FEATURE SPARSITY VIA MARGIN MAXIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Explaining the number of features in a network has been a long-desired result in mechanistic interpretability. Morwani et al. (2024) laid the baseline through their margin-maximization analysis, proving that quadratic networks maximizing margin under matched $L_{2,\nu}$ regularization use every frequency for modular addition. This restriction, however, excludes most models conventionally trained using L_2 regularization or a non-quadratic architecture, which are empirically sparse.

Our inquiry focuses specifically on modular addition in polynomial and multilinear models. Our primary theoretical finding centers on the cost of new frequencies—a trade-off between the margin of the solution and regularization. For interference-free product networks, we use this to characterize the maximum normalized margin and an optimal number of frequencies for a particular configuration of homogeneity degree (ν) and regularization. This recovers the dense frequency allocation seen in Morwani et al. (2024).

Beyond the interference-free class of models, we also prove upper and lower bounds on the maximum normalized margin of product networks. We also establish matching bounds on the optimal number of frequencies, and prove that an optimal spectral allocation uses $\Theta(\log p)$ frequencies. Finally, we prove a general asymptotic scaling law for normalized margin over product models.

1 INTRODUCTION

Morwani et al. (2024) use margin maximization (Wei et al., 2019; Lyu & Li, 2020) to characterize exactly the features of maximum-margin networks on algebraic tasks, and to bound the margin of the network. However, their arguments remain limited to the specific regularization ($L_{2,\nu}$, matched to the homogeneity degree ν) and architecture (quadratic models) chosen and remain hard to extend to more conventional regularizations; Li et al. (2025) extend the modular-addition result to n-operand power networks, again at the matched norm $L_{2,n+1}$. Their theory shows that under this configuration, the margin maximal solution would use all frequencies (dense) in the solution.

However, as observed by Nanda et al. (2023), Chughtai et al. (2023) and Zhong et al. (2023), under different architectures and regularizations, a sparse subset of frequencies (or irreducible representations) are utilized by the model, while shallow networks on one-hot inputs can use all of them (Morwani et al., 2024; McCracken et al., 2025). McCracken et al. (2025) similarly see the number of features transition from dense to sparse when depth or a trainable embedding is added to the network.

This paper generalizes the maximum-margin analysis of Morwani et al. (2024) to every $L_{2,b}$ norm with $2 \leq b \leq \nu$, including ordinary weight decay ($b = 2$), and across multilinear and polynomial architectures (homogeneity degree $\nu = n + 1$) for every network where there is no interference at the frequencies they use (no component of the logits at a used frequency that depends on anything other than the offset between a candidate and the correct answer; Definition 5).

This paper shows:

The explanation of the sparsity phenomenon and the optimal number of frequencies for a certain configuration

A global upper and lower bound on the normalized margin of models

A general scaling law for the optimal number of frequencies for a certain prime p (cf. the logarithmic growth observed empirically by McCracken et al. (2025))

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.

2 PRELIMINARIES

We consider one-hidden-layer networks with homogeneous neurons and no biases, $f(\theta, x) = \sum_{i=1}^{m} \phi(\omega_i, x)$ with $\theta = \{\omega_1, \ldots, \omega_m\}$, where $\phi(\lambda\omega, x) = \lambda^\nu \phi(\omega, x)$ for all $\lambda > 0$; ν is the *homogeneity degree*. For the networks below, a neuron is $\omega_i = (u_{i1}, \ldots, u_{in}, w_i) \in (\mathbb{R}^p)^{n+1}$, and $\|\omega_i\|_2$ is the Euclidean norm of the stacked vector.

The $L_{2,b}$ norm is $\|\theta\|_{2,b} = \big(\sum_i \|\omega_i\|_2^b\big)^{1/b}$: an ℓ_2 norm within each neuron and an ℓ_b norm across neurons, as in Morwani et al. (2024).

Modular addition For a prime $p > 2$ and $n \geq 1$ operands, the dataset is $D_p := \{((a_1, \ldots, a_n), a_1 + \cdots + a_n \bmod p) : a_1, \ldots, a_n \in \mathbb{Z}_p\}$; for an input x, $y(x)$ is its label, and $f(\theta, x)[c]$ is the logit of candidate $c \in \mathbb{Z}_p$. Inputs are one-hot, so the embedding vector $u \in \mathbb{R}^p$ acts on operand a as $u(a)$. The *product neuron* is $\phi(\omega_i, x)[c] = w_i(c) \prod_{t=1}^{n} u_{it}(a_t)$ and the *power neuron* is $\phi(\omega_i, x)[c] = w_i(c)\big(\sum_{t=1}^{n} u_{it}(a_t)\big)^n$, both with $\nu = n + 1$.

Fourier transform. For $v : \mathbb{Z}_p \to \mathbb{C}$, $\hat{v}(j) := \sum_k v(k)\rho^{-jk}$ with $\rho := e^{2\pi i/p}$. For real v, $\hat{v}(-j) = \overline{\hat{v}(j)}$, so the distinct non-zero frequencies are $F := \{1, \ldots, \frac{p-1}{2}\}$. The multidimensional transform of the logit table is $\hat{f}_\theta(k_1, \ldots, k_n; l) := \sum_{x,c} f(\theta, x)[c]\, \rho^{-(k_1 a_1 + \cdots + k_n a_n + lc)}$; for a product neuron it factorizes as $\hat{w}_i(l) \prod_{t=1}^{n} \hat{u}_{it}(k_t)$.

3 THEORETICAL APPROACH

Maximum margin As in Morwani et al. (2024), the margin of a datapoint x where the correct answer is y is $g(\theta, x, y) = f(\theta, x)[y] - \max_{y' \neq y} f(\theta, x)[y']$. The margin of the dataset is $h(\theta) = \min_{(x,y)\in D_p} g(\theta, x, y)$, and the maximum margin under $L_{2,b}$ norm is $\gamma_b^* := \max_{\|\theta\|_{2,b} \leq 1} h(\theta)$. By homogeneity the normalized margin of any θ is $h(\theta)/\|\theta\|_{2,b}^\nu$. By Wei et al. (2019, Thm. 4.1; Thm. C.1 for multi-class cross-entropy), the global minimizers of cross-entropy plus $\lambda\|\theta\|_{2,b}^b$ attain normalized margin tending to γ_b^* as $\lambda \to 0$. Hence, we study maximum margin solutions as ideal, theoretically achievable solutions for the network, without considering empirical achievability. Gradient flow on cross-entropy converges in direction to KKT points of the L_2 margin problem, which need not be global maxima (Lyu & Li, 2020; Vardi et al., 2022); with small weight decay see Lyu et al. (2024).

3.1 AVERAGED MARGINS AND FOURIER AMPLITUDES

The minimum margin is at most any averaged class-weighted margin (Morwani et al., 2024, Section 3.2); see also Li et al. (2025, Def. 3.6). We average over D_p and put all weight on the candidate with offset δ from the answer. Let

$$\eta_\theta(\delta) := \mathbb{E}_x\big[f(\theta, x)[y(x) - \delta]\big], \qquad \delta \in \mathbb{Z}_p,$$

Lemma 1 (average margin)**.** *For every θ and $\delta \neq 0$, $h(\theta) \leq \eta_\theta(0) - \eta_\theta(\delta)$.*

Proof. For every x, $y(x) - \delta$ is an incorrect candidate, so $f(\theta, x)[y(x)] - f(\theta, x)[y(x) - \delta] \geq h(\theta)$. Average over x. □

Averaging over inputs keeps only the part of the logit table that is a function of the offset $y(x) - c$ between label and candidate. Write $f(\theta, x)[c] = \eta_\theta(y(x) - c) + r_\theta(x, c)$. The remainder r_θ averages to zero over inputs at every fixed offset, so it carries no systematic preference for the correct answer; it is termed *interference* ('destructive interference' seen in Gromov, 2023). Interference cannot raise the margin: a network whose table is $\eta_\theta(y(x) - c)$ alone has margin exactly $\min_{\delta \neq 0}\big(\eta_\theta(0) - \eta_\theta(\delta)\big)$, and by Lemma 1 adding r_θ can only lower it.

η_θ is a real function on $\mathbb{Z}_p$, hence a sum of cosines (Appendix A.1):

$$\eta_\theta(\delta) = \bar{\eta}_\theta + \sum_{j \in F} \mathrm{Re}\big(\beta_j \rho^{j\delta}\big), \qquad \beta_j := 2p^{-(n+1)} \hat{f}_\theta(j, \ldots, j; -j). \tag{1}$$

We call $A_j := |\beta_j|$ the *amplitude* of frequency j; a frequency with $A_j > 0$ is a *feature*; this is the exact counterpart of the 'key frequencies' of Nanda et al. (2023) and of the frequencies neurons 'use' in Morwani et al. (2024). Since f is a sum over neurons, $\beta_j = \sum_i \beta_{ij}$ with β_{ij} computed from neuron i alone; for a product neuron $\beta_{ij} = 2p^{-\nu}\overline{\hat{w}_i(j)}\prod_{t=1}^{n}\hat{u}_{it}(j)$. Frequency j separates the label from the candidate δ away by the *coverage* $\kappa_j(\delta) := 1 - \cos(2\pi j\delta/p)$, which satisfies $\sum_{\delta\neq 0}\kappa_j(\delta) = p$ and is small where $j\delta$ is near a multiple of p. Hence, several frequencies are needed to cover every offset to maximize margin.

Lemma 2 (amplitude bound)**.** *For every θ and $\delta \neq 0$, $h(\theta) \leq \sum_{j\in F}\mathrm{Re}(\beta_j)\,\kappa_j(\delta) \leq \sum_{j\in F} A_j\,\kappa_j(\delta)$. (Proof: Appendix A.2.)*

The margin is thus bounded from above by the amplitudes of the frequencies alone, and hence, is largest when each β_j is real and positive and each frequency contributes a cosine peaked at the label. Averaging over δ recovers the uniformly class-weighted margin of Morwani et al. (2024, App. F.1); for aligned cosine logits the right-hand side is the familiar separation of Nanda et al. (2023, Sec. 4.2) and McCracken et al. (2025, Thm. 4.7).

3.2 The budget-splitting problem

For a model with homogeneity degree ν and regularization $L_{2,b}$, let $s := \nu/b \in [1, \nu/2]$.

The price of a feature. In a product neuron, the amplitude A of a frequency is proportional to the product of the Fourier magnitudes of the neuron's ν weight vectors at that frequency, while its $L_{2,b}$ penalty is the sum of their squared norms, raised to the power $b/2$. At a fixed product, the sum of squares is least when the ν factors have equal norm (AM–GM). So the least penalty that buys amplitude A is a constant multiple of $A^{b/\nu} = A^{1/s}$ (Appendix A.3).

A power neuron $w(c)\big(\sum_t u_t(a_t)\big)^n$ is not a product, but only the multilinear term $n!\,w(c)\prod_t u_t(a_t)$ of its expansion reaches the modes $(j,\dots,j;-j)$ that carry A (Appendix C). Its amplitude is therefore $n!$ times the same product of Fourier magnitudes, and the same price holds with a different constant. This is the familiar fact that ℓ_2 weight decay on a product of k factors acts as an $\ell_{2/k}$ quasi-norm on the product (Hoff, 2017; Tibshirani, 2021; Dai et al., 2021; Kolb et al., 2025), found in the Fourier domain for linear convolutional networks by Gunasekar et al. (2018); here the exponent is $1/s = b/\nu \leq 1$, and the new ingredient is its interaction with the coverage of offsets.

Splitting the budget. A network therefore splits one unit of budget over frequencies: spending $y_j^{1/s}$ on frequency j buys amplitude y_j, and by Lemma 2 the margin is at most the coverage of the worst offset. Since $1/s \leq 1$, concentrating the budget is discounted, which favours few features, while every offset must be covered, which favours many. The best split is

$$\Gamma_s(p) := \max\Big\{\min_{\delta\neq 0}\sum_{j\in F} y_j\,\kappa_j(\delta)\ :\ y\in\mathbb{R}^F_{\geq 0},\ \sum_{j\in F} y_j^{1/s}\leq 1\Big\}. \tag{2}$$

For a maximizer y^* we write $K^*(s,p) = |\mathrm{spt}(y^*)|$ for its number of features.[1] Equivalently, Γ_s is the largest spectral gap of a weighted circulant graph on $\mathbb{Z}_p$ under an $\ell_{1/s}$ budget on edge weights; at $s = 1$ this is convex edge-weight design (Boyd, 2006; Sun et al., 2006), and minimizing $\sum_j v_j^{1/s}$ over a polyhedron is a standard concave (ℓ_q, $q < 1$) problem (Ge et al., 2011). Concurrently, Kam et al. (2026) study the same worst-offset separation under an ℓ_2 budget, where the full-support optimum is uniform.

