different types. The NTK condition number at initialisation is $\kappa = 113$ for the initialisation these two runs share, against a median of 132 over 40 initialisations (Table 2), the seed-to-seed spread being wide. It is a property of the architecture and the sample, and is *identical* for the two runs, as it must be, since it does not involve the target at all. Conditioning accounts for how long a representable target takes. It can never account for a target that is not there. Only δ^2 separates a slow run from an impossible one.

E.6 Encoding ablation

The fixed roots-of-unity encoding hands the network a Fourier-friendly representation, and a reasonable worry is that the sharp staircase of Figure 2 is an artefact of that choice and not a property of the architecture. Proposition B.1 predicts what should happen if the encoding is instead *learned*: representability should broaden from the line $m + n = k$ to the rank condition $\mathrm{rank}(\omega^f) \le k + 1$, so that every linear-phase and every separable target becomes reachable at every degree, while ab, the one full-rank target, remains impossible. We test this by replacing the fixed inputs with trainable complex embeddings $e_1, e_2 : \mathbb{Z}_p \to \mathbb{C}$ (two p-entry complex tables), optimised jointly with W, v under the same AdamW recipe. Everything else in the protocol of Section 5 is unchanged. We consider two initialisations of the embeddings, random unit-modulus phases (LEARNED-RAND) and the roots-of-unity values themselves (LEARNED-ROOT), and run the full $39 \times 5 \times 3$ grid for each.

The outcome matches Proposition B.1 exactly. Under both initialisations, 38 of the 39 tasks become representable and are solved (train and test both exceed 99%) at every degree and every seed, all 35 linear-phase targets and the three separable targets $a^2 + b$, $a + b^2$, $a^2 + b^2$, resolving as INST in the large majority of cells with a small GROK minority quantified below. Meanwhile ab fails at every degree and every seed, its loss pinned at the rank-$(k + 1)$ floor of Corollary B.2. The $m + n = k$ staircase, the defining feature of the fixed-encoding phase diagram, dissolves completely: a learned embedding makes the entire linear region and all separable targets representable, leaving a single failing column at ab (Figure 13). The two initialisations behave near-identically, which is itself informative, the boundary is set by representability, not by how close the initial embedding sits to the analytic Fourier basis.

Broadening representability does *not* broadly reintroduce grokking, for two reasons. The successful cells are overwhelmingly INST and not GROK: of the 190 solved cells per initialisation (38 tasks $\times$ 5 degrees), 186 are INST under random-phase initialisation and 177 under roots-of-unity initialisation, with *no* MEM outcome anywhere in either sweep. Training and test accuracy cross together, empirically the same gapless signature as the fixed holomorphic network. This is an *observation* here, not a transferred theorem. Once the embeddings are trained, the reachable family $\mathcal{R}_k$ is no longer a fixed linear subspace, the restriction-map sandwich of Theorem 4 does not apply, and fitting a random subset of entries becomes a structured low-rank matrix-completion problem (Candès &

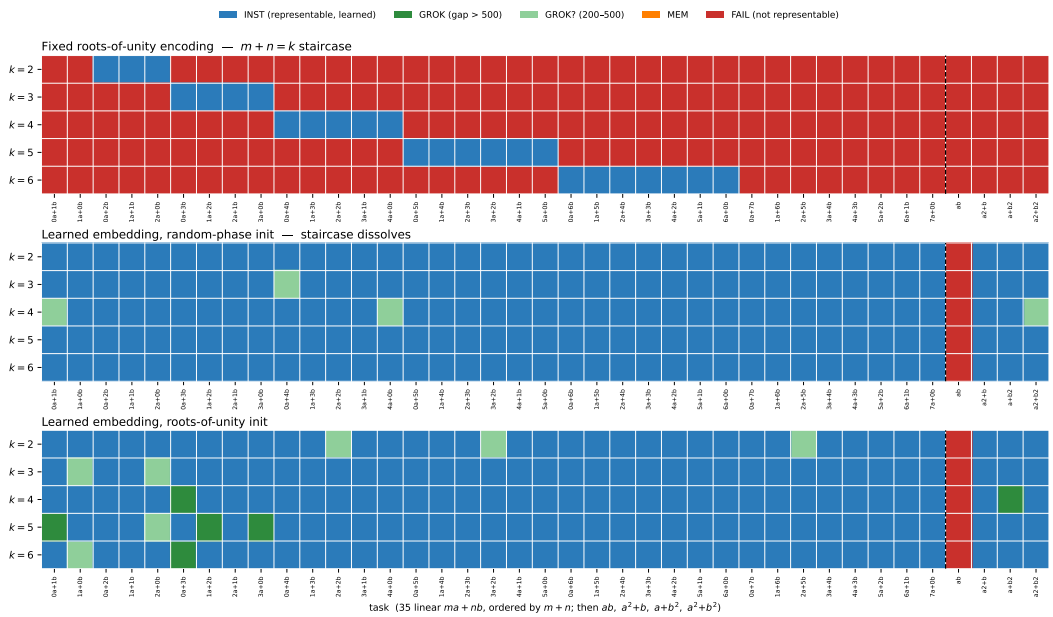

Figure 13: Phase diagrams under three encodings, across degrees k (rows) and the 39 tasks (columns, ordered as in Figure 2). *Top*: the fixed roots-of-unity encoding reproduces the $m + n = k$ staircase (the data of Figure 2). *Middle, bottom*: learned embeddings, initialised from random phases and from the roots of unity respectively, collapse the staircase, every linear and separable target is learned to INST at every degree, leaving a single failing column at ab, the only full-rank target, exactly as Proposition B.1 and Corollary B.2 predict.

