Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Exact spectral allocation in Fourier-feature neural tangent kernels

Bin Zhang

zhangbin@buaa.edu.cn

Beihang University

Weifei Jiang

Beihang University

Research Article

Keywords: Fourier features, Neural tangent kernels, Spectral bias, Kernel methods, Implicit neural representations, Dictionary design

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11338822/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Exact spectral allocation in Fourier-feature neural tangent kernels

Bin Zhang[1*] and Weifei Jiang[1]

[1*]Hangzhou International Innovation Institute, Beihang University, Hangzhou, China.

*Corresponding author(s). E-mail(s): zhangbin@buaa.edu.cn;
Contributing authors: weifeijiang@buaa.edu.cn;

Abstract

Fourier features determine the kernel spectrum of coordinate networks, yet bandwidth and harmonic support leave the learning rates of individual frequencies unspecified. We study fixed, finite integer Fourier dictionaries on the torus. For a dot-product kernel with nonnegative summable power-series coefficients, each Fourier eigenvalue is a mixture of signed lattice-walk probabilities. This separates architecture-dependent coefficients from dictionary-dependent multiplicities and yields exact support, finite-rank approximations with trace-norm error bounds, and certified comparisons of spectral design objectives. For the two-layer bias-free rectified linear unit (ReLU) neural tangent kernel (NTK), we characterize an even-walk subgroup that can exclude frequencies inside the generated lattice and derive the leading one-dimensional spectral tail. Two dictionaries matched in cardinality, maximum frequency, and base second moment have identical leading tails but a **19.51**-fold difference in their mode-**9** eigenvalues; finite-network decay rates track this difference. Across three target frequencies and 108 dictionary–target pairs, spectral predictions have rank correlations of **0.977**–**0.992** with fitted rates. Under a four-frequency budget for twenty target modes, spectral selection reduces off-grid finite-network mean squared error (MSE) by **12.3%** relative to direct top-energy selection on a mixed-spectrum signal. Fifty runs on five natural-image crops show close agreement with finite-grid MSE predictions and larger deviations in individual mode rates. These results quantify how a fixed Fourier dictionary allocates kernel mass and predicts learning in the NTK regime.

Keywords: Fourier features, Neural tangent kernels, Spectral bias, Kernel methods, Implicit neural representations, Dictionary design

1 Introduction

Fourier embeddings are widely used to improve the learning of oscillatory signals by coordinate-based neural networks. Tancik et al. [1] showed that the embedding converts a dot-product neural tangent kernel (NTK) into a stationary kernel and allows its bandwidth to be adjusted. Meanwhile, harmonic analyses of implicit neural representations (INRs) describe the combinations of base frequencies that a nonlinear network can generate [2–4]. Together, these perspectives explain why Fourier features help and why nonlinear synthesis extends beyond the supplied frequencies.

For a particular finite dictionary, how much kernel mass does each synthesized frequency receive? Harmonic support identifies which modes are present, while bandwidth summarizes the overall spectrum. The local eigenvalue determines the kernel-regime convergence rate of an individual mode. For example, $B_A = (1, 8, 10)$ and $B_B = (4, 7, 10)$ have equal size, maximum frequency, and sum of squares. Both generate every integer, yet their two-layer rectified linear unit (ReLU) NTKs assign mode 9 eigenvalues of approximately 0.06155 and 0.003155. Their leading high-frequency tails also agree. This pair demonstrates the variation in local learning speed that can remain after matching global spectral scale.

We start from the Fourier-feature kernel composition $K_B(\theta, \theta') = \Psi(g_B(\theta - \theta'))$, where g_B is the average cosine of the base frequencies. Expanding the scalar profile gives

$$\lambda_m(B) = \sum_{\ell \geq 0} a_\ell \frac{N_{\ell,B}(m)}{(2M)^\ell}, \qquad \Psi(u) = \sum_{\ell \geq 0} a_\ell u^\ell. \tag{1}$$

Here $N_{\ell,B}(m)$ counts ordered signed length-ℓ representations of m using the M dictionary entries. Fourier convolution gives this identity order by order. Summing the contributions separates the architecture coefficients from the dictionary multiplicities and provides the basis for the support, approximation, and design results below.

We develop three parts of this analysis. First, the allocation formula gives uniform and trace-norm truncation bounds, grid-aliasing control, and certified comparisons of spectral objectives. Second, the ReLU specialization gives exact even-walk support and a one-dimensional leading spectral tail, connecting arithmetic restrictions to both local eigenvalues and global decay. Third, numerical checks, 846 synthetic-target runs, and 50 natural-image runs test the learning predictions and fixed-budget design criterion. The experiments include multiple targets, finite-width diagnostics, and grid-sensitivity measurements. Code and complete numerical records are provided in Online Resource 1.

2 Related work

2.1 NTKs, spectral bias, and power-series kernels

Jacot et al. [5] and Lee et al. [6] connect wide-network training to kernel dynamics. Rahaman et al. [7] study Fourier spectral bias, while Cao et al. [8] analyze eigendirection-dependent convergence. Arora et al. [9] develop exact infinite-width kernel computations. Murray et al. [10] derive architecture-dependent NTK power-series coefficients; these provide the natural scalar input to our calculation. Related

work establishes spectral decay and parity properties for residual and other neural kernels [11, 12]. Our one-dimensional tail calculation concerns the composed kernel along an integer Fourier embedding, including its arithmetic repetition and antipodal singularities.

2.2 Fourier features and harmonic expansions

Random Fourier features approximate a prescribed stationary kernel [13, 14]; here a fixed dictionary is composed with a nonlinear dot-product kernel. Tancik et al. [1] diagonalize periodic-grid composed kernels by the fast Fourier transform (FFT) and use their spectra to study learning and reconstruction. Wang et al. [15] examine Fourier-feature NTK eigenvector bias in regression and multiscale partial differential equations (PDEs). Benbarka et al. [2] connect Fourier mappings to integer-lattice Fourier series, and Yüce et al. [3] derive structured harmonic dictionaries and discuss the weights and decay of generated harmonics. Supplementary Lemma I.4 of Ma et al. [4] gives a finite-order harmonic expansion of a Fourier-feature NTK in support of adaptive filtering. Building on these expansions, we make the normalized signed multiplicities explicit, control the summable infinite-order limit, and derive support and numerical certificates for a fixed dictionary.

2.3 Task-dependent encoding and learning consequences

FreSh [16] selects encoding hyperparameters by aligning the initial network output spectrum with the target spectrum. Molina et al. [17] use frequency-domain error dynamics to optimize Fourier-feature initialization distributions. CAFE [18] constructs and selects frequency combinations, while STAF [19] learns sinusoidal activation parameters. These methods adapt encodings or activations for INR fitting. Our task-dependent selection evaluates a spectral criterion for fixed integer dictionaries. The population learning and ridge formulas use established kernel principles, including task–model alignment [20] and effective dimension [21].

3 Exact allocation and numerical certificates

3.1 Periodic dictionary and kernel profile

Let $\mathbb{T}^d = (\mathbb{R}/2\pi\mathbb{Z})^d$ carry normalized Haar measure $\mathrm{d}\rho = (2\pi)^{-d}\,\mathrm{d}\theta$. The Fourier characters $\varphi_m(\theta) = e^{im^\mathsf{T}\theta}$, $m \in \mathbb{Z}^d$, form an orthonormal basis of $L^2(\rho)$. Let $B = (\omega_1, \ldots, \omega_M)$ be a multiset of nonzero integer vectors. Repetitions are allowed. Define

$$\gamma_B(\theta) = M^{-1/2}\big(\cos(\omega_j^\mathsf{T}\theta), \sin(\omega_j^\mathsf{T}\theta)\big)_{j=1}^{M}, \quad g_B(t) = \frac{1}{M}\sum_{j=1}^{M}\cos(\omega_j^\mathsf{T} t). \tag{2}$$

The embedding has unit norm and inner product $g_B(\theta - \theta')$. A scalar dot-product profile therefore induces the stationary kernel $K_B(\theta, \theta') = \kappa_B(\theta - \theta')$, with $\kappa_B = \Psi \circ g_B$.

Assumption 1 (Nonnegative summable profile) $\Psi(u) = \sum_{\ell \geq 0} a_\ell u^\ell$ for $u \in [-1, 1]$, where $a_\ell \geq 0$ and $A := \sum_{\ell \geq 0} a_\ell < \infty$.

Nonnegative coefficients give the support equivalence and one-sided approximation bounds below. The eigenvalue identity also holds for absolutely summable signed coefficients. Under Assumption 1, Ψ is analytic on $(-1, 1)$ and continuous at the endpoints; its endpoint derivative may diverge, as it does for the ReLU profile. We use a geometric profile to test the general positive-kernel calculation and the explicit ReLU NTK for finite-network predictions.