Proposition 3 (coverage bound)**.** *With $\|A\|_{1/s} := \big(\sum_j A_j^{1/s}\big)^s$, every θ satisfies $h(\theta) \leq \Gamma_s(p)\,\|A\|_{1/s}$.*

Proof. If $A = 0$, Lemma 2 gives $h(\theta) \leq 0$. Otherwise $y := A/\|A\|_{1/s}$ is feasible for equation 2; apply Lemma 2 and multiply by $\|A\|_{1/s}$. □

[1] We have not proved that all maximizers have the same support size; every statement is made for "a maximizer" and holds for each.

Let $\beta_j(\omega)$ be the coefficient of the single neuron $\phi(\omega, \cdot)$ and

$$C_b := \max_{\|\omega\|_2 \le 1} \sum_{j \in F} |\beta_j(\omega)|^{b/\nu},$$

so that, by homogeneity, $\sum_j |\beta_j(\omega)|^{b/\nu} \le C_b \|\omega\|_2^b$ for every neuron. For the product neuron $C_b^{\nu/b} = B_p := 2(2p\nu)^{-\nu/2}$ for all $2 \le b \le \nu$, and for the power neuron $C_b^{\nu/b} = n!\, B_p$ (Appendix C). At $b = \nu$ this is the single-neuron computation of Morwani et al. (2024, App. F.1) ($n = 2$) and Li et al. (2025, Lemmas D.7–D.8); at $b = 2$ it is the Fourier-domain representation cost of Gunasekar et al. (2018, Lemma 7).

Theorem 4 (universal bound)**.** *For any one-hidden-layer network with homogeneity constant ν, any width, and any $2 \le b \le \nu$, $\gamma_b^* \le C_b^{\nu/b}\, \Gamma_{\nu/b}(p)$.*

Proof. Let $r = b/\nu \le 1$. As $t \mapsto t^r$ is subadditive and $A_j \le \sum_i |\beta_{ij}|$, $\sum_j A_j^r \le \sum_{i,j} |\beta_{ij}|^r \le C_b \|\theta\|_{2,b}^b$. Raise to the power $1/r$ and apply Proposition 3. Taking the maximum with $\|\theta\|_{2,b}$ gives the result. □

At $b = \nu$ the bound recovers the results of Morwani et al. (2024, Thm. 7) and Li et al. (2025, Thm. 4.1): $\Gamma_1(p) = p/(p-1)$ with equal amplitude on every frequency, and the resulting value is exactly their maximum margin (Corollary 14, Appendix A.4). For $b < \nu$, including weight decay, we are not aware of an earlier upper bound.

4 Modular addition

Interference at the feature's own frequency. A product neuron builds $\cos(2\pi j(a_1 + \cdots + a_n - c)/p)$ by multiplying one cosine per operand, and the product-to-sum formulas also produce *sign-flipped* copies: for $n = 2$, $\cos\frac{2\pi ja}{p}\cos\frac{2\pi jb}{p} = \frac{1}{2}\cos\frac{2\pi j(a+b)}{p} + \frac{1}{2}\cos\frac{2\pi j(a-b)}{p}$, with the second term being interference. In general, for each non-empty $T \subseteq [n]$ the neuron carries the cosine of $2\pi j(\sum_{t \notin T} a_t - \sum_{t \in T} a_t - c)/p$, interference at the feature's own frequency. We measure it by $\beta_{j,T} := 2p^{-(n+1)} \hat{f}_\theta(k_1, \ldots, k_n; -j)$ with $k_t = -j$ for $t \in T$ and $k_t = j$ otherwise, so $\beta_{j,\emptyset} = \beta_j$ and $\beta_{j,T} = \sum_i \beta_{ij,T}$. These copies appear in the analytic solutions of Gromov (2023, Eqs. 10–11), which cancel them approximately through random phases with exact cancelling constructions seen in Morwani et al. (2024, App. F.2) and Li et al. (2025, App. E).

Definition 5 (interference-free network)**.** A network is *interference-free* if $\beta_{j,T} = 0$ for every $j \in F$ and every non-empty $T \subseteq [n]$.

The presence of interference is seen in the network's logit table: neurons may carry several frequencies and interference, which may cancel in the sum. Interference at other modes is unrestricted and costs the model per C_b. Lemma 1 removes this interference by averaging. We treat general n and price it (Lemma 7)- for one neuron the copies are rigidly tied to the feature.

Lemma 6. *For product and power neurons, for each neuron i and frequency j there are unit complex numbers $\xi_{i1j}, \ldots, \xi_{inj}$ with $\beta_{ij,T} = \beta_{ij} \prod_{t \in T} \xi_{itj}$ for every $T \subseteq [n]$. (Proof: Appendix C.)*

The lemma is Hermitian symmetry of the weights' Fourier transform (cf. Gromov, 2023, Eq. 11). Hence, a single neuron is never interference-free at a frequency it uses, and several neurons must cancel each other's copies. Relabel the neurons with $\beta_{ij} \ne 0$ as $l = 1, \ldots, N$, with unit factors ξ_{lt} and total $\beta = \beta_j$. The network is interference-free at j exactly when

$$\sum_{l=1}^{N} \beta_l \prod_{t \in T} \xi_{lt} \;=\; \beta\,[T = \emptyset] \qquad \text{for every } T \subseteq [n], \tag{3}$$

Lemma 7. *Let $\nu = n+1$, $|\xi_{lt}| = 1$, and let $\beta_1, \ldots, \beta_N$ satisfy equation 3. Then for every $r \in (0, 1]$, $\sum_l |\beta_l|^r \ge \nu^{1-r} |\beta|^r$. If $\beta = 0$, equality holds exactly when all $\beta_l = 0$. If $\beta \ne 0$ and $r < 1$, equality holds if and only if exactly ν of the β_l are non-zero, all of modulus $|\beta|/\nu$, with pairwise orthogonal chain vectors $m_l := (1, \xi_{l1}, \xi_{l1}\xi_{l2}, \ldots, \xi_{l1} \cdots \xi_{ln}) \in \mathbb{C}^\nu$.*

A non-zero clean feature requires at least ν contributing neurons; for $r < 1$, its least cost is attained by exactly ν equal-magnitude contributions, and more neurons never make it cheaper; the cancellation costs a factor ν^{1-r} over a single (interfering) neuron. At $r = 1$ the inequality is the triangle inequality, which is why the neuron count is pinned only for $b < \nu$.

Informal Proof (Appendix B.1). For $\beta \neq 0$, the matrix $H = \sum_l \beta_l m_l m_l^*$ has below-diagonal entries equal to twisted sums of equation 3 over intervals of operands, hence zero, and diagonal β; so $\det(I + tH/\beta) = (1+t)^\nu$. Expanding the determinant over subsets of neurons and bounding Gram determinants by one (Hadamard) gives $\prod_l (1 + x_l t) \geq (1+t)^\nu$ with $x_l = \nu|\beta_l|/|\beta|$, and integrating the logarithm against $t^{-1-r}dt$ gives $\sum_l x_l^r \geq \nu$. □

Given Lemma 15, the inequality also follows from the concave trace inequality of Rotfel'd (Thompson, 1976) with AM–GM on singular values; our route additionally characterizes equality. For $\nu = 3$ the equality configuration is the three-node cube-root-of-unity solution of Tian (2025).

Theorem 8. *Consider product networks with n operands ($\nu = n + 1$), let $2 \leq b \leq \nu$, $s = \nu/b$, and let y^* maximize $\Gamma_s(p)$. For width $m \geq \nu|\mathrm{spt}(y^*)|$, the maximum $L_{2,b}$-margin over interference-free networks on D_p is*

$$\gamma^*_{b,\mathrm{IF}} = \nu^{-(\nu-b)/b}\, B_p\, \Gamma_{\nu/b}(p).$$

For $b < \nu$, any interference-free network achieving it satisfies:

1. *its amplitudes are proportional to a maximizer $y^\star$ of equation 2, so it uses exactly $|\mathrm{spt}(y^\star)|$ frequencies;*
2. *every frequency it uses is carried by exactly ν neurons, with $|\beta_{ij}| = A_j/\nu$ and pairwise orthogonal chain vectors; no neuron carries an unused frequency.*

Informal Proof (Appendix B.2). By Lemmas 6 and 7, $\sum_i |\beta_{ij}|^r \geq \nu^{1-r} A_j^r$ at each j ($r = b/\nu$); summing and applying the single-neuron budget gives $\nu^{1-r} \sum_j A_j^r \leq C_b \|\theta\|_{2,b}^b$, and Proposition 3 gives the bound. Attainment is Proposition 9. At equality, Proposition 3 is tight (condition 1) and Lemma 7 is tight at each frequency (condition 2). □

The class value is the universal bound divided by $\nu^{(\nu-b)/b}$ ($\sqrt{3}$, 4, $5^{3/2}$ at $\nu = 3, 4, 5$ under L_2 weight decay; 1 at $b = \nu$): the price of splitting each feature over ν neurons.

Proposition 9. *For $j \in F$, $\lambda > 0$ and $\alpha_l = \pi l/\nu$ ($l = 0, \ldots, n$), the ν product neurons $u_t^{(l)}(a) = \lambda \cos(2\pi j a/p + \alpha_l)$, $w^{(l)}(c) = \lambda \cos(2\pi j c/p + n\alpha_l)$ have total logit table $\nu\lambda^\nu 2^{-n} \cos(2\pi j(a_1 + \cdots + a_n - c)/p)$ and $\|\omega^{(l)}\|_2^2 = \nu\lambda^2 p/2$. One block per $j \in \mathrm{spt}(y^*)$, with $\lambda_j^\nu = 2^n y_j^*/\nu$, attains Theorem 8.*

Proof: Appendix B.3.

The doubled phases $e^{2i\alpha_l}$ are the ν^{th} roots of unity, the equality case of Lemma 7: summed over the block, they cancel every sign-flipped copy. The block identity is the equally spaced case of the separated expansion of Mohlenkamp & Monzón (2005, Thm. 2); the same cancellation appears as a sufficient construction in Marchetti et al. (2026, Lemma B.8) and, for $\nu = 3$, as the three-node solution of Tian (2025) and a full-diversification configuration of He et al. (2026, Sec. 4). New here are the weights $\lambda_j^\nu \propto y_j^*$, attainment of the class value, and optimality (Lemma 7). The weights $\lambda_j^\nu \propto y_j^*$ matter, since optimal shares are not equal.

Logarithmic scaling. The programme also fixes how the optimum scales with the prime.

Theorem 10 (logarithmic scaling)**.** *Fix $s > 1$ and let $a_s := 3 + s/2$. For every odd prime p,*

$$\tfrac{1}{2}\lceil 8 \ln p \rceil^{1-s} \;\leq\; \Gamma_s(p) \;\leq\; \tfrac{3}{2}(2a_s)^s (\ln p)^{1-s}.$$

Consequently, for product networks with $2 \leq b < \nu$, write $\gamma^*_{b,m}(p)$ for the maximum $L_{2,b}$-margin at width m, with interference allowed. Uniformly over $m \geq \nu\lceil 8 \ln p \rceil$,

$$\gamma^*_{b,m}(p) = \Theta_{\nu,b}\big(p^{-\nu/2} (\ln p)^{1-\nu/b}\big).$$

(Proof: Appendix B.5.)