Recht, 2009) for which we prove no analogue of Theorem 6. Consistently with that caveat, a small minority of cells do exhibit a short grokking-style gap. Under random-phase initialisation the four cells that are not INST are all borderline, with a gap between 200 and 500 steps and no clean GROK. Under roots-of-unity initialisation the thirteen are six clean GROK cells and seven borderline. Exact gaplessness is therefore no longer guaranteed once the encoding is learned.

The asymmetry between the two initialisations runs the other way from what one would expect. Starting the embedding at the analytic Fourier basis is the better-informed choice, and it is the one that produces every clean GROK cell in the sweep, while random phases produce none. The budgets also differed, 20000 steps against 40000, so we repeat the roots-of-unity arm at the matched 40000 steps with everything else unchanged. The verdicts barely move: 180 of the 190 solved cells are INST against 177 before, the ten that are not are two clean GROK and eight borderline against six and seven before, every degree still solves 38 of the 39 tasks, and ab still fails everywhere. Three cells cross from borderline to INST and four clean GROK cells soften to borderline, so the asymmetry is a property of the initialisation rather than of run length. What both sweeps agree on is the boundary: 38 tasks solved and ab failing at every degree under either initialisation, which is the prediction Proposition B.1 makes and the only claim the ablation is designed to test. The dominant outcome is nonetheless a simultaneous fit. What a learned encoding changes is primarily *which* targets are representable. The gapless character of the outcome largely, though not universally, persists. What does grow, smoothly with k, is the *delay before the fit appears*. The network must first discover embedding phases compatible with the target's rank-one structure, an acquisition process whose length increases with the activation degree: the step at which 90% of representable cells are solved rises from ≈ 3800 at $k = 2$ to ≈ 7900 at $k = 6$ (Table 6). This delay is the learned-encoding counterpart of the representation-acquisition phase that grokking performs in the one-hot setting. Here it precedes a mostly *simultaneous* train–test fit, not a delayed one. Figure 14 shows this directly for three representative cells: for the representable targets $1a + 1b$ and $a^2 + b^2$ the train and test loss fall together and first cross 99% accuracy within a single 10-step logging interval (gap ≤ 10 steps), after a delay that lengthens from $k = 3$ to $k = 6$, while ab never leaves the rank-$(k{+}1)$ floor $1 - (k + 1)/p$ of Corollary B.2 at either degree.