Definition 1 (Signed multiplicity and step law) For $\ell \geq 1$, let

$$N_{\ell,B}(m) = \#\left\{(j_r, s_r)_{r=1}^{\ell} \in ([M] \times \{-1, 1\})^\ell : \sum_{r=1}^{\ell} s_r \omega_{j_r} = m\right\},$$

$$N_{0,B}(m) = \mathbf{1}\{m = 0\}.$$

Write $\mu_B = (2M)^{-1} \sum_j (\delta_{\omega_j} + \delta_{-\omega_j})$ and $\mathcal{D} = \{\ell : a_\ell > 0\}$.

The entries of the multiset are labeled in this definition: coincident signed frequencies accumulate their probabilities.

Theorem 1 (Exact allocation) *Under Assumption 1, the integral operator* $T_B f = \int K_B(\cdot, \theta') f(\theta') \, \mathrm{d}\rho(\theta')$ *is positive, self-adjoint, and trace class. Its eigenpairs are* (φ_m, λ_m), *where*

$$\boxed{\lambda_m(B) = \sum_{\ell \geq 0} a_\ell \mu_B^{*\ell}(m) = \sum_{\ell \geq 0} a_\ell \frac{N_{\ell,B}(m)}{(2M)^\ell}.} \tag{3}$$

The series converges uniformly in m, $\lambda_m = \lambda_{-m}$, *and*

$$\lambda_m > 0 \iff \exists \ell \in \mathcal{D} : \; N_{\ell,B}(m) > 0, \qquad \sum_m \lambda_m = \mathrm{Tr}(T_B) = A = \Psi(1). \tag{4}$$

Proof: Appendix A.1.

Since $g_B(t) = \sum_m \mu_B(m) e^{i m^\top t}$, its ℓth power convolves the step law ℓ times. Summability permits termwise integration. Architecture determines a_ℓ, and dictionary geometry determines the distribution of each order's contribution across modes. Changing the dictionary therefore redistributes a fixed total spectral mass.

Let $r(m) = \inf\{\ell \in \mathcal{D} : N_{\ell,B}(m) > 0\}$, with $r(m) = \infty$ if the set is empty. When $r(m) < \infty$,

$$a_{r(m)} \mu_B^{*r(m)}(m) \leq \lambda_m(B) \leq \sum_{\ell \geq r(m)} a_\ell. \tag{5}$$

Thus both arrival order and multiplicity matter. Also, $\lambda_m = 0$ outside the generated group $G_B = \mathrm{span}_{\mathbb{Z}} B$. With $\Omega = \max_j \|\omega_j\|_\infty$, a length-$\ell$ path cannot reach beyond $\ell\Omega$, giving $\lambda_m \leq \sum_{\ell \geq \lceil \|m\|_\infty / \Omega \rceil} a_\ell$. These direct consequences are proved in Appendix A.1.

3.2 Truncation, trace norm, and grid aliasing

Define $R_L = \sum_{\ell>L} a_\ell$, $\Psi_L(u) = \sum_{\ell=0}^{L} a_\ell u^\ell$, and $\lambda_{m,L} = \sum_{\ell=0}^{L} a_\ell \mu_B^{*\ell}(m)$. The corresponding operator is denoted $T_{B,L}$.

Theorem 2 (Certified finite approximation) *For every $L \geq 0$, $T_{B,L}$ has rank at most $(2L\Omega+1)^d$ and*

$$\|\kappa_B - \Psi_L \circ g_B\|_\infty = R_L, \qquad \|T_B - T_{B,L}\|_{\mathrm{op}} = \sup_m(\lambda_m - \lambda_{m,L}) \leq R_L, \tag{6}$$

$$\|T_B - T_{B,L}\|_1 = \sum_m (\lambda_m - \lambda_{m,L}) = R_L. \tag{7}$$

Here $\|\cdot\|_1$ is the trace norm. On a uniform N^d grid, the normalized kernel matrix has Fourier eigenvalues

$$\lambda_m^{[N]} = \sum_{k\in\mathbb{Z}^d} \lambda_{m+Nk}. \tag{8}$$

For $\|m\|_\infty < N/2$, setting $L_m = \lceil (N - \|m\|_\infty)/\Omega \rceil - 1$ gives

$$0 \leq \lambda_m^{[N]} - \lambda_m \leq R_{L_m}. \tag{9}$$

An order-L polynomial spectrum is free of aliasing on its entire support if $N > 2L\Omega$.

Proof: Appendix A.2.

The uniform kernel bound is attained at $t = 0$, and the trace-norm equality follows from nonnegative omitted eigenvalues. Equation (8) relates the discrete-grid and continuum spectra: a full-kernel FFT exactly diagonalizes the grid matrix and approximates the continuum coefficients through aliasing. For the geometric profile $\Psi_q(u) = (1-q)/(1-qu)$, $R_L = q^{L+1}$. ReLU coefficients have a slower tail; their numerical resolution is assessed separately.

4 ReLU support and spectral scale

4.1 Global moments and the ReLU endpoint

Proposition 3 (Spectral second moment) *If $\sum_\ell \ell a_\ell < \infty$ and $\Sigma_B = M^{-1}\sum_j \omega_j\omega_j^\mathsf{T}$, then*

$$\sum_m m m^\mathsf{T} \lambda_m(B) = \Psi'(1)\Sigma_B. \tag{10}$$

If $\sum_\ell \ell a_\ell = \infty$, the directional moment $\sum_m (v^\mathsf{T} m)^2 \lambda_m$ is infinite for each v with $v^\mathsf{T}\Sigma_B v > 0$.

Proof: Appendix A.3.

For the bias-free network

$$f_n(\theta) = \sqrt{2/n}\sum_{i=1}^{n} v_i \,\mathrm{ReLU}(w_i^\mathsf{T}\gamma_B(\theta)), \tag{11}$$

with independent standard Gaussian v_i, w_i and both layers trained, the limiting NTK profile is

$$\Psi_R(c) = \frac{\sqrt{1-c^2} + 2c(\pi - \arccos c)}{\pi}. \tag{12}$$

Its trace is $\Psi_R(1) = 2$, and its coefficients are

$$a_0 = \pi^{-1}, \quad a_1 = 1, \quad a_{2k} = \frac{2k+1}{\pi(2k-1)^2} \frac{\binom{2k}{k}}{4^k} \ (k \geq 1), \quad a_{2k+1} = 0 \ (k \geq 1). \tag{13}$$

In particular, $a_{2k} \sim (2\pi^{3/2})^{-1} k^{-3/2}$, so Assumption 1 holds and $\Psi_R'(1) = \infty$. Matching base second moments therefore does *not* match a finite ReLU spectral second moment. Appendix B.1 derives Equations (12)–(13) from the network parameterization.

4.2 Exact support beyond the generated lattice

Theorem 4 (Even-walk support of the two-layer ReLU NTK) *Let*

$$H_B = \mathrm{span}_{\mathbb{Z}}\{2\omega_1, \omega_2 - \omega_1, \ldots, \omega_M - \omega_1\}. \tag{14}$$

Then H_B is the set of all endpoints of even-length signed walks, $[G_B : H_B] \leq 2$, and

$$\mathrm{supp}(\lambda^R) = H_B \ \cup \ \{\omega_j, -\omega_j : 1 \leq j \leq M\}. \tag{15}$$

Proof idea. Every even signed walk decomposes into pairs whose endpoints lie in H_B, while each generator in Equation (14) is itself realized by a two-step walk. The positive even coefficients of the ReLU profile therefore contribute exactly H_B, and the linear coefficient contributes the base atoms $\pm\omega_j$. Nonnegativity prevents cancellation between degrees. *Proof:* Appendix B.1.

For $B = (1, 3)$, the generated lattice is $\mathbb{Z}$ but $H_B = 2\mathbb{Z}$. Odd modes other than $\pm 1, \pm 3$ have zero eigenvalue. This example separates a lattice obstruction from an active-degree obstruction. For $B = (1, 2)$, $H_B = \mathbb{Z}$ and every mode has positive mass. Architectural changes such as extra layers or biases can change the active degrees; Theorem 4 is specific to Equation (12).