Theorem 11 (logarithmic number of features)**.** *Fix $s > 1$ and let $a_s := 3 + s/2$. For every odd prime p, every maximizer of equation 2 uses $\Theta(\log p)$ frequencies: between $\ln p/\big(8a_s^2(3 \cdot 9^{s-1})^{1/s}\big)$ and $9 \cdot 4^{1/(s-1)} \ln p$ of them, and it spends a share of at least $4^{-1/(s-1)}/(9 \ln p)$ of its budget on each. (Proof: Appendix B.5.)*

Table 1: $K^*(\nu/b, p)$ under L_2 weight decay ($b = 2$), best found, not certified: supports are enumerated exhaustively for $K \leq 8, 7, 5$ at $p = 59, 71, 113$ and by local search above, with the amplitudes optimized on the leading supports.

p	$\nu = 3$ ($s = 3/2$)	$\nu = 4$ ($s = 2$)	$\nu = 5$ ($s = 5/2$)
59	5	4	3
71	5	4	3
113	7	4	4

The network rate in Theorem 10 does not require the global margin maximizer to be interference-free (Appendix D). The universal upper bound and a block construction using at most $\nu\lceil 8 \ln p\rceil$ neurons give the same dependence on p, with constants depending only on ν, b. When the width is sufficient to realize a programme maximizer, the comparison becomes

$$\nu^{1-s} B_p \Gamma_s(p) \leq \gamma^*_{b,m}(p) \leq B_p \Gamma_s(p)$$

With equal shares, $\min_{\delta \neq 0} \sum_{j \in S} \kappa_j(\delta)$ is $|S|$ minus half the second eigenvalue of $\mathrm{Cay}(\mathbb{Z}_p, \pm S)$, so the unweighted analogue of Theorem 11 is the classical fact that abelian Cayley graphs need, and admit, $\Theta(\log p)$ generators for a constant spectral gap (Alon & Roichman, 1994, Sec. 3); for fixed K see Friedman et al. (2006), and for arbitrary weights Wang & Ye (2025). Empirically, trained networks' frequency counts grow logarithmically, and $O(\log p)$ random frequencies suffice for an unnormalized margin (McCracken et al., 2025). New here are the two-sided bound under the concave budget, the support bound with its share floor, and the network rate. Even with interference, a near-optimal network must use sufficiently many features:

Corollary 12. *Let $2 \leq b < \nu$, $s = \nu/b > 1$, $a_s = 3 + s/2$, and $K \geq 1$. Let $\theta \neq 0$ be any network (product or power, any width, interference allowed) whose amplitudes A_j are non-zero on at most K frequencies, and let $x := \ln p/K$. Then*

$$\frac{h(\theta)}{\|\theta\|^{\nu}_{2,b}} \;\leq\; C_b^{\nu/b} \cdot \tfrac{3}{2}\Big(\frac{x}{e^{x/(2a_s)} - 1}\Big)^s (\ln p)^{1-s}.$$

For fixed ν, b, product networks with $K = o(\log p)$ therefore attain a vanishing fraction of $\gamma^*_{b,m}(p)$ at any comparison width $m \geq \nu\lceil 8 \ln p\rceil$. (Proof: Appendix B.5.)

At the level of the programme this is classical: the qualitative equal-share analogue is that abelian Cayley graphs with $o(\log n)$ generators do not expand (Alon & Roichman, 1994, Prop. 3); fixed K with arbitrary weights gives simultaneous Dirichlet approximation (Schmidt, 1980), with the sharper decay $p^{-2/K}$; and for $K = o(\log p)$ under the $\ell_{1/s}$ budget a pigeonhole argument with frequency-dependent radii (Minkowski's linear-forms theorem; Cassels, 1959) gives the same conclusion. The contribution is the transfer to network margins through Lemma 2 and Theorem 4.

How many features. Under L_2 weight decay the count falls as the degree rises and grows only slowly with p (Table 1). The dense network of Corollary 14 reaches 0.49–0.60 of the interference-free optimum at $\nu = 3$, 0.15–0.23 at $\nu = 4$ and 0.04–0.08 at $\nu = 5$: the preference for few features sharpens with each step in degree. Whether a network carrying interference can beat the class value is an open conjecture (Appendix D), however, empirically trained networks do carry interference, quantitatively measured by Section 5.

How the count grows with p. By Theorem 11, for fixed $2 \leq b < \nu$ every programme maximizer uses $\Theta(\log p)$ frequencies, and by Theorem 10 the same bounds determine the order of the unrestricted maximum margin.

The theorem also shows how narrow the dense regime is. By Corollary 14 every frequency is used at $b = \nu$. By the share floor $4^{-1/(s-1)}/\lceil 8 \ln p\rceil$ derived in the proof of Theorem 11 (Appendix B.5),

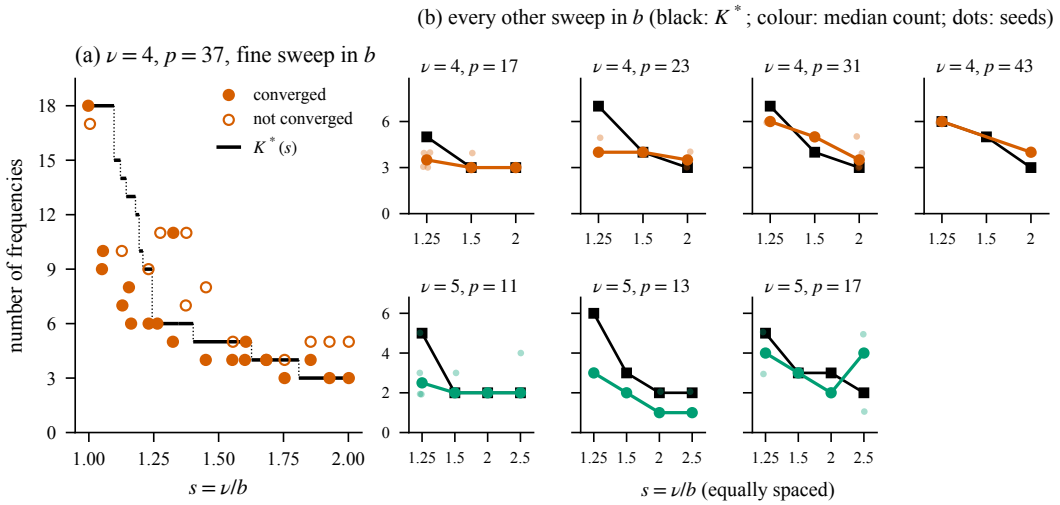


Figure 1: **The count follows the exponent.** (a) Product networks at $\nu = 4$, $p = 37$, sweeping b from 4 to 2 (s from 1 to 2). Black: $K^*(s)$, exact, with its twelve plateaus (the two at $K^* = 6$ differ in their amplitude ratios). Each marker is one run (19 converged, 13 not). (b) Every other sweep in b at fixed (ν, p) with at least two converged cells ($\nu = 4$, $p = 59$, where 8 of 12 runs diverged, has one and is omitted). Black: K^*; colour: median count over converged seeds; dots: individual seeds. The s values are drawn equally spaced.

however, for $(p-1)/2 > \lceil 8 \ln p \rceil$ (that is, $p \geq 73$), a programme maximizer can use all $\frac{p-1}{2}$ frequencies only if $s - 1 \leq \ln 4 / \ln\left(\frac{p-1}{2\lceil 8 \ln p \rceil}\right)$, which is $O(1/\log p)$; this envelope is loose. So the range of b below ν on which the programme optimum stays dense shrinks as p grows. Computed exactly, the optimum is dense up to $s = 1.548$, 1.345, 1.285, 1.217, 1.194 and 1.160 at $p = 7$, 11, 13, 17, 19 and 23.

5 Experiments

We train the networks of Section 2 with full-batch gradient descent with momentum and ask whether the number of features tracks the programme equation 2. Theorem 8 is a statement about interference-free maximizers. Yet trained networks are neither interference-free nor exact maximizers, hence, we predict trends in the number of frequencies observed and not exact counts. Training details, the convergence criterion and the count definition are in Appendix E.

5.1 Where interference enters trained networks

Trained networks carry some interference: most of their logit energy lies in the clean modes $(j, \ldots, j; -j)$ (median 0.97), and less of it as s grows. They are also not maximizers of the programme. Their learned allocations reach about 0.87 of the optimum $\Gamma_s(p)$, well above a random allocation or the network at initialization, but the individual budget shares do not match y^*. No trained network exceeded the universal bound of Theorem 4. The numbers behind this section are in Appendix E.

5.2 The exponent $s = \nu/b$

Sweeping the norm exponent at fixed degree and prime (Figure 1a), the trained count falls with K^*, from every frequency at $b = \nu$ to three at $b = 2$, although it does not resolve the finer plateaus of K^*. The same holds in every other sweep in b (Figure 1b and, at $\nu = 3$, Figure 2). Across all pairs of settings that share ν and p but differ in b, the trained count moves in the same direction as K^* in 30 of 36 (Appendix E).

At the dense end, $b = \nu$ ($s = 1$, the left end of Figure 1a), every converged network at $\nu = 3$ and $\nu = 4$ uses every frequency, most with near-equal amplitudes, as Corollary 14 predicts; density alone is not diagnostic, since the network at initialization is already dense, but flatness is. At $\nu = 5$ the picture is mixed.

5.3 Moving the degree and the norm separately

Theorem 8 says that the optimal spectral allocation depends on ν and b only through s. At equal s, trained networks of different degree use the same number of frequencies to within one in 12 of 17 comparisons: gradient descent falls further below K^*, and further from the optimum, as the degree grows (Appendix E).

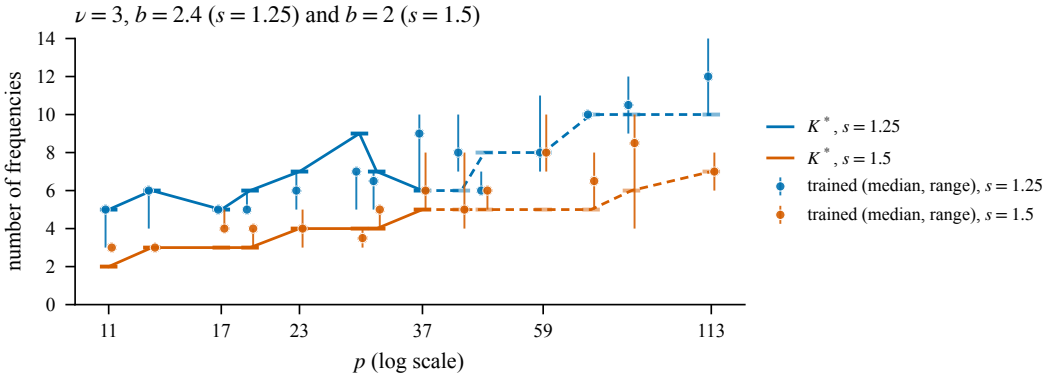


Figure 2: **The count across primes**, product networks at $\nu = 3$. Lines: K^* (solid where exact, $p \leq 37$; dashed where best found, Appendix E). Markers: median over converged seeds, with the range. K^* is not monotone in p (at $s = 1.25$ it is $9, 7, 6$ at $p = 29, 31, 37$); the trained counts do not follow these steps.

Across primes (Figure 2), the count stays within one of K^* in most (p, s) cells (20 of 28, against 16 for the best constant), but it does not follow the non-monotone steps of K^*.

6 Discussion

For modular addition, the number of features a maximum-margin network uses is set by the exponent $s = \nu/b$: every frequency at $b = \nu$, and $\Theta(\log p)$ frequencies for every $b < \nu$ (Theorems 8 and 11). The same programme recovers the dense maxima of Morwani et al. (2024) and Li et al. (2025) at $b = \nu$, and in trained networks the count moves with s in the direction the programme predicts (Section 5).

Limitations. The exact optimum of Theorem 8 is proved for interference-free networks. For all networks we prove only the universal bound and the order of the optimum (Theorems 4 and 10); whether interference can raise the margin for $2 \leq b < \nu$ is open (Appendix D). The results concern maximum-margin solutions, which regularized cross-entropy reaches only at its global minimizers as $\lambda \to 0$ (Wei et al., 2019); gradient descent need not find them. Accordingly, trained networks carry some interference and are not maximizers of the programme, so our experiments test trends in the count rather than exact values: they do not resolve the finer plateaus of K^* or its non-monotone steps in p, and they fall further below K^* as the degree grows. For power networks we prove the same upper bounds (with constant $n!B_p$), but they are known to be attained only at $b = \nu$ (Li et al., 2025), so the exact optimum and the logarithmic laws are proved for product networks only (Appendix D). Finally, the theory covers sums of such neurons over the cyclic group $\mathbb{Z}_p$; networks with a shared linear layer in front of the neurons are priced by a nuclear norm and fall outside it (Appendix C), and other groups are not treated here.