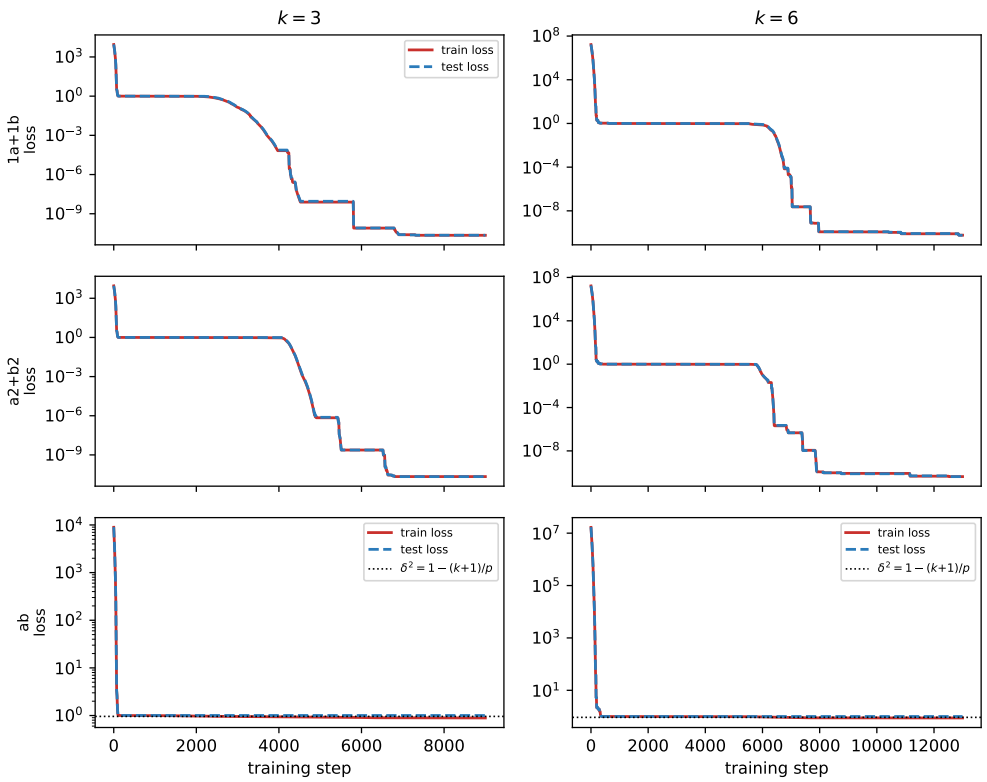

Figure 14: Train (red) and test (blue dashed) loss under a learned random-phase embedding, at 10-step resolution, for three representative cells and two degrees. Curves show best-so-far loss, which removes the post-convergence decoding jitter without altering the trajectory. For the representable targets $1a + 1b$ and $a^2 + b^2$ the train and test loss descend together and reach zero essentially simultaneously (they first exceed 99% accuracy within a single 10-step interval), after an acquisition delay that grows from $k = 3$ to $k = 6$. There is no memorisation-without-generalisation gap. For ab, which is unrepresentable at any degree, the loss plateaus at the rank-$(k+1)$ floor $1 - (k + 1)/p$ of Corollary B.2 (dotted line) and training accuracy never leaves chance.

Table 6: Encoding ablation. For each activation degree k: the number of tasks (out of 39) reaching INST under the fixed roots-of-unity encoding (the $m + n = k$ staircase of Figure 2) versus a learned complex embedding, and the training step at which 90% of representable cells are solved under LEARNED-RAND (LEARNED-ROOT is within a few hundred steps). Each cell is scored by the majority verdict over its three seeds, so the first two columns count tasks and not seeds, and the seed-level rates are in Table 4. Under the fixed encoding every solved cell is INST, so the second column counts both. Under a learned embedding the third column counts solved tasks: of the $38 \times 5 = 190$ solved cells, 186 are INST under LEARNED-RAND and 177 under LEARNED-ROOT, the remainder being GROK or borderline. The task ab fails under every encoding and every degree.

k	fixed (INST/39)	learned (solved/39)	step to 90% solved
2	3	38	3800
3	4	38	6100
4	5	38	7000
5	6	38	7200
6	7	38	7900

E.7 THE RANK CONDITION IS NECESSARY BUT NOT SUFFICIENT

Proposition B.1 shows $\operatorname{rank}(\omega^f) \le k + 1$ is necessary for representability under a learned encoding, and we proved it sufficient for rank-one targets. It is *not* sufficient in general: the reachable set $\mathcal{R}_k = \{\sum_l c_l\, e_1(\cdot)^l e_2(\cdot)^{k-l}\}$ is a Veronese/secant-type variety strictly contained in the matrices of rank $\le k + 1$ (Appendix B.1). A dimension count makes the generic form of this claim rigorous.

Proposition E.1. *Let $2 \le r \le k + 1$ and $p > (k + 5)/2$. The set of matrices of rank $\le r$ that lie in the closure $\overline{\mathcal{R}_k}$ is a proper Zariski-closed subvariety of the determinantal variety $\mathcal{M}_r = \{Y \in \mathbb{C}^{p\times p} : \operatorname{rank} Y \le r\}$. Consequently a generic rank-r target—any T outside a measure-zero subset of $\mathcal{M}_r$, e.g. almost surely under Gaussian factor sampling—has strictly positive distance to $\mathcal{R}_k$: the rank condition is necessary but not sufficient.*

Proof. $\mathcal{R}_k$ is the image of the polynomial map $\Psi : \mathbb{C}^p \times \mathbb{C}^p \times \mathbb{C}^{k+1} \to \mathbb{C}^{p\times p}$, $(e_1, e_2, c) \mapsto \big(\sum_l c_l\, e_1(a)^l e_2(b)^{k-l}\big)_{a,b}$, hence a constructible set whose Zariski closure is an irreducible variety of dimension at most $2p + k + 1$, the dimension of the source (Harris, 1992). For constructible sets over $\mathbb{C}$ the Zariski and Euclidean closures coincide, so membership in $\overline{\mathcal{R}_k}$ is exactly approximability by reachable outputs. The determinantal variety $\mathcal{M}_r$ is irreducible of dimension $r(2p - r)$ (Harris, 1992). For $r \ge 2$ the latter is minimised at $r = 2$, where $r(2p-r) = 4p-4 > 2p+k+1 \ge \dim \overline{\mathcal{R}_k}$ precisely when $p > (k + 5)/2$. Hence $\mathcal{M}_r \cap \overline{\mathcal{R}_k}$ has dimension strictly below $\dim \mathcal{M}_r$ and is a proper closed subvariety; its complement is Zariski-open and dense in $\mathcal{M}_r$ and has full measure under any absolutely continuous sampling. Any T in this complement lies outside the Euclidean closure of $\mathcal{R}_k$, so $\inf_{Y\in\mathcal{R}_k} \|Y - T\| > 0$. □

The proposition guarantees a positive residual generically but says nothing about its size. The following experiment validates the prediction numerically and quantifies the residual. For each target rank $r = 1, \dots, k + 2$ we draw generic complex rank-r targets, fit them with the learned-embedding reachable form, and compare against the unconstrained Eckart–Young rank-$(k+1)$ residual, which is zero whenever $r \le k + 1$. Settings are in Appendix F.