4.3 A shared spectral tail can conceal distinct learning rates

Theorem 5 (One-dimensional ReLU tail) *Let $d = 1$, $h = \gcd(|\omega_1|, \ldots, |\omega_M|)$, and $\sigma_B = (M^{-1} \sum_j \omega_j^2)^{1/2}$. Set $\chi_B = 1$ when every ω_j/h is odd, and $\chi_B = 0$ otherwise. Off $h\mathbb{Z}$, $\lambda_m^R = 0$. Along nonzero $m \in h\mathbb{Z}$, as $|m| \to \infty$,*

$$\lambda_m^R = \frac{h\sigma_B}{\pi^2 m^2} \left[1 + \chi_B(-1)^{m/h}\right] + o(|m|^{-2}). \tag{16}$$

When $\chi_B = 1$, odd multiples of h outside the base atoms are exactly absent, as in Theorem 4.

Proof: Appendix B.2.

The coefficient in Equation (16) comes from cusps of $\Psi_R(g_B(t))$ at $g_B(t) = 1$ and, when present, $g_B(t) = -1$. It supplies a global ReLU scale even though the spectral second moment diverges. For $B_A = (1, 8, 10)$ and $B_B = (4, 7, 10)$, $h = 1$, $\chi_B = 0$,

and $\sigma_B = \sqrt{55}$, so both have leading tail $\sqrt{55}/(\pi^2 m^2)$. Their local allocations differ already at the first active order. Dictionary B_A reaches 9 at order two through $1+8$ and $10-1$ and their ordered permutations, giving $N_{2,B_A}(9) = 4$. Dictionary B_B first reaches 9 at active order four, for example $10+10-7-4 = 9$. The first contributions are therefore $a_2(4/36)$ and $a_4 N_{4,B_B}(9)/6^4$, respectively. Summing all orders quantifies the local eigenvalue contrast.

5 Learning and dictionary design

Under population kernel gradient flow, $\dot{f}_t = T_B(f^\star - f_t)$, each residual coefficient obeys $e_m(t) = e^{-t\lambda_m} e_m(0)$. Population kernel ridge regression with penalty $\eta > 0$ gives coefficients $\widehat{f}_\eta(m) = \lambda_m \widehat{f^\star}(m)/(\lambda_m + \eta)$. These follow by diagonalization and are proved, together with the reproducing kernel Hilbert space (RKHS) characterization and effective-dimension bounds, in Appendix C.1.

For a normalized periodic target $f^\star(\theta) = \sum_{m\in S} \sqrt{2w_m}\cos(m\theta + \phi_m)$ with $S \subset \mathbb{Z}_{>0}$, $w_m \geq 0$ and $\sum_m w_m = 1$, zero-initialized population flow has exact squared error

$$E_t(B) = \sum_{m\in S} w_m e^{-2t\lambda_m(B)}. \tag{17}$$

This gives a fixed-budget design criterion: choose M base frequencies to minimize E_t at a specified horizon. The retention score $J_\eta(B) = \sum_{m\in S} w_m \lambda_m/(\lambda_m + \eta)$ measures retained amplitude weighted by target energy. Squared ridge error instead weights each mode by $(\eta/(\lambda_m + \eta))^2$; both expressions are derived in Appendix C.1.

Proposition 6 (Objective certificates) *For $t \geq 0$, $\eta > 0$, a finite candidate family, and nonnegative target weights with sum W, define $J_{\eta,L}$ and $E_{t,L}$ by replacing λ_m with $\lambda_{m,L}$. Uniformly over the family,*

$$0 \leq J_\eta - J_{\eta,L} \leq W R_L/\eta, \qquad 0 \leq E_{t,L} - E_t \leq 2tW R_L. \tag{18}$$

A maximizer of $J_{\eta,L}$ has true-score regret at most WR_L/η. If its truncated score exceeds every other candidate by more than that bound, it is also the unique maximizer of J_η. The analogous assertions hold for minimizing E_t with bound $2tWR_L$.

Proof: Appendix C.2.

Signed convolution evaluates a candidate by iterating $p_0 = \delta_0$, $p_{\ell+1} = p_\ell * \mu_B$ and accumulating $a_\ell p_\ell$. In one dimension, dense recursion costs $O(ML^2\Omega)$ arithmetic operations and $O(L\Omega)$ working memory; in higher dimensions the support can grow rapidly. An alias-free FFT of the truncated composed kernel provides an alternative, independently checkable calculation. Exhaustive search costs additionally $\binom{C}{M}$ candidate evaluations for C admissible frequencies. We use it as a small-budget reference, and evaluate a cheaper greedy construction with one-frequency exchanges as a practical comparator.

These learning formulas use uniform periodic measure and integer frequencies. On a finite grid, the spectrum follows Equation (8); nonuniform samples generally

couple Fourier modes. For finite networks, the empirical kernel evolves during training. Appendix C.3 gives a perturbation bound for comparison with a fixed kernel and a stability bound for changes in feature amplitudes.

6 Experiments

6.1 Protocol and independent checks

All experiments are deterministic given the recorded seeds and use double-precision (float64) arithmetic on a central processing unit (CPU). The finite model is Equation (11), trained in both layers by full-batch gradient descent on N equispaced points with loss $(2N)^{-1}\sum_i (f(\theta_i) - y_i)^2$. We use the centered output $\widetilde{f}_\vartheta = f_\vartheta - f_{\vartheta_0}$, with the initial function frozen and subtracted at training and test locations. This preserves the tangent kernel and gives the same initial residual for every dictionary. For the one-dimensional experiments, a time t denotes optimizer step size times iteration count. Their reported *decay rates* are fitted slopes of log absolute cosine residual over $0 \le t \le 10$, distinct from the optimizer step size.

The one-dimensional targets are analytically constructed periodic signals. Their test mean squared error (MSE), evaluated at 1,024 offset locations unseen during fitting, estimates integrated reconstruction error for these noiseless targets. Section 6.6 evaluates natural-image reconstruction on the fitted pixels. Standard deviations (SDs) summarize seed variability within each condition, with initialization seeds paired across dictionary rules where applicable. Full target, initialization, fitting, grid, and search specifications are in Appendix D; the image protocol is in Appendix F.

Table 1 gives the training configuration and seed counts for the 846 one-dimensional runs. A further 50 runs evaluate natural-image reconstruction in Section 6.6.

Table 1 One-dimensional training configurations. The 72 additional initial-kernel measurements used for width convergence involve no training; the 50 natural-image runs are specified in Section 6.6

Training experiment	Width	Grid N	Seeds per case	Steps / step size	Runs
Matched pair	256, 1,024	256	8	400 / 0.2	32
Support restrictions	512	256	5	200 / 0.2	20
Four-task design	512	256	8	200 / 0.1	224
Cross-dictionary prediction	512	128	5	100 / 0.1	540
Grid sensitivity	512	128, 256, 512	5	100 / 0.1	30
Total					846

Figure 1 summarizes the independent spectral and truncation checks. For $B = \{(1,0),(0,1),(1,2)\}$ and Ψ_q with $q = 0.55$, a 256^2 grid is alias-free at order $L = 24$. Signed convolution and direct FFT agree to floating-point precision, with maximum discrepancy 5.55×10^{-17}. The omitted trace is $R_{24} = 3.23 \times 10^{-7}$, and the truncated

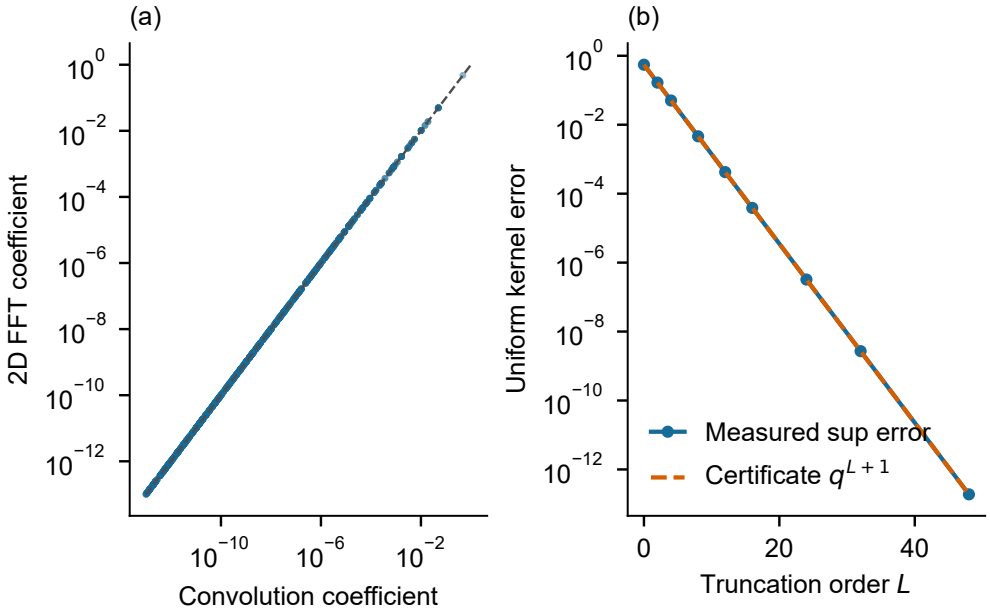


Fig. 1 Independent spectral and truncation checks. (a) All order-24 coefficients exceeding 10^{-13} for a two-dimensional dictionary: signed-convolution probabilities versus a direct FFT of the polynomial kernel. (b) Measured kernel sup error and its certificate, including the maximizer $t = 0$. The overlap is predicted by Theorem 2

covariance agrees with $\Sigma_B \sum_{\ell=0}^{24} \ell a_\ell$. Finite differences check the analytic network gradients, and an explicit Jacobian product checks the empirical Gram matrix. Dense linear solves validate ridge amplitudes and expose the continuum/grid distinction in Appendix E.