AI USE STATEMENT

In this work, we used generative AI tools to design, and give feedback on, the research methodology and experiments, implement methods, provide critical ingredients for proving mathematical claims, interpret results and assist in the writing of proofs. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims and propose or refine hypotheses, and generating synthetic data sets, assisting with translation, cleaning and reformatting datasets, and qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, create the figures, suggest experimental parameters, build the supplementary material, draft parts of the paper (Section 5 and Appendix E), edit the paper for readability, search for, identify, summarize and analyse existing literature, format references, brainstorm, identify gaps and suggest a structure for the paper. We have reviewed all AI-assisted work. All proofs that were AI-assisted were verified by the authors and Lean 4, a literature survey was undertaken to prevent plagiarism and LLM-generated code was hand-reviewed. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All theorems are stated with their hypotheses in Sections 3 and 4 and proved in full in Appendices A and B; Appendix C checks the assumptions for concrete networks. The data are the full table D_p of Section 2, with no split or preprocessing. Appendix E gives the training recipe, the convergence criterion and the definition of the feature count. The anonymized supplementary material contains the code, with scripts that regenerate the predictions, runs and analyses; the trained weights and full configuration of every run; a list of every launched run with its status; the preregistrations, with the post-hoc analyses reported separately; an audit of the training code against the stated recipe; and a file, `results.json`, that links every number in the paper, including the numerical checks of Section 4, to the script that computes it.

REFERENCES

Noga Alon and Yuval Roichman. Random Cayley graphs and expanders. *Random Structures & Algorithms*, 5(2):271–284, 1994. doi: 10.1002/rsa.3240050203.

Stephen Boyd. Convex optimization of graph Laplacian eigenvalues. In *Proceedings of the International Congress of Mathematicians*, volume 3, pp. 1311–1319, 2006. URL https://web.stanford.edu/~boyd/papers/cvx_opt_graph_lapl_eigs.html.

J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer, Berlin, Heidelberg, 1959. doi: 10.1007/978-3-642-62035-5.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6243–6267, 2023. URL https://proceedings.mlr.press/v202/chughtai23a.html. arXiv:2302.03025.

Sebastian M. Cioabă. Closed walks and eigenvalues of Abelian Cayley graphs. *Comptes Rendus Mathématique. Académie des Sciences. Paris*, 342:635–638, 2006. doi: 10.1016/j.crma.2006.03.005.

Zhen Dai, Mina Karzand, and Nathan Srebro. Representation costs of linear neural networks: Analysis and design. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26884–26896, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/e22cb9d6bbb4c290a94e4fff4d68a831-Abstract.html.

DLMF. NIST digital library of mathematical functions. https://dlmf.nist.gov/, Release 1.2.8 of 2026-09-15, 2026. URL https://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973. URL https://dml.cz/handle/10338.dmlcz/101168.

Joel Friedman, Ram Murty, and Jean-Pierre Tillich. Spectral estimates for Abelian Cayley graphs. *Journal of Combinatorial Theory, Series B*, 96(1):111–121, 2006. doi: 10.1016/j.jctb.2005.06.012.

Dongdong Ge, Xiaoye Jiang, and Yinyu Ye. A note on the complexity of L_p minimization. *Mathematical Programming*, 129(2):285–299, 2011. doi: 10.1007/s10107-011-0470-2.

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

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/0e98aeeb54acf612b9eb4e48a269814c-Abstract.html. arXiv:1806.00468.

Jianliang He, Leda Wang, Siyu Chen, and Zhuoran Yang. On the mechanism and dynamics of modular addition: Fourier features, lottery ticket, and grokking. *arXiv preprint arXiv:2602.16849*, 2026. URL https://arxiv.org/abs/2602.16849.

Warren M. Hirsch and Alan J. Hoffman. Extreme varieties, concave functions, and the fixed charge problem. *Communications on Pure and Applied Mathematics*, 14(3):355–369, 1961. doi: 10.1002/cpa.3160140313.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Peter D. Hoff. Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017. doi: 10.1016/j.csda.2017.06.007.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012. doi: 10.1017/CBO9781139020411.

Chon-Fai Kam, Miloud Bessafi, and Frédéric Cadet. Symmetry without a manifold: intrinsic dimension on orbits. *arXiv preprint arXiv:2609.17926*, 2026. URL https://arxiv.org/abs/2609.17926.

Chris Kolb, Tobias Weber, Bernd Bischl, and David Rügamer. Deep weight factorization: Sparse learning through the lens of artificial symmetries. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2502.02496. arXiv:2502.02496.

Chenyang Li, Yingyu Liang, Zhenmei Shi, Zhao Song, and Tianyi Zhou. Fourier circuits in neural networks and transformers: A case study of modular arithmetic with multiple inputs. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 523–531, 2025. URL https://proceedings.mlr.press/v258/li25b.html. arXiv:2402.09469.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

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. URL https://arxiv.org/abs/2311.18817. arXiv:2311.18817.

Giovanni Luca Marchetti, Daniel Kunin, Adele Myers, Francisco Acosta, and Nina Miolane. Sequential group composition: A window into the mechanics of deep learning. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2602.03655. arXiv:2602.03655.

Gavin McCracken, Gabriela Moisescu-Pareja, Vincent Létourneau, Doina Precup, and Jonathan Love. Uncovering a universal abstract algorithm for modular addition in neural networks. In *Advances in Neural Information Processing Systems*, volume 38, pp. 49512–49573, 2025. doi: 10.52202/085713-1476. URL https://openreview.net/forum?id=zuHs6RHQwT. arXiv:2505.18266.

Martin J. Mohlenkamp and Lucas Monzón. Trigonometric identities and sums of separable functions. *The Mathematical Intelligencer*, 27(2):65–69, 2005. doi: 10.1007/BF02985795.

Depen Morwani, Benjamin L. Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham Kakade. Feature emergence via margin maximization: case studies in algebraic tasks. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2311.07568. arXiv:2311.07568.

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.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1376–1401, 2015. URL https://proceedings.mlr.press/v40/Neyshabur15.html.

R. A. Rankin. The closest packing of spherical caps in n dimensions. *Proceedings of the Glasgow Mathematical Association*, 2(3):139–144, 1955. doi: 10.1017/S2040618500033219.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton Mathematical Series. Princeton University Press, 1970. doi: 10.1515/9781400873173.

René L. Schilling, Renming Song, and Zoran Vondraček. *Bernstein Functions: Theory and Applications*, volume 37 of *De Gruyter Studies in Mathematics*. De Gruyter, 2 edition, 2012. doi: 10.1515/9783110269338.

Wolfgang M. Schmidt. *Diophantine Approximation*, volume 785 of *Lecture Notes in Mathematics*. Springer, Berlin, Heidelberg, 1980. doi: 10.1007/978-3-540-38645-2.

David Slepian. Group codes for the Gaussian channel. *Bell System Technical Journal*, 47(4):575–602, 1968. doi: 10.1002/j.1538-7305.1968.tb02486.x.

Nathan Srebro, Jason Rennie, and Tommi S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, volume 17, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/hash/e0688d13958a19e087e123148555e4b4-Abstract.html.

Jun Sun, Stephen Boyd, Lin Xiao, and Persi Diaconis. The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM Review*, 48(4):681–699, 2006. doi: 10.1137/S0036144504443821.

Robert C. Thompson. Convex and concave functions of singular values of matrix sums. *Pacific Journal of Mathematics*, 66(1):285–290, 1976. doi: 10.2140/pjm.1976.66.285.

Yuandong Tian. Composing global solutions to reasoning tasks via algebraic objects in neural nets. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2410.01779. arXiv:2410.01779.

Ryan J. Tibshirani. Equivalences between sparse models and neural networks. Working notes, April 15, 2021, 2021. URL https://www.stat.berkeley.edu/~ryantibs/papers/sparsitynn.pdf.

Gal Vardi, Ohad Shamir, and Nathan Srebro. On margin maximization in linear and ReLU networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 37024–37036, 2022. doi: 10.52202/068431-2683. URL https://arxiv.org/abs/2110.02732.

Lixia Wang and Ke Ye. An estimate for positive definite functions on finite abelian groups and its applications. *arXiv preprint arXiv:2503.06908*, 2025. URL https://arxiv.org/abs/2503.06908.

Colin Wei, Jason D. Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8744cf92c88433f8cb04a02e6db69a0d-Abstract.html. arXiv:1810.05369.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.17844. arXiv:2306.17844.

APPENDIX

Appendix organization. Appendix A proves the results of Section 3, including the price of a feature (Lemma 13). Appendix B proves the results of Section 4: the cost of a feature (Lemma 7), Theorem 8, the ν-block (Proposition 9), two facts about the programme (Lemmas 18 and 19) and logarithmic scaling (Theorems 10 and 11, Corollary 12). Appendix C computes the single-neuron budget and checks the sign rule for the product and power networks. Appendix D states the conjecture that the interference-free value is the global maximum, and what is known about it.

Throughout, for $v : \mathbb{Z}_p \to \mathbb{R}$ we use $\hat{v}(j) = \sum_a v(a)\rho^{-ja}$, so that $v(a) = p^{-1}\sum_j \hat{v}(j)\rho^{ja}$, $\sum_{j\in\mathbb{Z}_p} |\hat{v}(j)|^2 = p\|v\|_2^2$ (Parseval), $\hat{v}(-j) = \overline{\hat{v}(j)}$, and hence $\sum_{j\in F} |\hat{v}(j)|^2 \le \frac{p}{2}\|v\|_2^2$.

A PROOFS FOR SECTION 3

A.1 THE SPECTRUM OF THE AVERAGED LOGIT

Write the logit table through its multidimensional Fourier transform, $f(\theta, x)[c] = p^{-(n+1)} \sum_{k_1,\dots,k_n,l} \hat{f}_\theta(k_1, \dots, k_n; l)\, \rho^{k_1 a_1 + \cdots + k_n a_n + lc}$. Substituting $c = y(x) - \delta = a_1 + \cdots + a_n - \delta$ and averaging over x,

$$\eta_\theta(\delta) = p^{-(n+1)} \sum_{k,l} \hat{f}_\theta(k; l)\, \rho^{-l\delta} \prod_{t=1}^{n} \mathbb{E}_{a_t}\big[\rho^{(k_t+l)a_t}\big].$$

Each average is 1 if $k_t = -l$ and 0 otherwise, so only the modes $(j, \dots, j; -j)$ survive (with $j = -l$):

$$\eta_\theta(\delta) = p^{-(n+1)} \sum_{j\in\mathbb{Z}_p} G_j\, \rho^{j\delta}, \qquad G_j := \hat{f}_\theta(j, \dots, j; -j).$$

Since the table is real, $G_{-j} = \overline{G_j}$, and pairing j with $-j$ gives equation 1 with $\bar{\eta}_\theta = p^{-(n+1)}G_0$ and $\beta_j = 2p^{-(n+1)}G_j$. The averaged form used in Lemma 2 is

$$\eta_\theta(0) - \tfrac{1}{2}\big(\eta_\theta(\delta) + \eta_\theta(-\delta)\big) = \sum_{j\in F} \mathrm{Re}(\beta_j)\big(1 - \cos(2\pi j\delta/p)\big) = \sum_{j\in F} \mathrm{Re}(\beta_j)\, \kappa_j(\delta),$$

because $\mathrm{Re}(\beta\rho^{j\delta}) + \mathrm{Re}(\beta\rho^{-j\delta}) = 2\,\mathrm{Re}(\beta)\cos(2\pi j\delta/p)$. The same computation for one neuron defines β_{ij}. For a product neuron $\hat{f}$ factorizes, so $G_{ij} = \hat{u}_{i1}(j) \cdots \hat{u}_{in}(j)\, \hat{w}_i(-j)$, and $p^{-(n+1)} = p^{-\nu}$.

A.2 PROOF OF LEMMA 2

Fix $\delta \neq 0$. Then $-\delta \neq 0$ as well, so Lemma 1 applies at both offsets:

$$h(\theta) \le \eta_\theta(0) - \eta_\theta(\delta), \qquad h(\theta) \le \eta_\theta(0) - \eta_\theta(-\delta).$$

Averaging the two and using the identity of Appendix A.1,

$$h(\theta) \le \eta_\theta(0) - \tfrac{1}{2}\big(\eta_\theta(\delta) + \eta_\theta(-\delta)\big) = \sum_{j\in F} \mathrm{Re}(\beta_j)\, \kappa_j(\delta).$$

This is the first inequality. For the second, $\kappa_j(\delta) = 1 - \cos(2\pi j\delta/p) \ge 0$ and $\mathrm{Re}(\beta_j) \le |\beta_j| = A_j$, so each term is at most $A_j\kappa_j(\delta)$. Equality in the second step requires $\mathrm{Re}(\beta_j) = A_j$, i.e. β_j real and non-negative, at every j with $\kappa_j(\delta) > 0$. □

A.3 THE PRICE OF A FEATURE

This makes precise the price used in Section 3.2; Appendix C proves the general single-neuron budget, of which this is the single-frequency case.