Figure 15 shows the outcome. Rank-one targets are fit to $\lesssim 10^{-8}$, a positive control confirming the optimiser reaches the global optimum, so a nonzero residual elsewhere signals genuine unreachability and not optimisation failure. For $2 \le r \le k + 1$ the Eckart–Young residual is zero (the rank bound is satisfied) yet the learned-embedding residual is bounded away from zero and grows with r: generic targets that meet the rank condition are nonetheless outside $\mathcal{R}_k$. For $r > k + 1$ both residuals are positive, the ordinary rank barrier. The phenomenon is degree- and p-independent. The emphasis on *generic* targets is essential, since special rank-r targets whose factors are consecutive powers of a single embedding remain reachable. This is the experimental counterpart of the secant-variety caveat in Appendix B.1, made rigorous for generic targets by Proposition E.1.

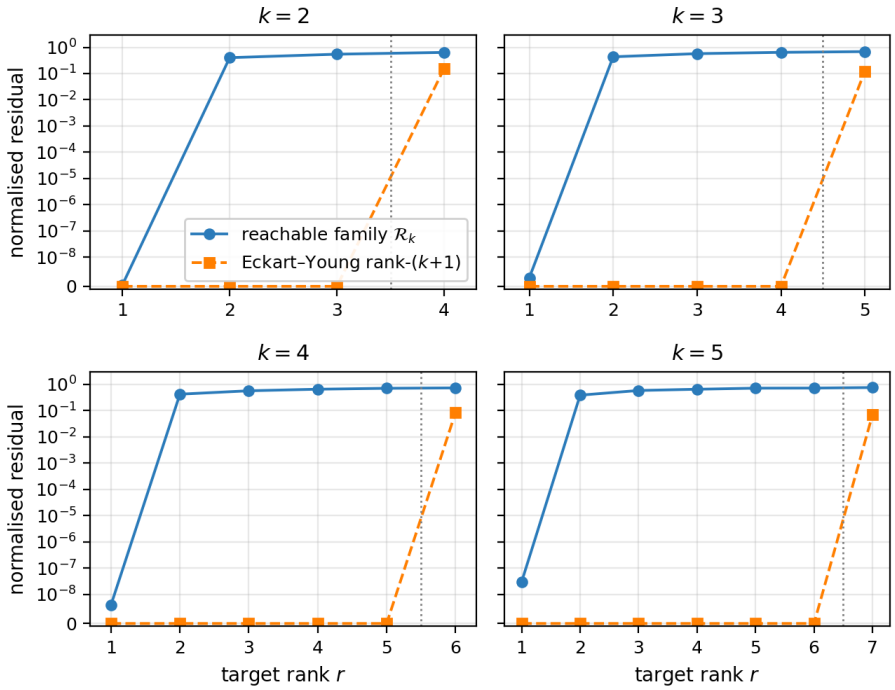

Figure 15: **Rank is necessary, not sufficient.** Generic complex targets of rank r fit with the learned-embedding reachable form $\mathcal{R}_k$ (blue) against the unconstrained Eckart–Young rank-$(k{+}1)$ residual (orange), median over 8 targets per rank. Rank-one fits to $\lesssim 10^{-8}$, a positive control. For $2 \le r \le k+1$, left of the dotted line, the rank bound is met and the Eckart–Young residual is numerically zero, yet the target is unreachable and the blue residual grows with r. For $r > k+1$, right of the line, the ordinary rank barrier is active and both are positive. Shown for $k = 2, \ldots, 5$ at $p = 97$; the pattern is p-independent.

E.8 Bottleneck sweep

The sweep of Section 5.3 inserts a linear bottleneck of width d between the one-hot input and a width-256 ReLU layer, so that d caps the rank of the feature map while leaving everything else at the ReLU baseline settings. Seven of the eight cells are unanimous across their eight seeds. The exception is $d = 8$, which splits four MEM to four PART with final test accuracies spanning 0.006 to 0.862, a spread no other cell approaches. At $d = 4$ the test curves have stopped climbing by the end of the budget: fitting a line to the last fifth of each run gives a gain below 0.021 over 10,000 steps in every seed, so that cell is not a GROK run censored by the budget. Retraining the four representative cells of Figure 17 from the recorded settings on a later PyTorch release reproduces the sweep exactly: the fit steps (18500 at $d = 4$, 2000 at $d = 8$, 750 at $d = 16$), the gap (15000 at $d = 16$) and the final test accuracies all agree to the logged precision, so the plotted curves are the runs the table counts. Figure 16 plots the gap against d with interquartile ranges, Table 7 records the verdicts and medians, and Figure 17 shows one representative pair of curves per regime.

The 200000-step check. The grokking gap grows steeply as d falls toward 16, so the outcomes at $d \in \{4, 8\}$ could be delays the 50000-step budget cuts off rather than terminal states. We rerun both widths at 200000 steps, all eight seeds, everything else unchanged. The two cells behave differently.

At $d = 4$ nothing moves. Every seed stays PART, the median final test accuracy goes from 0.747 to 0.760, the per-seed values run 0.742 to 0.879, and no seed has a positive trend over the last fifth of the run by the slope test above. All eight seeds in fact end below their own maximum, by 0.0004 to 0.0124, so test accuracy peaks inside the budget and then settles back. Quadrupling the budget buys 0.013 of median accuracy, so this is a ceiling and not a delay.