6.2 Matched scales, distinct local dynamics

For the geometric profile, the mode-9 eigenvalues are 0.0190638 and 0.00127517, a ratio of 14.95. For the ReLU NTK, they are 0.0615497 and 0.00315501, a ratio of 19.51. The network predictions in Figure 2 use the ReLU eigenvalues. Increasing the spectral quadrature grid from 65,536 to 131,072 checks their numerical stability.

We train eight seeds at each width $n \in \{256, 1024\}$ using $N = 256$, step size 0.2, and 400 steps. At width 1,024, fitted early decay rates are 0.06092 ± 0.00212 and 0.00315 ± 0.00015. The empirical NTK's relative Frobenius drift by $t = 80$ averages 1.40% and 1.53%, respectively. An independent eigendecomposition provides frozen empirical-kernel trajectories. At Fourier mode 1,000, both $m^2\lambda_m$ values are close to 0.7514, consistent with Theorem 5.

A separate initial-kernel width diagnostic for $B = (1, 4)$ averages relative Rayleigh errors over seven Fourier modes and 12 seeds per width. The mean error against the population spectrum decreases from approximately 12.20% at width 64 to 2.10% at

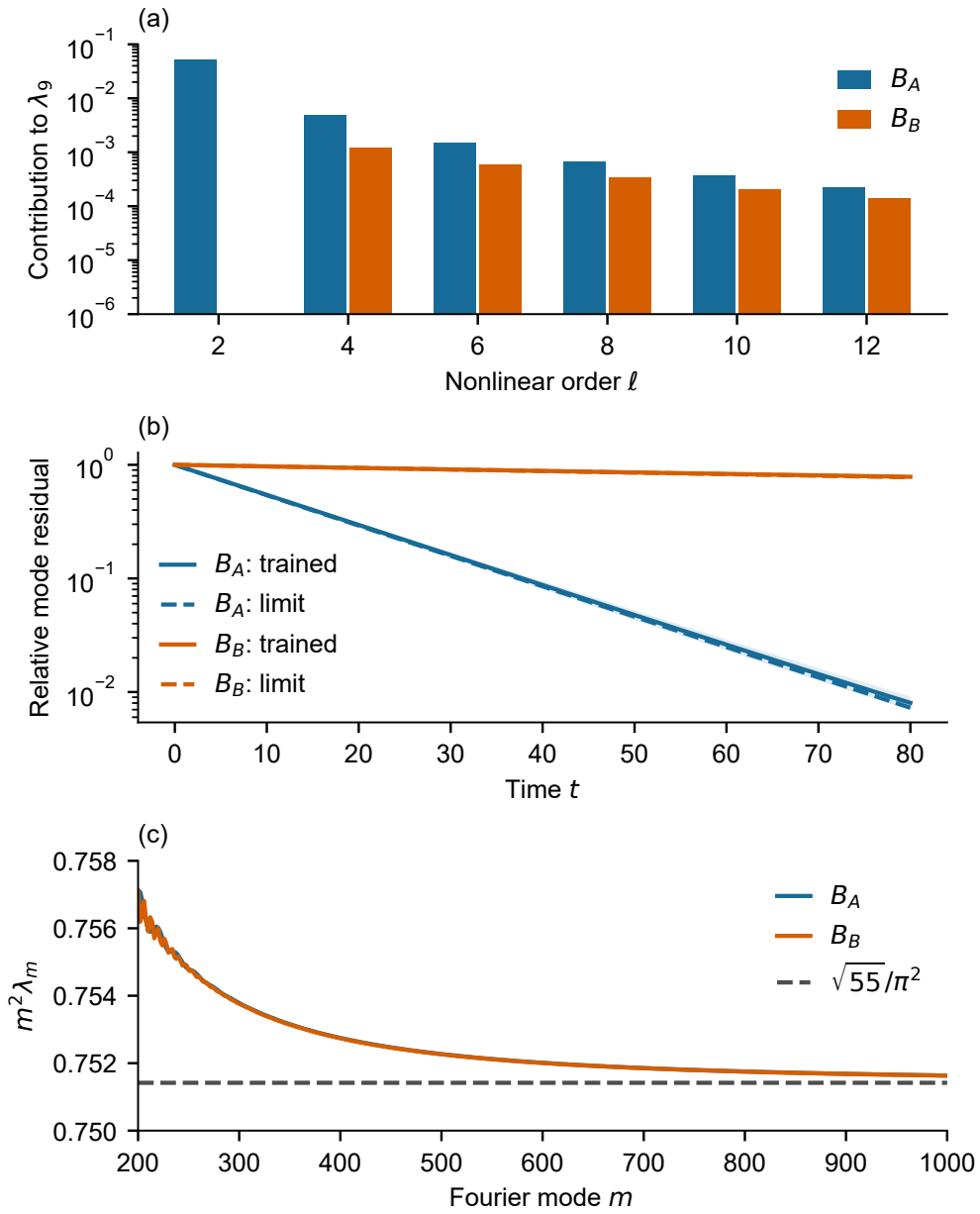


Fig. 2 Matched dictionaries $B_A = (1, 8, 10)$ and $B_B = (4, 7, 10)$. (a) ReLU order contributions to mode 9; the absent order-two contribution of B_B lies below the plotted range. (b) Centered width-1,024 networks, mean $\pm$ one SD over eight seeds; dashed lines are population NTK predictions. (c) For $m \geq 200$, scaled coefficients approach the common tail constant $\sqrt{55}/\pi^2$. The local contrast survives agreement of the leading asymptotic tail

width 2,048. These directional errors and the terminal Frobenius drift quantify finite-width agreement, but do not establish a uniform operator-norm bound along training. Appendix E gives the full width and drift curves.

6.3 Hard lattice and active-degree restrictions

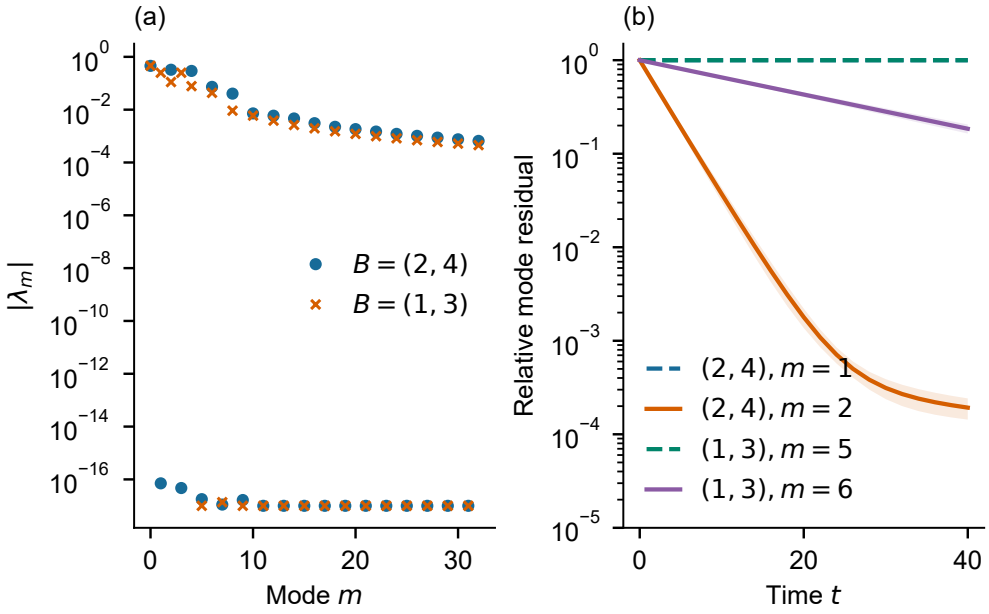


Fig. 3 Two distinct support restrictions. For $B = (2, 4)$, odd frequencies lie outside the generated lattice. For $B = (1, 3)$, mode 5 is in the generated lattice but excluded by the ReLU active degrees. Eigenvalues are floored at 10^{-17} for display. Training curves use five seeds, width 512 and $N = 256$; the two forbidden-mode curves overlap at one

Figure 3 separates the two restrictions. The normalized residual stays at one for both $B = (2, 4), m = 1$ and $B = (1, 3), m = 5$, while the reachable controls $m = 2$ and $m = 6$ decay. For $B = (1, 3)$, shifting θ by π negates $\gamma_B(\theta)$. Since $\mathrm{ReLU}(z) - \mathrm{ReLU}(-z) = z$, the odd component of every finite network is a linear combination of the base features. Mode 5 is therefore absent at every width, giving an exact representational explanation for its constant residual.