Lemma 13 (price of an amplitude)**.** *Let $\omega = (u_1, \dots, u_n, w)$ be a product neuron whose $\nu = n+1$ weight vectors are real and supported on the frequencies $\pm j$ for one $j \in F$, and let $A = |\beta_j(\omega)|$. Then for every $b > 0$,*

$$\|\omega\|_2^b \;\geq\; \big(A/B_p\big)^{b/\nu}, \qquad B_p = 2(2p\nu)^{-\nu/2},$$

with equality if and only if the ν weight vectors have equal norm. So the least $L_{2,b}$ penalty that buys amplitude A is $B_p^{-1/s} A^{1/s}$ with $s = \nu/b$.

Proof. Write $v_1, \dots, v_\nu$ for $u_1, \dots, u_n, w$ and $x_t = \|v_t\|_2$. A real vector supported on $\pm j$ has $|\hat{v}(-j)| = |\hat{v}(j)|$ and no other Fourier content, so Parseval gives $x_t^2 = \frac{2}{p}|\hat{v}_t(j)|^2$. By Appendix A.1, $A = 2p^{-\nu}\prod_t |\hat{v}_t(j)| = 2p^{-\nu}(p/2)^{\nu/2}\prod_t x_t$. By AM–GM, $\|\omega\|_2^2 = \sum_t x_t^2 \geq \nu\big(\prod_t x_t\big)^{2/\nu}$, with equality if and only if all x_t are equal. Hence

$$\|\omega\|_2^2 \;\geq\; \nu\Big(\frac{A}{2p^{-\nu}(p/2)^{\nu/2}}\Big)^{2/\nu} = \Big(\frac{A}{2p^{-\nu}(p/(2\nu))^{\nu/2}}\Big)^{2/\nu} = \big(A/B_p\big)^{2/\nu},$$

and raising to the power $b/2$ gives the claim, since $t \mapsto t^{b/2}$ is increasing. □

For a power neuron with the same weight vectors, only the multilinear term of $\big(\sum_t u_t(a_t)\big)^n$ reaches the mode $(j, \dots, j; -j)$, so $A = n! \cdot 2p^{-\nu}\prod_t |\hat{v}_t(j)|$ (Appendix C); the same proof gives $\|\omega\|_2^b \geq \big(A/(n!\,B_p)\big)^{b/\nu}$.

The inner norm of $L_{2,b}$ must be ℓ_2 for this argument: Parseval converts it to Fourier magnitudes, and AM–GM then acts on the squared factor norms. The outer exponent b only raises each neuron's cost to the power $b/2$, which turns the product of ν factors into the exponent $b/\nu = 1/s$.

A.4 The dense endpoint $b = \nu$

Corollary 14 ($b = \nu$: every frequency is used)**.** *$\Gamma_1(p) = p/(p-1)$, with unique maximizer $y_j = 2/(p-1)$ for all j, so at $b = \nu$ any network attaining Theorem 4 has equal amplitude on every frequency. For the quadratic network of Morwani et al. (2024) the bound is $2B_p\frac{p}{p-1} = \sqrt{2/27}\,/\,(p^{1/2}(p-1))$, exactly their maximum $L_{2,3}$-margin (Morwani et al., 2024, Thm. 7; width $m \geq 4(p-1)$).*

For the n-operand power network, $n!\,B_p\,p/(p-1)$ is the maximum $L_{2,n+1}$-margin of Li et al. (2025, Thm. 4.1), and equal amplitude on every frequency is already implicit in Morwani et al. (2024, App. F.3). The value $\Gamma_1(p) = p/(p-1)$ and its flat maximizer are the regular-simplex bound for cyclic codes (Rankin, 1955; Slepian, 1968).

Proof of Corollary 14.

For feasible y (at $s = 1$, $\sum_j y_j \leq 1$),

$$\min_{\delta \neq 0} \sum_j y_j \kappa_j(\delta) \leq \frac{1}{p-1}\sum_{\delta \neq 0}\sum_j y_j \kappa_j(\delta) = \frac{p}{p-1}\sum_j y_j \leq \frac{p}{p-1},$$

using $\sum_{\delta \neq 0}\kappa_j(\delta) = p$. The flat choice $y_j = 2/(p-1)$ attains this, as we now show together with uniqueness (the averaging step is the regular-simplex bound of Rankin, 1955; cf. Fiedler, 1973, §3.5–3.6; the inversion step parallels Morwani et al., 2024, App. F.3 and Li et al., 2025, Lemma F.2). Equality forces $\sum_j y_j = 1$ and the coverage $\sum_j y_j \kappa_j(\delta)$ to be constant in $\delta \neq 0$, i.e. $\zeta(\delta) := \sum_{j \in F} y_j \cos(2\pi j\delta/p)$ equals a constant c_0 for $\delta \neq 0$, with $\zeta(0) = 1$. So $\zeta = c_0 + (1-c_0)[\delta = 0]$. Its normalized Fourier coefficients $\frac{1}{p}\sum_\delta \zeta(\delta)\rho^{-k\delta}$ are $c_0 + (1-c_0)/p$ at $k = 0$ and $(1-c_0)/p$ at $k \neq 0$. From the cosine form they are 0 at $k = 0$ and $y_j/2$ at $k = \pm j$. Hence $c_0 = -1/(p-1)$ and $y_j = 2(1-c_0)/p = 2/(p-1)$ for every j.

For the product network, attainment is Proposition 9 at $b = \nu$ with y^* flat. The equality conditions are those of Proposition 3 (flat amplitudes) and of the triangle inequality (at each j, all non-zero β_{ij} have the same phase). For the quadratic network of Morwani et al. (2024), $C_3 = 2B_p$ (Appendix C), and $2B_p \cdot \frac{p}{p-1} = 4(6p)^{-3/2}\frac{p}{p-1} = \frac{2}{3\sqrt{6}} \cdot \frac{1}{p^{1/2}(p-1)} = \sqrt{\frac{2}{27}} \cdot \frac{1}{p^{1/2}(p-1)}$. □

B Proofs for Section 4

B.1 Proof of Lemma 7

Lemma 15 (chain lemma)**.** *$\|m_l\|_2^2 = \nu$, and $H := \sum_l \beta_l m_l m_l^*$ is upper triangular with every diagonal entry equal to β.*

Proof. Index the chain coordinates by $0, \dots, n$. Every entry of m_l has modulus 1, so $\|m_l\|_2^2 = \nu$ and the diagonal of H is $\sum_l \beta_l = \beta$. For $t > t'$, the (t, t') entry of $m_l m_l^*$ is $\prod_{u=t'+1}^{t} \xi_{lu}$. Its sum against β_l vanishes by equation 3, applied to the non-empty interval $T = \{t'+1, \dots, t\}$. □

The entries above the diagonal need not vanish: H is triangular, not necessarily Hermitian. Triangularity is all the proof needs.

Proof of Lemma 7. If $\beta = 0$, the inequality is immediate, with equality exactly when every $\beta_l = 0$. Otherwise divide the coefficients by β; equation 3 is preserved with total 1. We prove the normalized claim and restore the factor $|\beta|^r$ at the end. For $r = 1$, the claim is simply $\sum_l |\beta_l| \geq |\sum_l \beta_l| = 1$. Henceforth let $0 < r < 1$.

Step 1: compare determinants. Put $u_l = m_l/\sqrt{\nu}$, $d_l = \nu\beta_l$ and $x_l = |d_l|$, so that the u_l are unit vectors and $H = \sum_l d_l u_l u_l^*$. Let $U = [u_1 \cdots u_N]$ and $\Delta = \operatorname{diag}(d_1, \dots, d_N)$. Sylvester's determinant theorem $\det(I + AB) = \det(I + BA)$ (Horn & Johnson, 2012, Ch. 1) and expansion by principal minors give

$$\det(I_\nu + tH) = \det(I_N + t\Delta U^* U) = \sum_{L \subseteq [N]} t^{|L|} \Big(\prod_{l \in L} d_l \Big) G_L, \qquad G_L := \det \operatorname{Gram}(u_l : l \in L). \tag{4}$$

Gram positivity and Hadamard's inequality (Horn & Johnson, 2012, Ch. 7) give $0 \leq G_L \leq 1$, with $G_L = 1$ exactly when the vectors in L are orthonormal. By Lemma 15, $\det(I + tH) = (1+t)^\nu$. Taking absolute values in equation 4 therefore yields, for $t \geq 0$,

$$(1+t)^\nu \leq \sum_{L \subseteq [N]} t^{|L|} \prod_{l \in L} x_l = \prod_{l=1}^{N} (1 + t x_l). \tag{5}$$

The same comparison coefficient by coefficient gives $e_k(x) \geq \binom{\nu}{k}$ for $0 \leq k \leq \nu$, where e_k is the kth elementary symmetric polynomial; in particular, at least ν coefficients are non-zero.

Step 2: integrate the logarithm. Define

$$J_r(x) := \int_0^\infty \log(1 + xt)\, t^{-1-r}\, dt, \qquad x \geq 0.$$

For $x > 0$ the integrand is $O(t^{-r})$ near zero and $O(t^{-1-r} \log t)$ near infinity, so the integral converges. Substitution gives $J_r(x) = x^r J_r(1)$, also valid at $x = 0$, with $0 < J_r(1) < \infty$. Integrating the logarithm of equation 5 against $t^{-1-r} dt$ and cancelling $J_r(1)$ gives

$$J_r(1) \sum_l x_l^r = \sum_l J_r(x_l) \geq \nu J_r(1), \qquad \text{hence} \qquad \sum_l x_l^r \geq \nu.$$

Restoring the original coefficients, $x_l = \nu|\beta_l|/|\beta|$, proves $\sum_l |\beta_l|^r \geq \nu^{1-r} |\beta|^r$.

Step 3: characterize equality. At equality, the continuous non-negative function $\sum_l \log(1 + x_l t) - \nu \log(1+t)$ has zero integral against a strictly positive density, so it vanishes for every $t > 0$. Thus $\prod_l (1 + x_l t) = (1+t)^\nu$ as polynomials. Their roots show that the non-zero x_l are exactly ν copies of 1. Let L_0 be their support. Comparing the coefficient of t^ν in equation 4 gives $1 = (\prod_{l \in L_0} d_l) G_{L_0}$. Since $|\prod_{l \in L_0} d_l| = 1$, we have $G_{L_0} = 1$: the corresponding u_l are orthonormal, and the chain vectors m_l are orthogonal. Conversely, exactly ν non-zero coefficients of modulus $|\beta|/\nu$ satisfying equation 3 give $\sum_l |\beta_l|^r = \nu^{1-r} |\beta|^r$. □

Remark 16 (orthogonality as roots of unity). Two chain vectors $m_l, m_{l'}$ are orthogonal if and only if $\sum_{t=0}^{n} \prod_{u \le t} \xi_{l'u}\overline{\xi_{lu}} = 0$. If the ratios $\xi_{l'u}\overline{\xi_{lu}} = \zeta$ agree across operands, this reads $1+\zeta+\cdots+\zeta^n = 0$: ζ is a non-trivial ν^{th} root of unity. This is the configuration of Proposition 9.

Remark 17 (other penalties). For $0 < r < 1$, $x \mapsto x^r$ is a positive integral of the functions $x \mapsto \log(1+xt)$. The same proof, together with $\sum_l x_l \ge \nu$, gives $\sum_l \psi(\nu|\beta_l|/|\beta|) \ge \nu\,\psi(1)$ for every $\psi(x) = cx + \int_0^\infty \log(1+xt)\,\mu(dt)$ with $c \ge 0$ and $\mu \ge 0$, whenever these integrals are finite and $\beta \ne 0$. The result is not special to power penalties (the admissible ψ are Thorin–Bernstein functions, Schilling et al., 2012, Ch. 8; for the inequality alone see also Thompson, 1976).

B.2 Proof of Theorem 8

Let $r = b/\nu$, so $1/r = s$.

Upper bound. Let θ be interference-free. At each $j \in F$, the coefficients $(\beta_{ij})_i$ satisfy equation 3 with total β_j: for $T = \emptyset$ this is $\beta_j = \sum_i \beta_{ij}$, and for $T \ne \emptyset$ it is $0 = \beta_{j,T} = \sum_i \beta_{ij} \prod_{t\in T} \xi_{itj}$ by Lemma 6. Lemma 7 gives $\sum_i |\beta_{ij}|^r \ge \nu^{1-r} A_j^r$. Summing over j and using the single-neuron budget,

$$\nu^{1-r} \sum_j A_j^r \;\le\; \sum_i \sum_j |\beta_{ij}|^r \;\le\; C_b \sum_i \|\omega_i\|_2^b = C_b \|\theta\|_{2,b}^b. \tag{6}$$

Raising to the power $1/r$, $\|A\|_{1/s} \le \nu^{(r-1)/r} C_b^{1/r} \|\theta\|_{2,b}^{b/r} = \nu^{-(\nu-b)/b} B_p \|\theta\|_{2,b}^{\nu}$, and Proposition 3 gives $h(\theta) \le \nu^{-(\nu-b)/b} B_p \Gamma_s(p) \|\theta\|_{2,b}^{\nu}$.