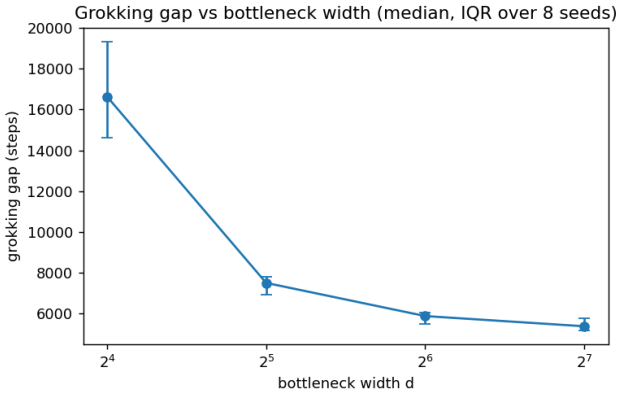

Figure 16: Grokking gap versus bottleneck width d (median and interquartile range over eight seeds, GROK cells only). The gap decreases monotonically as capacity grows, with non-overlapping interquartile ranges between $d = 16$ and $d \geq 32$.

Table 7: Bottleneck sweep: outcome, median step at which the training threshold is crossed, median final test accuracy, and median generalisation gap with interquartile range, as the bottleneck dimension d varies. Eight seeds per d, at the 50000-step budget. Every cell is unanimous except $d = 8$, which splits four to four. Rerunning $d \in \{4, 8\}$ at 200000 steps leaves $d = 4$ unchanged and moves $d = 8$ to six PART against two MEM (Appendix E.8).

d	verdict	fit step	final test	median gap	IQR (25–75%)
1	FAIL	—	0.19	—	—
2	FAIL	—	0.76	—	—
4	PART	12,500	0.75	—	—
8	MEM ×4, PART ×4	2000	0.04	—	—
16	GROK	750	1.00	16,600	14,600–19,300
32	GROK	500	1.00	7500	6900–7800
64	GROK	500	1.00	5900	5500–6100
128	GROK	250	1.00	5400	5200–5750

At $d = 8$ the picture changes. The median rises from 0.036 to 0.838 and the MEM count falls from four seeds to two. Of the four seeds at chance after 50000 steps, two climb to 0.566 and 0.798, the first still rising steeply when the run ends. The other two, seeds 1 and 7, finish at 0.007 and 0.008, and over the whole 200000 steps their test accuracy never exceeds 0.0096, short of the chance rate 0.0103 itself. The remaining four seeds, already above chance at 50000 steps, reach 0.879 to 0.949. Most of the $d = 8$ dip in Figure 18 is therefore an artefact of the shorter budget, and what survives is the seed dependence together with two seeds that do not escape.

The escape times spread over two orders of magnitude. Taking the step at which test accuracy first passes 5%, the six escaping seeds leave at 1250, 10750, 12500, 48750, 85500 and 116750 steps, and two never leave inside the budget. The last of the six is still climbing steeply when the run ends, gaining 0.54 over the final fifth, so 200000 steps settles the median but not every seed. A spread of this shape is what a heavy-tailed escape time produces, and Corollary D.16 derives one for the learned-embedding model; that analysis does not cover this architecture, so we record the resemblance without claiming the mechanisms are the same.

Neither width reaches the 99% test threshold at 200000 steps, so no GROK verdict appears at $d \leq 8$ under a fourfold budget. We report the 50000-step verdicts in Table 7 because they are the ones the rest of the sweep shares, and record the extension here.