6.4 Prospective dictionary design and finite-network transfer

The target set is $S = \{3, 4, 6, 7, 8, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25, 27\}$. We construct four target energy patterns: mixed, flat, low-weighted ($w_m \propto m^{-1/2}$), and high-weighted ($w_m \propto m^{1/2}$), each with unit population energy. We select four distinct base frequencies from $\{1, \ldots, 27\}$ by minimizing the ReLU objective E_{20}. All 17,550 candidates are evaluated on grids of size 4,096 and 8,192; the selected dictionary is unchanged for every task. The leading 32 candidates are refined at 65,536 points.

This ReLU search is evaluated numerically; the rigorous truncation certificate is tested separately with the geometric profile.

We compare direct top-energy insertion, a greedy construction followed by improving one-frequency exchanges, low-frequency and spread dictionaries, uniform random subsets, and a target-tuned Gaussian-scale baseline. For the latter, five Gaussian scales are scored using 64 pilot dictionaries per scale. The scale with the best mean predicted objective supplies eight fresh dictionaries for network evaluation. Gaussian draws are quantized and clipped to the same integer range, with repetitions retained, applying bandwidth tuning within the periodic fixed-budget setting. Flat-spectrum top-energy ties are broken by uniform random selection for each seed. Search costs and baseline definitions appear in Appendix D.

Table 2 Off-grid MSE at $t = 20$, mean $\pm$ one SD over eight seeds. Every network has width 512 and four Fourier entries and uses 200 steps of size 0.1 on $N = 256$ training points. For random dictionary rules, variability includes dictionary draws. Top-energy ties in the flat task are randomized. Lower is better

Dictionary rule	Mixed	Flat	Low-weighted	High-weighted
Exhaustive spectral	0.525 ± 0.004	0.521 ± 0.004	0.458 ± 0.012	0.549 ± 0.006
Greedy + swap	0.529 ± 0.010	0.532 ± 0.006	0.474 ± 0.011	0.549 ± 0.009
Top-energy	0.599 ± 0.005	0.643 ± 0.053	0.529 ± 0.005	0.705 ± 0.002
Gaussian, scale-tuned	0.732 ± 0.099	0.728 ± 0.094	0.675 ± 0.086	0.749 ± 0.081
Uniform random	0.669 ± 0.050	0.668 ± 0.048	0.660 ± 0.063	0.676 ± 0.045
Low-frequency	0.809 ± 0.004	0.782 ± 0.004	0.676 ± 0.005	0.856 ± 0.004
Spread	0.785 ± 0.007	0.786 ± 0.006	0.815 ± 0.006	0.763 ± 0.006

Figure 4 compares training and off-grid errors for the prospective dictionary designs. The exhaustive designs are $(6, 10, 13, 21)$ for mixed and flat targets, $(3, 4, 10, 15)$ for the low-weighted target, and $(3, 7, 18, 20)$ for the high-weighted target. On the mixed task, mean off-grid MSE is 12.3% lower than top-energy insertion and 28.2% lower than the Gaussian-scale baseline. Table 2 reports all methods and tasks. The greedy alternative uses 195–287 spectral evaluations and nearly matches the exhaustive result on the high-weighted task. The comparison demonstrates transfer of the spectral criterion to finite-network training and the benefit of a cheaper search at this budget.

For the geometric-profile retention score in Proposition 6, $q = 0.55$, $L = 24$, and $\eta = 0.04$ select $(7, 8, 11, 12)$. Its score gap to the runner-up is 0.02343, exceeding the one-sided objective error bound 1.31×10^{-4}. This certifies a unique optimum for the full geometric profile over the specified 17,550-element family.

6.5 Predictivity across dictionaries and targets

Figure 5 compares population eigenvalues with fitted early decay rates across dictionaries. For each target $m \in \{5, 9, 17\}$, we enumerate $B = (a, b, 12)$ with $1 \le a < b < 12$, $\gcd(B) = 1$, and $m \notin B$. We train five seeds per dictionary at width 512, $N = 128$, step

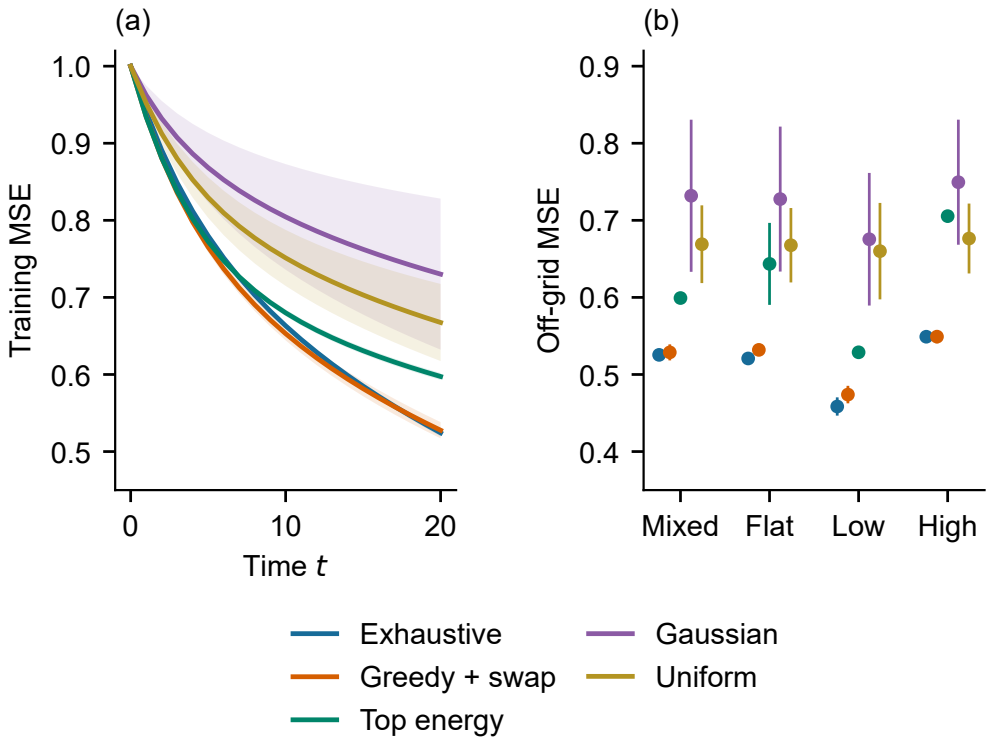


Fig. 4 Network validation of prospective design. (a) Training MSE for the mixed-spectrum target, with one-SD bands. (b) Off-grid MSE for the same five rules across four target energy patterns. Error bars are one SD, and colors match panel (a). All dictionaries are selected before evaluating their finite-network training results

size 0.1, for 100 steps. The local eigenvalue has Spearman correlations 0.981, 0.992, and 0.977 with mean fitted rates. RMS base frequency has correlations approximately -0.386, -0.359, and -0.082. Table 3 reports the correlations, dictionary-bootstrap intervals, and absolute calibration errors. First-arrival depth and the order-two contribution are also informative; their comparisons and the interval construction details remain in Appendix E.

The panels contain 108 dictionary–target pairs and 540 networks. Bootstrap resampling reweights dictionaries within each enumerated panel; the intervals describe this fixed family rather than uncertainty over an independently sampled dictionary population. Replacing population eigenvalues by the grid eigenvalues from Equation (8) reduces the median relative calibration errors from 3.07%, 5.25%, and 8.90% to 1.73%, 1.51%, and 1.90%, respectively. Thus strong rank prediction coexists with a grid-dependent correction to absolute decay rates. Online Resource 1 also contains matched-seed grid checks and finite-width Rayleigh measurements.