Attainment. Proposition 9 gives an interference-free network with $\nu|\mathrm{spt}(y^*)|$ neurons attaining the value; any further neurons are set to zero.

Structure ($b < \nu$). Let θ be interference-free and attain the value. The value is positive, so $h(\theta) > 0$ and $A \ne 0$. The two inequalities $h(\theta) \le \Gamma_s(p)\|A\|_{1/s}$ and $\|A\|_{1/s} \le \nu^{-(\nu-b)/b} B_p \|\theta\|_{2,b}^{\nu}$ must both be equalities. For the first, write $y = A/\|A\|_{1/s}$; by Lemma 2, $h(\theta) \le \|A\|_{1/s} \min_{\delta\ne 0} \sum_j y_j \kappa_j(\delta) \le \|A\|_{1/s}\Gamma_s(p)$, so equality forces $\min_\delta \sum_j y_j\kappa_j(\delta) = \Gamma_s(p)$: y is a maximizer of equation 2, which is condition 1 with $c = \|A\|_{1/s}$. For the second, the first inequality of equation 6 is an equality. It is a sum over j of the inequalities $\sum_i |\beta_{ij}|^r \ge \nu^{1-r} A_j^r$, so each of these is an equality. At a frequency with $A_j > 0$, the equality case of Lemma 7 ($r < 1$) is condition 2. At a frequency with $A_j = 0$ it forces $\beta_{ij} = 0$ for every i: no neuron carries an unused frequency. □

B.3 Proof of Proposition 9

Identity. Write $\vartheta_t = 2\pi j a_t/p$ and $\vartheta_c = 2\pi j c/p$, and expand each cosine into exponentials. Neuron l contributes

$$2^{-\nu}\lambda^\nu \sum_{\varepsilon\in\{\pm1\}^n,\ \varepsilon_c=\pm1} \exp\big(i(\varepsilon\cdot\vartheta + \varepsilon_c\vartheta_c)\big)\, \exp\Big(i\alpha_l\big(\textstyle\sum_t \varepsilon_t + n\varepsilon_c\big)\Big).$$

With $N_- = |\{t : \varepsilon_t = -1\}|$ we have $\sum_t \varepsilon_t = n - 2N_-$, so the l-dependent phase is $\exp(2\pi i\, lu/\nu)$ with $u = n - N_-$ if $\varepsilon_c = +1$ and $u = -N_-$ if $\varepsilon_c = -1$. Summing over $l = 0, \dots, n$ gives $\nu\,[u \equiv 0 \bmod \nu]$. As $0 \le N_- \le n$, this happens only for $(N_- = n, \varepsilon_c = +1)$ and $(N_- = 0, \varepsilon_c = -1)$, i.e. for $e^{\mp i(\sum_t \vartheta_t - \vartheta_c)}$, whose sum is $2\cos(\sum_t \vartheta_t - \vartheta_c)$. The block's table is therefore $\nu\lambda^\nu 2^{-n} \cos\big(2\pi j(a_1 + \cdots + a_n - c)/p\big)$. For the cost, $\sum_a \cos^2(2\pi j a/p + \alpha) = p/2$ for $j \not\equiv 0$, so each of the ν weight vectors of neuron l has squared norm $\lambda^2 p/2$.

Attainment. Let $r = b/\nu$ and $s = 1/r$. Scaling y^* up increases the objective, so $\sum_j (y_j^*)^r = 1$. With $\nu\lambda_j^\nu 2^{-n} = y_j^*$ the network's table is exactly $f(\theta, x)[c] = \sum_j y_j^* \cos\big(2\pi j(y(x) - c)/p\big)$. It depends only on $y(x) - c$, so it is interference-free, and on every input $f(\theta,x)[y(x)] - f(\theta,x)[y(x) - \delta] = \sum_j y_j^* \kappa_j(\delta)$; hence $h(\theta) = \min_{\delta\ne0} \sum_j y_j^* \kappa_j(\delta) = \Gamma_s(p)$. For any non-negative allocation y, the same construction has penalty

$$P_b := \|\theta\|_{2,b}^b = \sum_j \nu(\nu p/2)^{b/2} (2^n y_j/\nu)^r = \nu^{1-r} B_p^{-r} \sum_j y_j^r.$$

For $y = y^*$ the last sum is one. Thus $\|\theta\|_{2,b}^{\nu} = P_b^{1/r} = \nu^{s-1} B_p^{-1}$, and

$$\frac{h(\theta)}{\|\theta\|_{2,b}^{\nu}} = \nu^{1-s} B_p \Gamma_s(p),$$

as claimed. More generally, any y with $\sum_j y_j^r = 1$ gives normalized margin $\nu^{1-s} B_p \min_{\delta \neq 0} \sum_j y_j \kappa_j(\delta)$. □

B.4 Two facts about the programme

Throughout, $s > 1$, $r = 1/s < 1$, and $\phi(v) := \sum_j v_j^r$. Rescaling a feasible y so that its worst coverage is one puts it in the polyhedron

$$\mathcal{P}_p := \Big\{ v \in \mathbb{R}^F_{\geq 0} : \ \sum_{j \in F} v_j \, \kappa_j(\delta) \geq 1 \ \text{ for all } \delta \neq 0 \Big\}.$$

If y is feasible for equation 2 with worst coverage $m > 0$, then $v = y/m \in \mathcal{P}_p$ and $\phi(v) = m^{-r} \sum_j y_j^r \leq m^{-r}$. Conversely, for $v \in \mathcal{P}_p$ the vector $y = v/\phi(v)^s$ is feasible with worst coverage at least $\phi(v)^{-s}$. Hence $\Gamma_s(p) = (\min_{\mathcal{P}_p} \phi)^{-s}$. A maximizer has $\sum_j y_j^r = 1$, since scaling up raises the objective. So y is a maximizer if and only if $y = \Gamma_s(p)\, v$ with v a minimizer of ϕ over $\mathcal{P}_p$.

Lemma 18 (maximizers are vertices)**.** *Let $s > 1$. Every maximizer of equation 2 is $y = \Gamma_s(p)\, v$ for a vertex v of $\mathcal{P}_p$ that minimizes $\sum_j v_j^{1/s}$.*

Proof. This is the classical fact that concave functions attain their minima over a polyhedron at vertices (Hirsch & Hoffman, 1961; Rockafellar, 1970, §32). The polyhedron $\mathcal{P}_p$ is nonempty and closed. Since $v_j \leq \phi(v)^{1/r}$ for every j, the sublevel sets of ϕ in $\mathcal{P}_p$ are compact, so a minimum exists. If a minimizer v were not a vertex, it would be the midpoint of distinct $v', v'' \in \mathcal{P}_p$. Strict concavity of $t \mapsto t^r$ on $[0, \infty)$ would give $\phi(v) > \frac{1}{2}(\phi(v') + \phi(v'')) \geq \min \phi$, a contradiction. □

The vertex set is finite, so K^* can be computed exactly by enumerating the supports of vertices (Table 1); in general this minimization is NP-hard (Ge et al., 2011).

Lemma 19 (removing a feature)**.** *Let $s > 1$. Let y be feasible for equation 2 with $\sum_j y_j^{1/s} = 1$ and worst coverage $m \geq (1 - \varepsilon)\, \Gamma_s(p)$, where $\varepsilon \in [0, 1)$. For $j \in \operatorname{spt}(y)$, let $w_j := y_j^{1/s}$ be the share of the budget spent on j, and let m_{-j} be the worst coverage after setting $y_j = 0$. Then*

$$m - 2y_j \ \leq \ m_{-j} \ \leq \ \frac{(1 - w_j)^s}{1 - \varepsilon}\, m.$$

In particular, removing feature j from a maximizer costs at least the fraction $1 - (1 - w_j)^s \in [w_j, s\, w_j]$ of the worst coverage.

Proof. Removing frequency j decreases every offset's coverage by at most $2y_j$, proving the lower bound. For the upper bound, if $w_j = 1$ then $m_{-j} = 0$. Otherwise, set $y_j = 0$ and multiply the remaining amplitudes by $(1 - w_j)^{-s}$. The resulting allocation has budget one, so its worst coverage obeys $(1 - w_j)^{-s} m_{-j} \leq \Gamma_s(p) \leq m/(1 - \varepsilon)$. Finally, $(1 - w)^s \leq 1 - w$ and Bernoulli's inequality $(1 - w)^s \geq 1 - sw$ give $w \leq 1 - (1 - w)^s \leq sw$. □

In the network of Proposition 9, deleting the ν neurons of frequency j removes exactly $y_j \cos(2\pi j(a_1 + \cdots + a_n - c)/p)$ from the table, so these bounds apply to its margin directly. Lemma 19 is used in the proof of Theorem 11.

B.5 Proof of Theorems 10 and 11 and Corollary 12

Write $F_y(\delta) := \sum_{j \in F} y_j \kappa_j(\delta)$ and $m(y) := \min_{\delta \neq 0} F_y(\delta)$. Let $r = 1/s$ and $a_s = 3 + s/2$; logarithms are natural. The constants below are chosen for simplicity, not sharpness.

Lower bound. Let $L = \lceil 8 \ln p \rceil$ and draw $J_1, \dots, J_L$ independently and uniformly from F. For fixed $\delta \neq 0$,

$$\mathbb{E} \cos(2\pi J_1 \delta / p) = \frac{2}{p-1} \sum_{j \in F} \cos(2\pi j \delta / p) = -\frac{1}{p-1}.$$

By Hoeffding's inequality for variables in $[-1, 1]$ (Hoeffding, 1963, Thm. 2), $\Pr\left[\frac{1}{L}\sum_\ell \cos(2\pi J_\ell \delta/p) > \frac{1}{2}\right] \le e^{-L/8} \le \frac{1}{p}$. A union bound over the $p-1$ offsets leaves positive probability that $\sum_\ell \kappa_{J_\ell}(\delta) \ge L/2$ for every $\delta \neq 0$; fix such a draw. Let n_j be the multiplicity of j in the draw and set $y_j = n_j L^{-s}$. Then

$$q := \sum_j y_j^r = L^{-1} \sum_j n_j^r \le 1, \qquad F_y(\delta) = L^{-s} \sum_\ell \kappa_{J_\ell}(\delta) \ge \tfrac{1}{2} L^{1-s}.$$

Here $q > 0$. Replacing y by y/q^s makes the budget exactly one, preserves its support, and can only increase coverage. Thus the lower bound is achieved by a feasible allocation using at most L frequencies. Since $L \le 9 \ln p$ for $p \ge 3$, also $\Gamma_s(p) \ge \frac{1}{2}(9 \ln p)^{1-s}$. This is the cyclic case of the random Cayley graph argument of Alon & Roichman (1994, Sec. 3, Thm. 2); cf. McCracken et al. (2025, Thm. 4.7).

A positive-Fourier inequality. If every non-zero offset had large coverage, the average of $e^{-tF_y(\delta)}$ would be small. Positivity of Fourier coefficients forces the same average to remain large under a bounded amplitude budget; comparing the two estimates will bound the margin.

Let $I_0(z) = \frac{1}{\pi}\int_0^\pi e^{z\cos\vartheta} d\vartheta$ and $\psi(z) = z - \ln I_0(z)$. For $z \ge 0$, the identity

$$e^{z\cos\vartheta} = \exp\left((z/2)e^{i\vartheta}\right) \exp\left((z/2)e^{-i\vartheta}\right)$$

and the absolutely convergent exponential series show that all Fourier coefficients are non-negative, with constant coefficient $I_0(z)$ (DLMF, Eqs. 10.32.1, 10.35.2). Expand each factor of $e^{-tF_y(\delta)} = \prod_j e^{-ty_j} e^{ty_j \cos(2\pi j\delta/p)}$ and average over $\delta \in \mathbb{Z}_p$. The average keeps the products whose frequencies sum to $0 \bmod p$. All are non-negative, and they include the product of the constant terms. Hence, for $t \ge 0$,

$$\frac{1}{p}\sum_{\delta \in \mathbb{Z}_p} e^{-tF_y(\delta)} \;\ge\; \prod_j e^{-ty_j} I_0(ty_j) = \exp\Big(-\sum_j \psi(ty_j)\Big). \tag{7}$$

No independence of the phases $j\delta$ is used: resonances only add non-negative terms (cf. the closed-walk comparison of Cioabă, 2006).