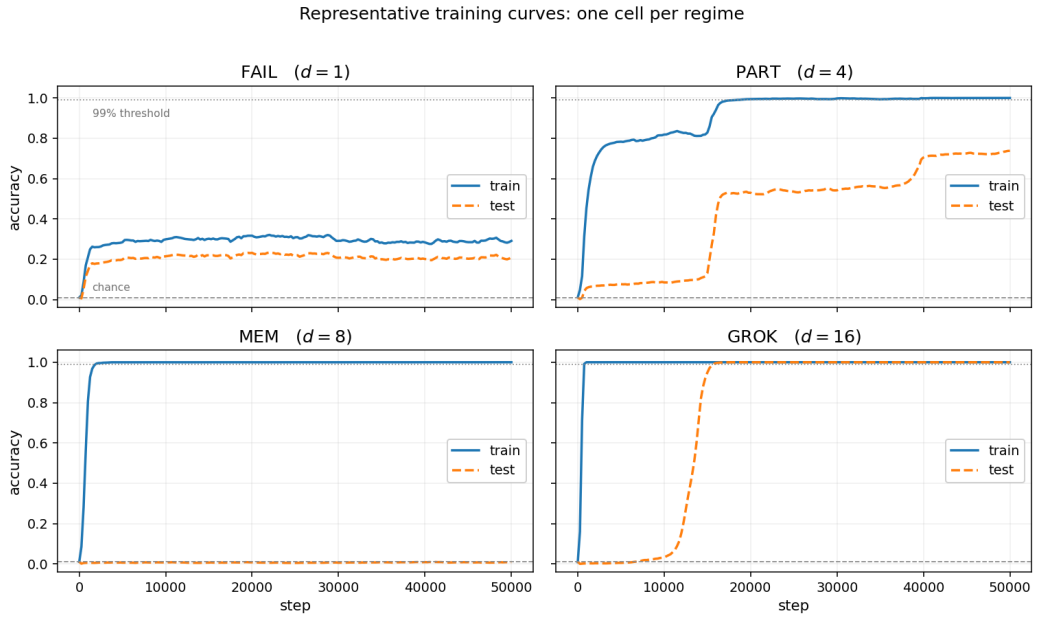

Figure 17: Representative train (solid) and test (dashed) accuracy curves, one cell per regime. FAIL ($d = 1$): training never crosses the 99% threshold (upper dotted line) and test tracks it closely. Panels share both axes; the lower dashed line is the chance rate $1/97$. PART ($d = 4$): training fits after about 12,500 steps and test accuracy climbs in steps to a plateau near 0.74, far above the chance rate (lower dashed line). MEM ($d = 8$): training fits within 2000 steps while test accuracy never leaves chance, peaking at 0.0101 against a chance rate of 0.0103 and ending at 0.0090 with a negative trend over the last fifth of the run. GROK ($d = 16$): training fits early and test follows only after the long generalisation delay.

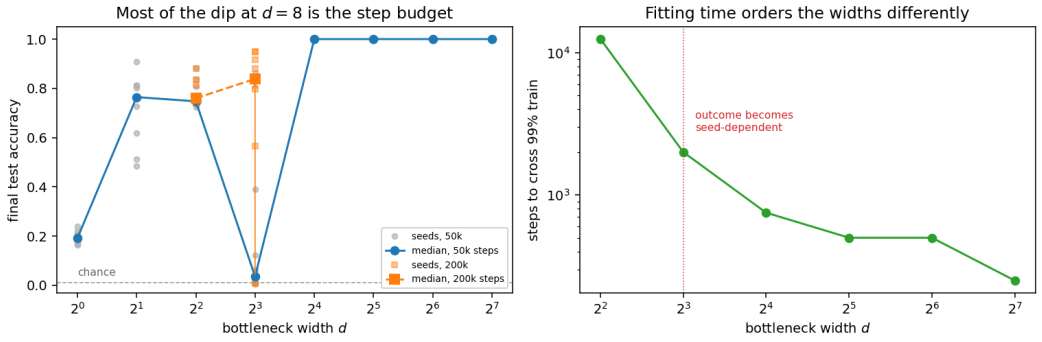

Figure 18: Left: final test accuracy against bottleneck width, medians joined and per-seed values as points, at both budgets. At 50000 steps (circles) the dip at $d = 8$ falls below the value at $d = 2$. At 200000 steps (squares, arrows marking the change) the $d = 4$ median moves only from 0.747 to 0.760 while $d = 8$ rises from 0.036 to 0.838, so the $d = 4$ plateau is a ceiling and most of the $d = 8$ dip is the budget. Two $d = 8$ seeds stay at chance under both (Appendix E.8). Right: median step at which training accuracy first crosses 99%, on a log scale, over the widths that cross it at all. The width whose outcome depends on the seed is the narrowest that fits the training set within 2000 steps.

target	optimiser	$\rho = \tau(k{=}6)/\tau(k{=}2)$	trend
$1a + 1b$	vanilla GD	0.27	decreases
	AdamW, $wd = 10^{-3}$	2.47	grows
	AdamW, $wd = 10^{-2}$	2.54	grows
	AdamW, $wd = 10^{-1}$	2.57	grows
$1a + 2b$	vanilla GD	0.27	decreases
	AdamW, $wd = 10^{-3}$	2.33	grows
	AdamW, $wd = 10^{-2}$	2.31	grows
	AdamW, $wd = 10^{-1}$	2.44	grows
$2a + 3b$	vanilla GD	0.27	decreases
	AdamW, $wd = 10^{-3}$	2.60	grows
	AdamW, $wd = 10^{-2}$	2.56	grows
	AdamW, $wd = 10^{-1}$	2.61	grows
$1a + 5b$	vanilla GD	0.27	decreases
	AdamW, $wd = 10^{-3}$	2.20	grows
	AdamW, $wd = 10^{-2}$	2.20	grows
	AdamW, $wd = 10^{-1}$	2.34	grows

Table 8: Robustness of the optimiser-induced reversal. Learned-embedding model, $H = 128$, $k \in \{2, \dots, 6\}$, 8 seeds per cell, acquisition threshold 0.99 train accuracy, no cell censored. $\rho > 1$ means the acquisition delay grows with the activation degree. The sign is set by the optimiser in all sixteen cells, and under AdamW it is insensitive to the weight decay, so the growth is not a regularisation effect. Under gradient descent ρ is independent of the target, as Corollary D.16 predicts for factorable targets. Its value 0.27 exceeds the gradient-flow ratio 0.11, so the decrease is slower than the idealised flow (see text).

E.9 Optimiser reversal

Section 5.4 changes the optimiser on one target at one weight decay. Table 8 repeats that comparison over four factorable targets and three weight decays, with the model, the width, the seeds and the acquisition threshold held fixed, so that the optimiser and the weight decay are the only quantities that move.

F Reproducibility details

This appendix collects the implementation details needed to reproduce the experiments.