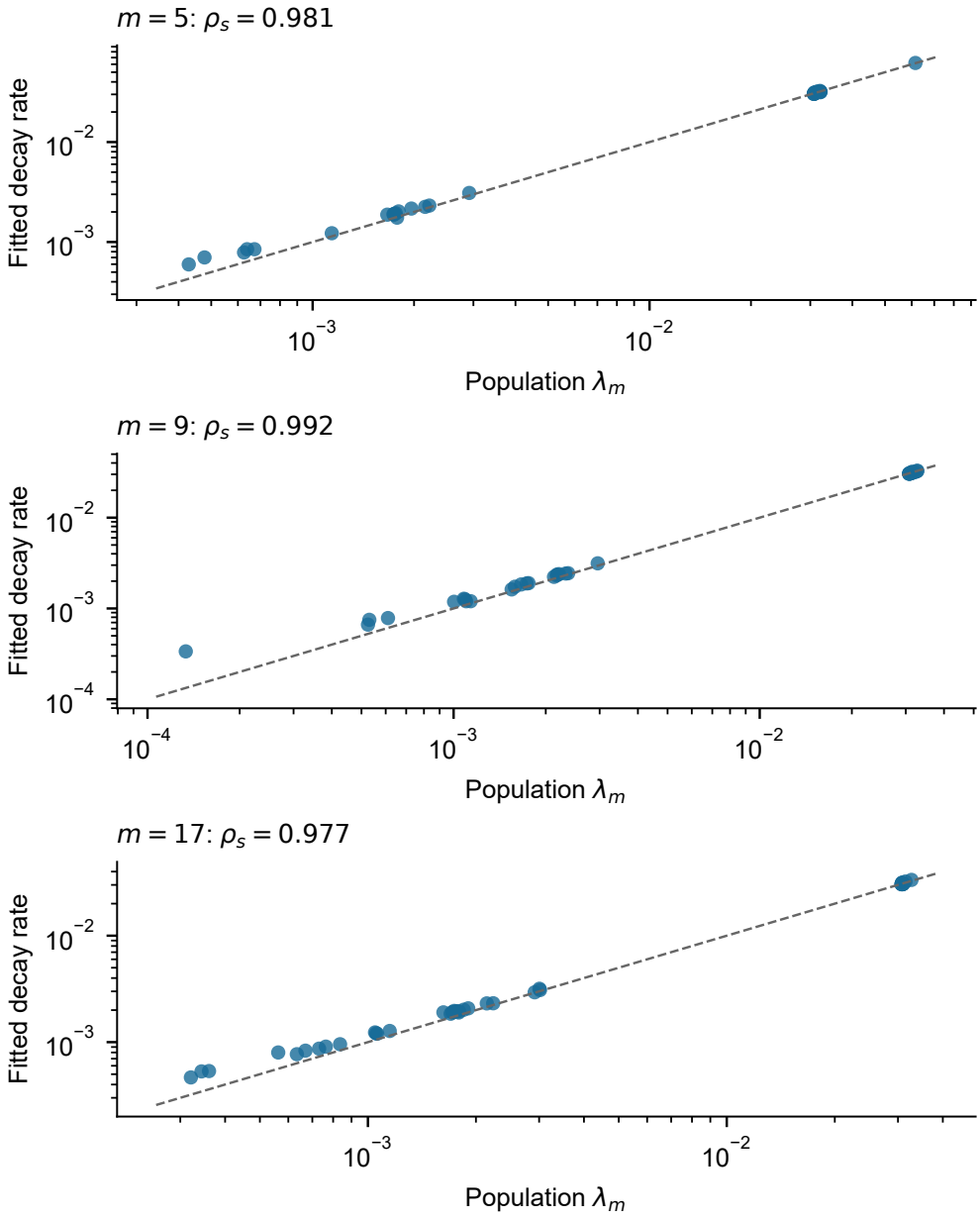


Fig. 5 Cross-dictionary prediction of early finite-network decay. Each point is a dictionary, averaging five width-512 seed runs; the dashed diagonal marks equality with the population eigenvalue. Targets $m = 5, 9, 17$ use 32, 34, and 42 dictionaries, respectively, all with cardinality three and maximum frequency 12. Small-rate deviations include finite-grid aliasing and finite width

Table 3 Cross-dictionary prediction statistics, using mean fitted rates over five seeds per dictionary. Spectrum intervals are percentile intervals from 2,000 dictionary-bootstrap resamples. Calibration error is $|\widehat{r} - \lambda|/\lambda$, using population or $N = 128$ grid eigenvalues as indicated

Statistic	$m = 5$	$m = 9$	$m = 17$
Number of dictionaries	32	34	42
Spectrum: Spearman ρ	0.981	0.992	0.977
Spectrum: 95% bootstrap interval	[0.936,0.995]	[0.968,0.999]	[0.925,0.994]
RMS base frequency: ρ	-0.386	-0.359	-0.082
Median relative error: population	3.07%	5.25%	8.90%
Median relative error: grid	1.73%	1.51%	1.90%

6.6 Natural-image coordinate regression

We test the allocation predictions on the first five lexicographically ordered BSDS500 test images [22]. Each image supplies a 64×64 center crop, converted to grayscale, centered, and normalized to unit mean-square energy. All 4,096 pixels are used for fitting and reconstruction-error evaluation. Each crop is treated as a discrete periodic signal, including its boundary discontinuities.

We fix two four-entry dictionaries before training:

$$B_A = \{(1,0),(8,0),(10,0),(0,1)\}, \qquad B_B = \{(4,0),(7,0),(10,0),(0,1)\}.$$

The first coordinate denotes the image row. Both dictionaries have maximum frequency 10 and covariance $\Sigma_B = \mathrm{diag}(165/4, 1/4)$. Their exact 64×64 grid eigenvalues at $(9,0)$ are 0.03341285 and 0.000911895, a ratio of 36.64. Using the discrete-grid spectrum in Equation (8) and the unitary target discrete Fourier transform (DFT) $\widehat{y}_m$, the zero-initial-output kernel prediction after s steps is

$$E_s^{[64]}(B) = \frac{1}{4096} \sum_m |\widehat{y}_m|^2 (1 - \alpha \lambda_m^{[64]}(B))^{2s}, \qquad \alpha = 0.1. \tag{19}$$

We train centered width-1,024 networks in both layers for 200 full-batch steps, using five paired initialization seeds per image and dictionary (50 runs).

Figure 6 summarizes the natural-image comparisons. At step 200, observed normalized MSEs are 0.6321 ± 0.0010 for B_A and 0.7818 ± 0.0004 for B_B, compared with grid predictions 0.6329 and 0.7814. Within image–dictionary conditions, rank correlations between predicted and seed-averaged fitted mode rates range from 0.761 to 0.976. Absolute rate calibration is weaker: median relative errors are 25.8% and 24.3%, and 9 and 10 of the 100 seed-averaged rates per dictionary are negative. Close aggregate MSE agreement thus coexists with finite-width deviations in individual mode dynamics. Appendix F gives all per-image results and the complete rate-fitting protocol.