A scalar bound. We claim $0 \le \psi(z) \le a_s \ln(1 + z^r)$ for $z \ge 0$.

- Non-negativity: $I_0(z) \le e^z$.
- For $z \le 1$: $I_0 \ge 1$ gives $\psi(z) \le z$. Also $\ln(1+z^r) \ge z^r/2 \ge z/2$ and $a_s \ge 2$.
- For $z \ge 1$: integrating only over $\vartheta \le z^{-1/2}$ and using $\cos\vartheta \ge 1 - \vartheta^2/2$ gives $I_0(z) \ge e^{z-1/2}/(\pi\sqrt{z})$, so $\psi(z) \le \ln\pi + \frac{1}{2} + \frac{1}{2}\ln z$. Meanwhile $3\ln(1+z^r) \ge 3\ln 2 > \ln\pi + \frac{1}{2}$, and $\frac{s}{2}\ln(1+z^r) \ge \frac{1}{2}\ln z$.

If $K \ge 1$ and y has at most K non-zero entries and $\sum_j y_j^r \le 1$, concavity of the logarithm then gives

$$\sum_j \psi(ty_j) \;\le\; a_s \sum_j \ln(1 + t^r y_j^r) \;\le\; a_s K \ln(1 + t^r/K) \;\le\; a_s t^r. \tag{8}$$

Upper bounds. Since $F_y(0) = 0$ and $F_y(\delta) \ge m(y)$ otherwise,

$$\frac{1}{p}\sum_\delta e^{-tF_y(\delta)} \le \frac{1}{p} + e^{-tm(y)}. \tag{9}$$

- Take $t = (\ln p/(2a_s))^s$. By equation 7–equation 9, $e^{-tm} \ge p^{-1/2} - p^{-1} \ge p^{-1/2}/3$, so $m(y) \le (\frac{1}{2}\ln p + \ln 3)/t \le \frac{3}{2}(2a_s)^s (\ln p)^{1-s}$ for every feasible y.

- Keeping K in equation 8 and taking $t^r = K(p^{1/(2a_sK)} - 1)$ instead gives, for every feasible y with at most K non-zero entries,

$$m(y) \;\le\; \tfrac{3}{2}(\ln p)^{1-s}\Big(\frac{x}{e^{x/(2a_s)}-1}\Big)^s, \qquad x = \ln p/K. \tag{10}$$

Number of frequencies (Theorem 11).

- Lower bound. A maximizer has $m = \Gamma_s(p) \ge \frac{1}{2}\,9^{1-s}(\ln p)^{1-s}$. With equation 10 and $e^z - 1 \ge z^2/2$ this gives $\frac{1}{2}\,9^{1-s} \le \frac{3}{2}(8a_s^2/x)^s$, i.e. $K = \ln p/x \ge \ln p/\big(8a_s^2(3\cdot 9^{s-1})^{1/s}\big)$.
- Upper bound. Lemma 19 with $\varepsilon = 0$ gives $m - 2y_j \le (1-w_j)^s m$. So $2y_j \ge m(1-(1-w_j)^s) \ge m\,w_j = m\,y_j^r$, i.e. $w_j \ge (m/2)^{1/(s-1)}$. As $\sum_j w_j = 1$, $K \le (2/m)^{1/(s-1)} \le 4^{1/(s-1)}L \le 9{\cdot}4^{1/(s-1)}\ln p$, and every share is at least $4^{-1/(s-1)}/L \ge 4^{-1/(s-1)}/(9\ln p)$. Keeping $L = \lceil 8\ln p\rceil$ gives the dense-regime bound in the main text whenever $|F| > L$.

Networks. Write $\gamma^*_{b,m}(p)$ for the maximum margin at width m. The universal bound gives $\gamma^*_{b,m}(p) \le B_p\Gamma_s(p)$ for every m. The normalized random allocation above uses at most L frequencies, so the block construction gives, for every $m \ge \nu L$,

$$\tfrac{1}{2}\nu^{1-s}B_pL^{1-s} \le \gamma^*_{b,m}(p) \le B_p\Gamma_s(p).$$

Together with the programme upper bound and $B_p = 2(2p\nu)^{-\nu/2}$, this proves the stated order uniformly over these widths. If m is also large enough to realize a programme maximizer, Theorem 8 gives the stronger lower bound $\nu^{1-s}B_p\Gamma_s(p)$.

Corollary 12. If $A = 0$, Lemma 2 gives $h(\theta) \le 0$, so the claim is immediate. Otherwise $y = A/\|A\|_{1/s}$ has budget one and at most K non-zero entries. Lemma 2, the budget estimate in Theorem 4, and equation 10 give

$$h(\theta) \le \|A\|_{1/s}m(y), \qquad \|A\|_{1/s} \le C_b^{\nu/b}\|\theta\|_{2,b}^{\nu},$$

which prove the displayed bound. For product networks, $C_b^{\nu/b} = B_p$. At any comparison width $m \ge \nu L$, the lower bound just proved shows that the ratio to $\gamma^*_{b,m}(p)$ is at most

$$3(9\nu)^{s-1}\Big(\frac{x}{e^{x/(2a_s)}-1}\Big)^s.$$

This tends to zero as $x = \ln p/K \to \infty$. □

B.6 Further remarks on Section 4

Equal shares. Restricting equation 2 to equal shares on a set of K frequencies, with $M_K(p) := \max_{|S|=K}\min_{\delta\neq 0}\sum_{j\in S}\kappa_j(\delta)$, gives $\Gamma_s(p) \ge \max_K K^{-s}M_K(p)$, a cheap approximation. Since the worst offset is no better than the average, $M_K \le pK/(p-1)$ (cf. Slepian, 1968, Thm. 2), so $K^{-s}M_K \le \frac{p}{p-1}K^{1-s}$ decays in K for $s > 1$. In Table 1 the equal-share argmax agrees with K^* except at $(p,\nu) = (59,3)$, where it gives 6, and $(113,3)$, where it gives 5; where adjacent equal-share values are within a few percent, the honest statement is a band of width one.

Shares in the block. Optimal shares are not equal (at $p = 59$, $s = 3/2$ they range over 0.15–0.25), and a ν-block with equal amplitudes on the optimal support does not attain Theorem 8.

Weaker cancellation suffices. The inequality of Lemma 7 uses only the $\binom{\nu}{2}$ interval patterns $T = \{t'{+}1,\dots,t\}$ of equation 3. The structure statement of Theorem 8 still needs the full condition: under the interval patterns alone, the equality set contains configurations that are not interference-free (at $\nu = 4$, the pattern $T = \{1,3\}$ can be non-zero at equality).

Other networks. The upper bound of Theorem 8 uses only the single-neuron budget and the sign rule. For the power network it holds with B_p replaced by $n!\,B_p$; the product-network attainment construction does not establish sharpness for the power network (Appendix D). For a network whose neurons satisfy the sign rule with n operand factors, the factor $\nu^{-(\nu-b)/b}$ becomes $(n+1)^{-(\nu-b)/b}$: it counts factors, not degree.

Scope. (i) Lemma 1 averages over every input; with a training fraction below one, the training-set margin need not obey the bound. (ii) Interference-freeness is a condition on the network, not a fact about trained networks, which carry interference at the sign-flipped modes (Section 5). (iii) K^* is the support size of a maximizer of equation 2.

C THE SINGLE-NEURON BUDGET AND THE SIGN RULE FOR CONCRETE NETWORKS

Product neuron. $\phi(\{u_1, \dots, u_n, w\}, x) = u_1(a_1) \cdots u_n(a_n)\, w$, $\nu = n + 1$. By Appendix A.1, $\beta_j = 2p^{-\nu} \prod_t \hat{u}_t(j)\, \overline{\hat{w}(j)}$. The coefficient at the mode flipping T is $2p^{-\nu} \prod_{t \notin T} \hat{u}_t(j) \prod_{t \in T} \hat{u}_t(-j)\, \overline{\hat{w}(j)}$, and $\hat{u}_t(-j) = \xi_{tj} \hat{u}_t(j)$ with $\xi_{tj} = \overline{\hat{u}_t(j)}/\hat{u}_t(j)$ (:= 1 if $\hat{u}_t(j) = 0$); this is Lemma 6. For the budget, let $m_j = \sum_t |\hat{u}_t(j)|^2 + |\hat{w}(j)|^2$. AM–GM over the ν factors gives $|\prod_t \hat{u}_t(j) \hat{w}(j)| \le (m_j/\nu)^{\nu/2}$. Since $b/2 \ge 1$, $t \mapsto t^{b/2}$ is superadditive, so $\sum_j m_j^{b/2} \le (\sum_j m_j)^{b/2}$; and Parseval gives $\sum_{j \in F} m_j \le \frac{p}{2} \|\omega\|_2^2$. Hence

$$\sum_{j \in F} |\beta_j(\omega)|^{b/\nu} \le (2p^{-\nu})^{b/\nu} \Big(\frac{p}{2\nu}\Big)^{b/2} \|\omega\|_2^b,$$

$$C_b^{\nu/b} \le 2p^{-\nu} \Big(\frac{p}{2\nu}\Big)^{\nu/2} = B_p.$$

For a non-zero neuron, equality in the budget inequality holds if and only if it is balanced (all ν Fourier magnitudes equal at every $j \in F$), has no constant component, and, for $b > 2$, uses a single frequency; at $b = 2$ the superadditivity step is an identity, so a neuron may carry several frequencies at no extra cost. The zero neuron also gives equality. A neuron whose ν weight vectors are cosines of one frequency with equal norms attains the budget, so $C_b^{\nu/b} = B_p$.

Power neuron. $\phi(\{u_1, \dots, u_n, w\}, x) = \big(\sum_t u_t(a_t)\big)^n w$, $\nu = n + 1$. Expand the power into monomials $\prod_t u_t(a_t)^{e_t}$ with $\sum_t e_t = n$. A monomial with some $e_t = 0$ has no Fourier content in slot t, and one with some $e_t \ge 2$ has some other $e_{t'} = 0$; so only the multilinear monomial, with coefficient $n!$, reaches a mode with all ν entries non-zero. Hence $\beta_j = n! \cdot 2p^{-\nu} \prod_t \hat{u}_t(j)\, \overline{\hat{w}(j)}$, Lemma 6 holds with the same ξ_{tj}, and $C_b^{\nu/b} = n!\, B_p$: the upper bound follows from the product budget, and a balanced single-frequency cosine neuron attains it. In particular $C_3 = 2B_p$ at $n = 2$. This is a single-neuron budget identity, not a network-level attainment claim. Theorem 4 and the upper bound of Theorem 8 apply. Attainment requires additional control of the monomials with a zero operand slot: Definition 5 permits them, but they can reduce the worst-case margin. Consequently the product-block identity alone does not give an attaining power network.

Where the approach does not apply. A network with a shared linear layer in front of the neurons is not a sum of neurons each paying its own norm: the shared layer is priced by a nuclear norm (Srebro et al., 2004), which does not penalize carrying many frequencies in one neuron, and the single-neuron budget does not bound it. Such networks are outside this paper.

D THE GLOBAL CONJECTURE, AND WHAT IS KNOWN

Conjecture 20. *For odd prime p, $n \ge 1$ and $2 \le b \le \nu$, the maximum $L_{2,b}$-margin over* all *product networks equals the interference-free value of Theorem 8: no network carrying interference exceeds it.*

Proved here. At $b = \nu$, Corollary 14 and Proposition 9 give matching unrestricted upper and lower bounds (at sufficient width). For $2 \le b < \nu$, Theorem 10 determines the order of the unrestricted optimum, while Theorem 8 gives the exact interference-free value. Equality of these two optima throughout this range remains open.

Power networks. The universal and interference-free upper bounds apply with single-neuron constant $C_b^{\nu/b} = n! B_p$. Their attainment is a separate question from the product-network result. At

$b = \nu$ the bound is attained for every n (Li et al., 2025, Thm. 4.1) (at $n = 2$, Morwani et al., 2024, Thm. 7), with the value of Corollary 14.

b outside $[2, \nu]$. For $b > \nu$ the normalized margin is unbounded as the width grows (cf. the unit-duplication argument of Neyshabur et al., 2015, Sec. 3.2). For $b < 2$ the superadditivity step of Appendix C fails for the product network (one neuron carrying every frequency exceeds it by $|F|^{1-b/2}$). The available bound is the $b = 2$ problem with a rescaled budget, which is not attained, so nothing about the number of features is claimed there.