Initialisation. All complex weights are circularly-symmetric complex Gaussian: $\mathbb{E}|W_{ij}|^2 = 1$ (real and imaginary parts i.i.d. $\mathcal{N}(0, \frac{1}{2})$) and $\mathbb{E}|v_j|^2 = 1/H$, matching the convention of Appendix C. Learned embeddings are initialised either as random unit-modulus phases (LEARNED-RAND) or at the roots-of-unity values (LEARNED-ROOT).

Decoding and run classification. Predicted residue: the nearest root of unity,

$$c = \arg\min_{c \in \mathbb{Z}_p} \left| \hat{y}(\omega^a, \omega^b) - \omega^c \right| = \left(\lfloor p \arg \hat{y}/2\pi \rceil \right) \bmod p,$$

i.e. $\arg \hat{y}$ rounded to the nearest integer multiple of $2\pi/p$, with ties resolved by round-half-to-even (`torch.round`). Run labels (INST/GROK/PART/MEM/FAIL) use the thresholds of Section 5.1. Best-so-far smoothing is applied only for display in Figure 14. Classification uses the raw per-log-step train and test accuracies.

Budgets and optimiser. Main phase-diagram grid (the 585 runs = 39 tasks $\times$ five degrees $k \in \{2, \dots, 6\} \times 3$ seeds): full-batch AdamW, learning rate 10^{-3}, weight decay 10^{-2}, 40% train fraction (data split under seed $1000 + \text{seed}$), gradient clipping at norm 1.0, accuracy logged every 100 steps. Step budget by encoding: 5000 steps for the fixed roots-of-unity encoding, 20000 for learned embeddings initialised at the roots, and 40000 for randomly initialised learned embeddings. All

degrees $k = 2, \ldots, 6$ use the same per-mode budget. The single $a + 5b$ shortfall at $k = 6$ is re-run to 8000 steps. The fine-resolution learned-embedding traces of Figure 14 are logged every 10 steps with budgets 9000 ($k = 3$) and 13000 ($k = 6$). ReLU baseline: width 256, cross-entropy, AdamW, learning rate 10^{-3}, weight decay 1.0, 8000 steps (matching the grokking-literature recipe of Appendix E.4). Bottleneck sweep: target $f(a, b) = a + b \bmod 97$, one-hot input, bottleneck widths $d \in \{1, 2, 4, 8, 16, 32, 64, 128\}$ feeding a width-256 ReLU hidden layer, cross-entropy, AdamW, learning rate 10^{-3}, weight decay 1.0 (applied to all parameters, including the bottleneck layer), 50000 steps, accuracy logged every 250 steps, eight seeds per width, data split under seed 123. The extension of Appendix E.8 repeats $d \in \{4, 8\}$ at 200000 steps with every other setting unchanged, 71.6 min on an A100 for its sixteen runs. The matched-budget check of Appendix E.6 repeats the roots-of-unity arm of the encoding ablation at 40000 steps rather than 20000, 50.8 min on an A100 for its 585 runs. Every cell is unanimous across seeds except $d = 8$, which splits four MEM to four PART (Table 7).

Optimiser comparison (Section 5.4 and Table 8). Learned-embedding model, $H = 128$, $p = 97$, 40% train fraction, eight seeds per cell, acquisition threshold 0.99, gradient clipping at norm 1.0 in both arms. AdamW arm: learning rate 10^{-3}, weight decay 10^{-2} for the single-configuration run and $10^{-1}, 10^{-2}, 10^{-3}$ across the grid, budget 40000 steps. Vanilla gradient-descent arm: full-batch SGD with learning rate 2×10^{-1}, zero momentum, zero weight decay, budget 200000 steps. The gradient-descent learning rate was raised from 5×10^{-2}, at which seven of eight seeds at $k = 2$ reached the budget without acquiring; the same value is then used at every k, so it rescales the clock uniformly and leaves the ratio ρ unaffected (Corollary D.17). Acquisition is evaluated every 50 steps in the single-configuration run and every 100 in the batched grid, and the ratio ρ in Table 8 is the median acquisition step at $k = 6$ divided by the median at $k = 2$, each median taken over the eight seeds of that cell on its own evaluation grid. No seed is censored in either arm at these settings.

Convex baseline (Table 5). Fixed roots-of-unity encoding, $H = 128$, $p = 97$, full domain, Adam at learning rate 3×10^{-3}, budget 5000 steps with early stopping at training loss below 10^{-8}, one run per degree under a single global seed. The convex arm is a direct least-squares solve in the coefficient space of $\mathcal{F}_k$ and carries no optimiser.

Generic-rank experiment. For each target rank $r = 1, \ldots, k+2$ we draw 8 generic complex rank-r targets $T = UV^\dagger$ ($U, V \in \mathbb{C}^{p \times r}$ with i.i.d. complex Gaussian entries), each Frobenius-normalised. Every target is fit with the learned-embedding reachable form directly: $e_1, e_2 \in \mathbb{C}^p$ are trainable (each real and imaginary part initialised i.i.d. $\mathcal{N}(0, 0.7^2)$) and the coefficient vector c is solved in closed form at every step by regularised least squares, $c = (\Phi^\dagger \Phi + 10^{-8} I)^{-1} \Phi^\dagger T$, which is differentiable in e_1, e_2. Parametrising $\mathcal{R}_k$ directly ($2p$ complex, hence $4p$ real, parameters) is far better conditioned than the full (W, v) network, so a positive residual signals unreachability and not optimisation failure. The embeddings are optimised with Adam at learning rate 2×10^{-2} for 20000 steps. We report the best of 12 random restarts and the median normalised residual over the 8 targets. The Eckart–Young baseline is the rank-$(k{+}1)$ singular-value tail energy $\sum_{i>k+1} \sigma_i^2 / \sum_i \sigma_i^2$.