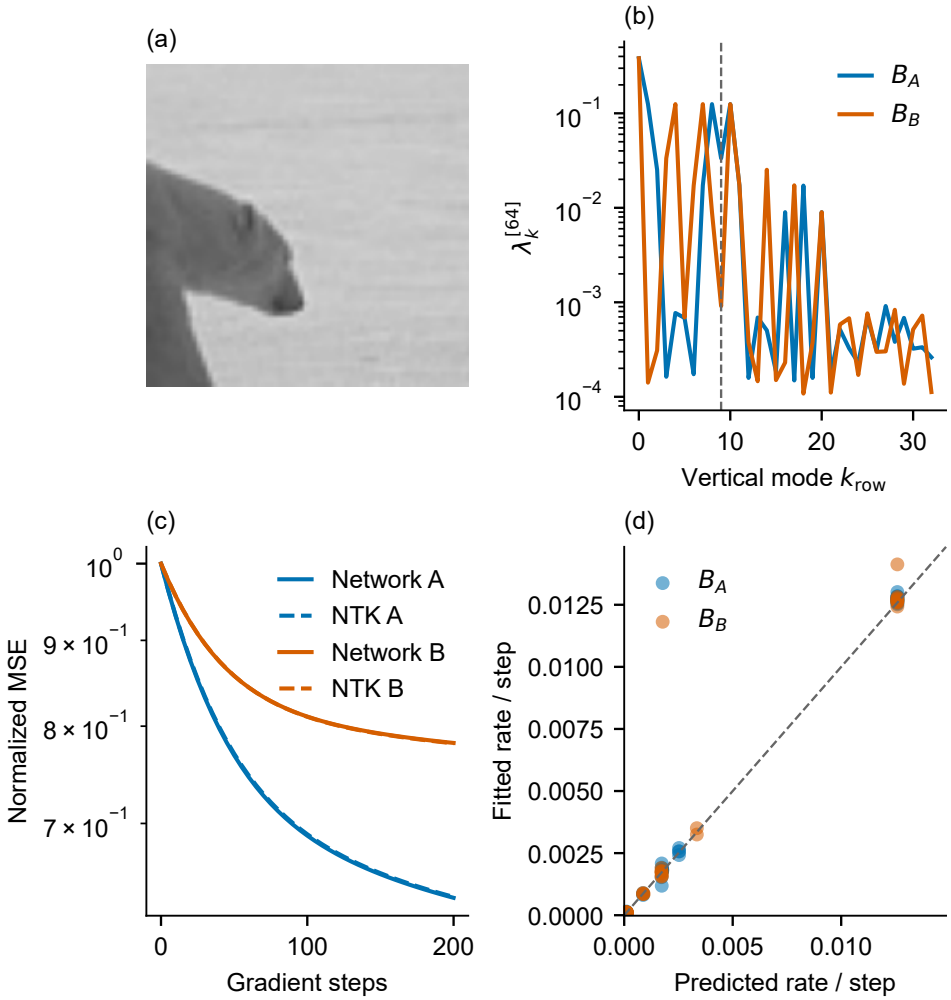


Fig. 6 Natural-image validation on five fixed BSDS500 crops. (a) The first crop in lexicographic image identifier order. (b) Exact grid eigenvalues along the vertical frequency axis. (c) MSE averaged over five images and five seeds; bands are one SD across seed-level image means, and dashed curves are infinite-width grid predictions. (d) Predicted versus fitted amplitude-decay rates over steps 0–50 for the 20 highest-energy non-DC conjugate pairs per image. Each of the 200 points averages five seeds; negative fitted rates are retained

7 Discussion and conclusion

Exact spectral allocation links harmonic synthesis to kernel learning rates. Ordered signed paths distribute each nonlinear-order coefficient across frequencies, and their summed contributions determine convergence and regularized retention. The ReLU calculation identifies frequencies excluded by active degrees and shows that identical leading spectral tails can coexist with nearly twenty-fold local eigenvalue differences.

The finite-network experiments cover noiseless full-batch training of centered two-layer networks with fixed integer frequencies. The five natural-image crops extend the target class while retaining a discrete periodic domain and reconstruction on fitted pixels. Larger image collections and unseen-pixel evaluation would be needed to assess broader image-representation performance. Learned frequencies, deep feature learning, nonperiodic domains, and nonuniform sampling introduce evolving or non-diagonal Fourier operators that require further analysis. Exhaustive design also has a combinatorial candidate cost; the greedy baseline demonstrates the resulting tradeoff between search cost and reconstruction error.

Within this setting, the allocation formula, ReLU support and tail analysis, and truncation certificates provide a quantitative description of an individual Fourier dictionary. The finite-network experiments show how this description can guide dictionary selection and explain local learning differences that bandwidth and harmonic support alone leave unresolved.

Supplementary information. Online Resource 1 contains the experiment and analysis code, dependency specifications, synthetic-target and natural-image numerical records, dataset provenance, and instructions for reproducing all ten figures and the reported statistics.

Statements and Declarations

Funding. The authors did not receive support from any organization for the submitted work.

Conflict of interest. The authors declare no competing interests.

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. The BSDS500 dataset analysed during the current study is publicly available at https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/resources.html. All other data generated during this study are included in this published article and its supplementary information files.

Materials availability. Not applicable.

Code availability. The code is available in Online Resource 1.

Author contribution. B.Z. and W.J. contributed to the study conception and design. B.Z. conceived the idea and performed the experiments and analysis. B.Z. and W.J. wrote the first draft of the manuscript. All authors read and approved the final manuscript.

Appendix A Proofs of the allocation and approximation results

A.1 Exact allocation, trace, and first-arrival bounds

Proof of Theorem 1 Since $|g_B(t)| \le 1$, the Weierstrass test gives absolute uniform convergence of $\sum_\ell a_\ell g_B(t)^\ell$. Euler's identity yields

$$g_B(t)^\ell = (2M)^{-\ell} \sum_{(j_r,s_r)_{r=1}^\ell} e^{i(\sum_r s_r \omega_{j_r})^\mathsf{T} t} = \sum_m \mu_B^{*\ell}(m) e^{im^\mathsf{T} t}.$$

For every ℓ, $\sum_m \mu_B^{*\ell}(m) = 1$. Hence $\sum_{\ell,m} a_\ell \mu_B^{*\ell}(m) = A < \infty$, so the double Fourier expansion converges absolutely and can be integrated termwise. Translation invariance of Haar measure gives

$$T_B \varphi_m(\theta) = e^{im^\mathsf{T}\theta} \int \kappa_B(t) e^{-im^\mathsf{T} t} \, \mathrm{d}\rho(t) = \lambda_m \varphi_m(\theta),$$

and coefficient extraction gives Equation (3). The tail is bounded by R_L uniformly in m. Nonnegative summands give the support equivalence, and Tonelli gives $\sum_m \lambda_m = A$. The complete Fourier basis therefore diagonalizes a bounded nonnegative trace-class operator. Its real even kernel also shows self-adjointness. Reflection of every step is a bijection between paths to m and $-m$, yielding spectral symmetry. □

For signed coefficients with $\sum_\ell |a_\ell| < \infty$, the same argument applies after replacing a_ℓ by $|a_\ell|$ in the absolute-convergence estimates. Support equivalence and the nonnegative tail identities require the sign assumption used in the paper.

For Equation (5), every active contribution below $r(m)$ is zero; retaining the first positive term gives the lower bound and using $\mu_B^{*\ell}(m) \le 1$ gives the upper bound. Each walk endpoint belongs to G_B and has infinity norm at most $\ell\Omega$. This proves both the generated-group restriction and the coefficient-tail frequency bound. The exact truncated rank is $|\bigcup_{\ell \le L : a_\ell > 0} \operatorname{supp}(\mu_B^{*\ell})|$, often smaller than the enclosing-box estimate.

A.2 Truncation and grid aliasing

Proof of Theorem 2 Every retained endpoint satisfies $\|m\|_\infty \le L\Omega$, proving the rank bound. Pointwise, $|\kappa_B - \Psi_L(g_B)| \le R_L$; at $t = 0$, $g_B(0) = 1$ and the omitted sum equals R_L, proving equality of the sup norm. The two operators have the same Fourier eigenbasis and nonnegative eigenvalue differences, so their operator-norm difference is the supremum of the differences. Summing those differences gives

$$\sum_m (\lambda_m - \lambda_{m,L}) = \sum_{\ell > L} a_\ell \sum_m \mu_B^{*\ell}(m) = R_L.$$

This is exactly the trace norm of the positive difference.

For the grid $t_j = 2\pi j/N$, $j \in \{0, \dots, N-1\}^d$, a character vector is an eigenvector of the matrix K/N^d. Its eigenvalue is $N^{-d} \sum_j \kappa_B(t_j) e^{-im^\mathsf{T} t_j}$. Substitute the absolutely summable Fourier series and apply discrete character orthogonality:

$$N^{-d} \sum_j e^{i(r-m)^\mathsf{T} t_j} = \mathbf{1}\{r - m \in N\mathbb{Z}^d\}.$$

This gives Equation (8). Every nonzero alias $m + Nk$ has infinity norm at least $N - \|m\|_\infty$. Consequently all aliases require walk order at least $L_m + 1$, and

$$\sum_{k\neq 0} \lambda_{m+Nk} = \sum_{\ell \geq L_m+1} a_\ell \sum_{k\neq 0} \mu_B^{*\ell}(m + Nk) \leq \sum_{\ell \geq L_m+1} a_\ell = R_{L_m}.$$

The inner sum is at most one, which controls all aliases jointly. Finally, if $N > 2L\Omega$, no two distinct lattice points in the support box are congruent modulo N in every coordinate. The polynomial coefficients are then recovered exactly by the grid transform. □

A.3 Global spectral moments

Proof of Proposition 3 Let $X \sim \mu_B$. Then $\mathbb{E}X = 0$ and $\mathbb{E}XX^\mathsf{T} = \Sigma_B$. For an independent length-ℓ walk S_ℓ, $\mathbb{E}S_\ell S_\ell^\mathsf{T} = \ell\Sigma_B$. Fix $v \in \mathbb{R}^d$. Tonelli, applied to nonnegative terms, gives

$$\sum_m (v^\mathsf{T} m)^2 \lambda_m = \sum_\ell a_\ell \mathbb{E}(v^\mathsf{T} S_\ell)^2 = \left(\sum_\ell \ell a_\ell\right) v^\mathsf{T}\Sigma_B v.$$