E EXPERIMENTAL DETAILS

Recipe. All networks have width $m = \nu(p-1)$, entries of every weight vector drawn i.i.d. from $N(0, 1)$, and no biases. They are trained by full-batch gradient descent with heavy-ball momentum 0.9 and constant learning rate $c\,p$ on mean cross-entropy over D_p plus $\lambda \sum_i \|\omega_i\|_2^b$, with constant $\lambda = \tilde{\lambda}\, B_p^{b/\nu}$. The penalty is part of the gradient (not decoupled). Training is in float32 with TF32 disabled; margins, spectra and ablations are computed in float64 (the relative gradient of the convergence criterion is computed in float32). The initial weights depend only on (n, p, m) and the seed, so runs with the same seed in different cells of one prime start from the same weights, and comparisons across b are paired. The constants below were fixed by a tuning phase with a declared mechanical rule: the most runs meeting the convergence criterion, then the smaller $\tilde{\lambda}$, then the lower median relative gradient. No prediction entered the rule, tuning runs are not evaluation data, and the constants were frozen before any evaluation run; evaluation runs with seed 0 at the tuning primes repeat the first 40k steps of a tuning run exactly. At $\nu = 4$, $p = 59$, where c was not tuned, the learning rate $c\,p$ made 8 of 12 runs diverge.

network	c	$\tilde{\lambda}$	steps	primes
product, $\nu = 3$	0.40	0.003	150k	11–113
product, $\nu = 4$	0.28	0.003	150k (90k at $p = 59$)	17–59
product, $\nu = 5$	0.035	0.003	300k	11–17
power, $\nu = 3$	0.10	0.01	150k	23–113

$\nu = 5$ was not run at $p \geq 23$: no tuning run converged at $p = 23$, and longer runs did not fit the compute limit set in advance (two hours per sweep).

Protocol. Product networks (and, at $b = \nu$, power networks) are trained on the full table D_p, with no train/test split. A run of T steps is *converged* if at step T its relative gradient $\|\nabla(\mathrm{CE} + \lambda R)\|/(\|\nabla \mathrm{CE}\| + \lambda\|\nabla R\|)$ is at most 10^{-3}, its total loss and its normalized margin differ from their values at step $0.9T$ by relative amounts of at most 10^{-4} and 10^{-3}, and its margin is positive; other runs are reported separately. Restricting to converged runs, as preregistered, favours sparser networks: in cells that have both, non-converged runs count 1.3 more frequencies on average. With them included, the direction test of Section 5 gives 36 of 41 pairs (shuffle null 38%, $p = 0.001$). K^* is exact for $p \leq 37$ (exhaustive vertex enumeration, Lemma 18) and best found, not certified, above. At two cells ($p = 113$, $s = 1.5$ and $p = 59$, $s = 2$) an earlier, independent support search found a higher value of $\Gamma_s(p)$ than the preregistered prediction (with $K^* = 7$ instead of 6 at the first); since Γ_s is a maximum we use the higher value there (the allocations were re-evaluated with the same code). In the main text a frequency is counted if $A_j \geq 0.2 \max_k A_k$. The grid, seeds, predictions and matching criteria were preregistered before any evaluation run.

Sweeps. Every launched run is listed with its status in the supplementary material; none is omitted.

- *Main sweep* (Figures 1b and 2). Product networks at $\nu = 3$ ($p \in \{23, 31, 43, 59, 83, 113\}$, $s \in \{1.25, 1.5\}$, 6 seeds), $\nu = 4$ ($p \in \{17, 23, 31, 43, 59\}$, $s \in \{1.25, 1.5, 2\}$, 4 seeds) and $\nu = 5$ ($p \in \{11, 13, 17\}$, $s \in \{1.25, 1.5, 2, 2.5\}$, 4 seeds). Of these runs, 126 converged, 46 did not, and 8 diverged. A further 48 trained runs were lost in transfer before their results were verified; they were marked invalid and rerun under new identifiers.
- *Sweep in p* (Figure 2). Product, $\nu = 3$, $p \in \{11, 13, 17, 19, 29, 37, 47, 71\}$, $s \in \{1.25, 1.5\}$, 3 seeds. 44 converged and 4 did not.

- *Fine sweep in b* (Figure 1a). Product, $\nu = 4$, $p = 37$, 16 values of s in $[1, 2]$ placed inside the exact plateaus of $K^*(s)$, 2 seeds, whose two initializations are shared by all 16 values of s. 19 converged and 13 did not.
- *Dense endpoint*. $b = \nu$: product $\nu = 3$ ($p = 23, 59, 113$), $\nu = 4$ ($p = 23, 43, 59$) and $\nu = 5$ ($p = 11, 13, 17$), and power $\nu = 3$ ($p = 23, 59, 113$), 4 seeds. 23 converged, 24 did not and 1 diverged; 12 further launches were stopped before training and rerun with identical configurations.
- *Relabelling*. Product, $\nu = 3$, $p = 23$, $b = 2$: 12 pairs of runs whose initial weights differ by the relabelling $j \mapsto uj$ of frequencies, and further seeds for the frequency marginals; 50 runs, 48 converged. Not used in the main text; its 6 unrelabelled runs coincide with main-sweep runs.

Numbers behind Section 5. Converged runs; counts use $A_j \geq 0.2 \max_k A_k$.

- *Clean energy.* Share of the logit table's energy in the modes $(j, \dots, j; -j)$: median 0.97, minimum 0.47 over all converged runs. At $\nu = 3$ the median is 0.98 ($s = 1.25$) and 0.96 ($s = 1.5$); at $\nu = 5$, 0.97 ($s = 1.25$) and 0.83 ($s = 2.5$); at $b = \nu$, 0.99. Spearman correlation with s within $\nu = 3, 4, 5$: -0.15, -0.37, -0.32.
- *Quality of the learned allocation* (main sweep, 126 runs). The programme value $\min_\delta \sum_j y_j \kappa_j(\delta)$ at $y = A/\|A\|_{1/s}$, divided by $\Gamma_s(p)$: median 0.87 (95% interval over seeds $[0.85, 0.88]$), and 12% of runs reach 0.95. Medians over cells of the nulls: 0.45 for 1000 random allocations on K^* frequencies, 0.67 for the network at initialization; the ν-block construction scores 1. Budget shares against y^* (preregistered): median total variation 0.42, against 0.52 for random allocations.
- *Universal bound.* Never exceeded in any of the 349 run records with a finite margin (converged or not; 6 of them duplicate main-sweep runs); the largest normalized margin is 0.76 of the bound for $b < \nu$ and 0.99999 at $b = \nu$.
- *Fine sweep in b* ($\nu = 4$, $p = 37$; Figure 1a). The converged run at $s = 1$ uses all 18 frequencies; every converged run at $s \geq 1.45$ is within one of K^*; between $s = 1.05$ and 1.25, where K^* passes through seven plateaus from 18 to 9, trained counts lie below K^*. The preregistered plateau-identity test matched in 2 of 19 converged runs.
- *Direction test*. Pairs of cells at fixed (ν, p) with different s and different K^* (main and p sweeps): the median count moves with K^* in 30 of 36 pairs, stays put in 4 and reverses in 2; shuffling counts across the cells of each (ν, p) gives 38% (95th percentile 53%; one-sided $p = 0.001$). Fine sweep: 83 of 96 pairs, against 47%. Slope of the change in count on the change in K^*: 0.49 (0.28–0.61 across count definitions and sweeps). Pooled per-run agreement within one of K^*: 71% of runs, against about half for the best constant.
- *Dense end* ($b = \nu$). Every converged product network at $\nu = 3$ ($p = 23, 59, 113$) and $\nu = 4$ ($p = 23$), and every converged power network at $\nu = 3$, $p = 23$, uses every frequency; per-cell medians of the smallest amplitude over the largest are 0.55–1.00 and of the total variation to uniform at most 0.034; these runs reach 0.87–1.00 of the bound. At $\nu = 5$, 2 of 7 converged runs (both at $p = 11$) are dense and flat; the others miss one to three frequencies or are far from flat (per-cell median total variation 0.10–0.41). No run converged at $\nu = 4$, $p \geq 43$, or for the power network at $p \geq 59$. At initialization all but at most two frequencies already exceed 10% of the largest amplitude.
- *Degree at fixed $b = 2$*. The median count moves as K^* predicts in 5 of 7 pairs, with one tie and one reversal ($p = 0.16$; 3 to 7 of 7 across count definitions).
- *Same s, different ν*. Over the 17 pairs of cells with the same (p, s) and different ν, the median counts agree to within one in 12; the higher-degree cell has the smaller count in 13, the same in 3 and the larger in 1 (sign test, two-sided $p = 0.002$; mean difference -1.1). Mean count minus K^* over cells with $s \leq 1.5$, $p \leq 43$: $+0.28$ ($\nu = 3$, 18 cells), -0.64 ($\nu = 4$, 7), -1.25 ($\nu = 5$, 6). Median quality of the learned allocation in the main sweep: 0.89, 0.88 and 0.79 at $\nu = 3, 4, 5$. For comparison, cells with the same ν and different s differ by 1.3 on average, where K^* differs by 2.3.

- *Across primes* ($\nu = 3$; Figure 2). The median count is within one of K^* in 20 of 28 (p, s) cells, against 16 for the best constant at each s. At $s = 1.5$ it exceeds K^* by 1.5 to 3 at $p = 59$–83 and matches it at $p = 113$.

Count definitions. We scored the count K of a trained network under thirteen definitions. The number of frequencies with $A_j \geq t \max_k A_k$ for $t = 0.2$ (used in the main text), 0.1, 0.05 and 0.01; the participation ratios $(\sum_j A_j)^2 / \sum_j A_j^2$ and $(\sum_j w_j)^2 / \sum_j w_j^2$ of the amplitudes and of the budget shares $w_j = A_j^{1/s} / \sum_k A_k^{1/s}$; the fewest frequencies carrying 90, 95 or 99% of the budget, the number with energy share $A_j^2 / \sum_k A_k^2 \geq 0.02$ or ≥ 0.05, and the fewest frequencies carrying 95% or 90% of $\sum_k A_k^2$. The direction tests of Section 5 under each definition (converged runs):

definition	s sweep, main + p sweeps same sign (ties)	slope	fine b sweep same sign	slope	ν at $b = 2$ same sign (ties)	same s / same ν mean $\lvert\Delta K\rvert$
$A_j \geq 0.2 \max$ (R)	30/36 (4)	0.49	83/96	0.59	5/7 (1)	1.12 / 1.31
$A_j \geq 0.1 \max$ (R)	30/36 (3)	0.48	68/96	0.59	4/7 (1)	0.97 / 1.29
$A_j \geq 0.05 \max$ (R)	30/36 (3)	0.49	68/96	0.59	4/7 (1)	0.88 / 1.36
$A_j \geq 0.01 \max$ (R)	31/36 (2)	0.49	68/96	0.61	4/7 (1)	0.85 / 1.37
participation, A (R)	34/36 (0)	0.45	83/96	0.56	6/7 (0)	1.01 / 1.18
participation, w (R)	31/36 (0)	0.39	73/96	0.49	4/7 (0)	0.93 / 1.10
90% of budget (R)	28/36 (5)	0.33	69/96	0.41	4/7 (1)	0.97 / 0.99
95% of budget (R)	30/36 (4)	0.44	64/96	0.44	4/7 (1)	0.97 / 1.19
99% of budget (R)	26/36 (5)	0.44	68/96	0.59	3/7 (1)	0.85 / 1.48
energy $\geq 2\%$	30/36 (4)	0.44	74/96	0.42	5/7 (1)	1.06 / 1.10
energy $\geq 5\%$	27/36 (8)	0.28	74/96	0.49	6/7 (1)	1.15 / 0.77
95% of energy	27/36 (8)	0.39	70/96	0.45	5/7 (2)	0.94 / 0.98
90% of energy	28/36 (7)	0.33	77/96	0.47	7/7 (0)	1.09 / 0.90

The null shuffles the counts across the cells of each group (1000 draws). Its mean agreement is 0.31–0.51 for the s sweeps and 0.43–0.52 for the fine b sweep, depending on the definition. One-sided p-values are 0.001 (the smallest possible with 1000 draws) for the s sweeps under every definition, 0.001–0.041 for the fine b sweep, and 0.001–0.66 for the ν comparison at $b = 2$.