Compute. All experiments run on a single GPU (NVIDIA A100-SXM4-40GB or Tesla T4). The batched encoding ablation trains all task×seed replicas of a given degree in one network, at roughly 10–45 s per degree on an A100. The full phase-diagram grid completes in 3–5 min and the encoding-comparison grid in 10–15 min. The bottleneck sweep dominates the budget at 70 min on an A100 for its 64 runs, so all experiments in the paper together come to about two GPU-hours. Runs use PyTorch 2.x with TF32 matrix multiplication left at the PyTorch default.

G Extended related work

Grokking and capacity

Grokking was introduced by Power et al. 2022, who observed that networks trained on small algorithmic datasets would memorise first and generalise much later, with the delay sensitive to weight decay and training fraction. Nanda et al. 2023 used mechanistic interpretability to trace the internal computation that emerges at the transition, identifying a Fourier-based algorithm that networks

discover when they generalise on modular addition. Liu et al. 2022 proposed a representation-learning account, identifying four learning phases, comprehension, grokking, memorisation, and confusion, determined by model size and weight decay. Their work also noted, without developing the point, that very small models may skip grokking and generalise immediately. Kumar et al. 2024 linked grokking to the transition from lazy (Jacot et al., 2018) to rich training dynamics, showing that the grokking transition coincides with a change in the effective alignment between learned features and the target function. Two theoretical accounts are closest to ours in setting. Mohamadi et al. 2024 prove that no kernel-regime model learns modular addition without seeing a constant fraction of the data, and that two-layer *quadratic* networks escaping the kernel regime generalise from far fewer points, a degree-two case adjacent to our monomial family, though they study sample complexity and not the exact expressible class. Lyu et al. 2024 induce grokking through a dichotomy of early- and late-phase implicit biases, the network lingering near a kernel predictor before a sharp transition to a min-norm solution.

Song and Ye 2026 give the most direct account of how model capacity shapes grokking timing. They separate grokking delay into a memorisation timescale and a generalisation timescale, both measurable functions of parameter count, and show that grokking emerges near the scale where these timescales intersect. Their analysis assumes throughout that the network is large enough to memorise the training data. The regime where memorisation is structurally impossible lies outside their framework. Our work sits in that regime, and Section 5.3 traces the continuous path from it to the capacity spectrum that Song and Ye characterise.

Rubin et al. 2023 model grokking as a first-order phase transition in two-layer networks, using statistical mechanics to identify competing basins corresponding to the memorised and generalised solutions. Žunkovič and Ilievski 2024 study solvable grokking models and derive closed-form critical exponents and grokking time distributions. Cullen et al. (2026) give a singular-learning-theory account of the same phenomenon, ranking the competing memorising and generalising basins by their local learning coefficient. Varma et al. 2023 attribute the transition to a competition between memorising and generalising circuits resolved by parameter efficiency. Boursier et al. 2025 derive grokking from optimisation alone, as a two-phase gradient flow that first interpolates and then, under small weight decay, drifts along a Riemannian norm-minimising flow. And Humayun et al. 2024 show empirically that grokking of the decision boundary is near-universal in deep networks. These analyses, like most in the grokking literature, operate in the universal-approximation regime where the network can always fit the training data.

Modular arithmetic and analytical solutions

Gromov 2023 derived closed-form weight solutions for two-layer MLPs on modular addition and showed that trained networks converge to qualitatively similar solutions at the grokking transition. Doshi et al. 2024 extended this to modular multiplication and general modular polynomials, and hypothesised a classification of polynomials into learnable and non-learnable classes that takes the same form as our Theorem 2. Their hypothesis is supported empirically but not proved, and they note the absence of a proof as an open challenge. The architectures differ in an important way: Doshi et al. use one-hot inputs with a real power activation, which is a universal approximator on any finite input set and therefore always capable of memorisation. In their setting, failure means memorisation without generalisation. In ours, it means the network cannot even fit the training data. Our Theorems 2 and A.5 prove the classification in the roots-of-unity-encoded holomorphic setting, and Theorem 6 converts non-representability into a width-independent training-loss floor.

Neural network expressivity

Universal approximation results (Cybenko, 1989; Hornik, 1991) establish that sufficiently wide networks with standard activations can approximate any target function, but say nothing about what a fixed architecture can express exactly. Work on depth separation (Telgarsky, 2016) and polynomial networks (Kileel et al., 2019) shows that specific architectural choices produce hard expressivity constraints. Our contribution is an exactly-computed expressible class for a specific holomorphic architecture, stated as a Fourier-domain membership condition and proved in both directions, together with the use of that characterisation as a pre-training *diagnostic*. Expressivity constraints are converted

into a finite-sample loss floor and a trajectory-level train–test coupling that decide, in advance, which learning regime a given task can occupy.