For $v^\mathsf{T}\Sigma_B v > 0$, this identity is valid as an extended nonnegative equality. If $v^\mathsf{T}\Sigma_B v = 0$, then $v^\mathsf{T} X = 0$ almost surely and the directional spectral moment is zero, regardless of the order moment; this case is treated separately to avoid an undefined product $0 \cdot \infty$. Under a finite first order moment, all entries of the matrix sum are absolutely summable by $|m_i m_j| \leq (m_i^2 + m_j^2)/2$. Polarization gives Equation (10). If the scalar order moment diverges, the same equality proves divergence in every nonzero-variance direction. □

Appendix B ReLU profile, support, and leading tail

B.1 Network normalization and coefficient expansion

For unit vectors x, x' with $x^\mathsf{T} x' = c$, let $U = w^\mathsf{T} x$, $V = w^\mathsf{T} x'$. These are correlated standard Gaussians. Differentiating the model in Equation (11), the output-weight and hidden-weight contributions to its limiting NTK are

$$2\mathbb{E}[U_+ V_+] = \frac{\sqrt{1-c^2} + c(\pi - \arccos c)}{\pi},$$
$$2\mathbb{E}[v^2] c \Pr(U > 0, V > 0) = \frac{c(\pi - \arccos c)}{\pi}.$$

Their sum is Equation (12). The Gaussian identities can be obtained by integrating in polar coordinates in the span of x, x'. At $c = 1$, each contribution equals one, fixing the trace normalization at two.

Write $C_k = 4^{-k}\binom{2k}{k}$. The expansions

$$\sqrt{1-c^2} = 1 - \sum_{k\geq 1} \frac{C_k}{2k-1} c^{2k},$$
$$\arcsin c = \sum_{k\geq 0} \frac{C_k}{2k+1} c^{2k+1}, \qquad \pi - \arccos c = \frac{\pi}{2} + \arcsin c.$$

give $a_0 = 1/\pi$, $a_1 = 1$, and zero higher odd coefficients. Since $C_{k-1} = 2kC_k/(2k-1)$, the coefficient of c^{2k} is

$$\frac{1}{\pi}\left[-\frac{C_k}{2k-1} + \frac{2C_{k-1}}{2k-1}\right] = \frac{(2k+1)C_k}{\pi(2k-1)^2}.$$

Stirling's formula gives $C_k \sim (\pi k)^{-1/2}$, yielding the stated coefficient asymptotic, absolute summability, and divergence of $\sum_\ell \ell a_\ell$. Equality of the series with Equation (12) extends to the endpoints by uniform convergence and continuity.

Proof of Theorem 4 Let $\mathcal{E}_B$ denote the union of all even-length endpoint sets, including zero. Concatenation adds endpoints and reversing every sign negates an endpoint; therefore $\mathcal{E}_B$ is a subgroup. It contains $2\omega_1$ and each $\omega_j - \omega_1$, which are length-two endpoints, so $H_B \subseteq \mathcal{E}_B$. Conversely, modulo H_B, every ω_j equals ω_1 and $-\omega_1 = \omega_1$, because $2\omega_1 = 0$ in the quotient. Every even signed sum is consequently zero modulo H_B. Thus $\mathcal{E}_B = H_B$.

The quotient G_B/H_B is generated by the class of ω_1, whose order is at most two. This proves the index bound. Every even degree, as well as degree one, has a strictly positive ReLU coefficient; all other degrees vanish. The support equivalence of Theorem 1 therefore gives precisely H_B together with the one-step endpoints. □

This result is also valid for repeated dictionary entries and signed base vectors. The choice of the first entry in Equation (14) does not affect the resulting subgroup because it equals the intrinsic even-walk endpoint set.

B.2 Proof of the one-dimensional tail

Proof of Theorem 5 Write $b_j = \omega_j/h$, so $\gcd(|b_1|, \ldots, |b_M|) = 1$. First, $g_B(t) = 1$ exactly at $t_r = 2\pi r/h$, $0 \le r < h$: an average of cosines equals one only if each cosine equals one, and the simultaneous conditions follow from Bézout's identity. Similarly, $g_B(t) = -1$ requires $\omega_j t \equiv \pi \pmod{2\pi}$ for all j. Such a point exists exactly when every b_j is odd; its locations are $s_r = (2r+1)\pi/h$.

To see the necessity of the odd condition directly, let $u = ht/\pi$ so $b_j u$ is an odd integer for every j. A Bézout combination of the b_j shows that u is an integer. Then u and every b_j must be odd. The converse is immediate by taking $u = 1$.

At a point t_0 with $g_B(t_0) = 1$, Taylor expansion gives

$$g_B(t_0 + s) = 1 - \tfrac{1}{2}\sigma_B^2 s^2 + O(s^4).$$

At a point with $g_B(t_0) = -1$, it gives $g_B(t_0+s) = -1 + \frac{1}{2}\sigma_B^2 s^2 + O(s^4)$. The scalar endpoint expansions are

$$\Psi_R(1-\varepsilon) = 2 - \frac{\sqrt{2\varepsilon}}{\pi} - 2\varepsilon + O(\varepsilon^{3/2}), \qquad \Psi_R(-1+\varepsilon) = -\frac{\sqrt{2\varepsilon}}{\pi} + O(\varepsilon^{3/2}).$$

Thus at either type of singular point, $\kappa_B(t_0 + s)$ has the cusp term $-\sigma_B|s|/\pi$. Its first derivative has jump $-2\sigma_B/\pi$. To justify differentiating the remainder, write $1 \mp g_B(t_0+s) = s^2 q(s^2)$ locally, where q is analytic and $q(0) = \sigma_B^2/2 > 0$. The convergent endpoint expansions of Ψ_R are analytic series in $\sqrt{1 \mp c}$. Their composition is therefore a sum of an analytic function of s^2 and $|s|$ times an analytic function of s^2. After subtracting the leading cusp, the latter starts at $|s|^3$, whose second derivative is locally integrable. Away from these finitely

many points the kernel is smooth. Hence its second distributional derivative has an integrable regular part. In distribution notation,

$$\kappa_B'' = r(t) - \frac{2\sigma_B}{\pi}\left(\sum_{r=0}^{h-1}\delta_{t_r} + \chi_B\sum_{r=0}^{h-1}\delta_{s_r}\right), \qquad r \in L^1([0,2\pi]).$$

Taking normalized Fourier coefficients and applying the Riemann–Lebesgue lemma to r yields

$$m^2\lambda_m = \frac{\sigma_B}{\pi^2}\left(\sum_{r=0}^{h-1} e^{-imt_r} + \chi_B\sum_{r=0}^{h-1} e^{-ims_r}\right) + o(1).$$

The first finite sum equals h for $h \mid m$ and zero otherwise. The second is $h(-1)^{m/h}$ when $h \mid m$. The exact vanishing off $h\mathbb{Z}$ already follows from generated-group support. Combining these facts proves Equation (16); exact vanishing of the unsupported odd multiples also follows from Theorem 4. □

The asymptotic is for a fixed dictionary. Its numerical illustration uses even m from 32 to 1,000, plots the range 200–1,000, and compares quadrature sizes 2^{19} and 2^{20}.

Appendix C Learning, design certificates, and perturbations

C.1 Population learning and RKHS consequences

Expand $e_t = f^\star - f_t = \sum_m e_m(t)\varphi_m$. From $\dot{e}_t = -T_B e_t$, boundedness and diagonalization give $\dot{e}_m = -\lambda_m e_m$, hence $e_m(t) = e^{-t\lambda_m}e_m(0)$. A positive-eigenvalue mode reaches relative residual $\delta \in (0,1)$ at time $\log(1/\delta)/\lambda_m$; a null mode is constant.

The population ridge equation $(T_B + \eta I)f_\eta = T_B f^\star$ gives

$$\widehat{f}_\eta(m) = \frac{\lambda_m}{\lambda_m + \eta}\widehat{f^\star}(m), \qquad \|f_\eta - f^\star\|_2^2 = \sum_m \left(\frac{\eta}{\lambda_m+\eta}\right)^2 |\widehat{f^\star}(m)|^2.$$

For the real target in Equation (17), the two coefficients at m and $-m$ each have squared magnitude $w_m/2$, proving Equation (17) by Parseval. Its phases do not affect the population error.

The associated RKHS is

$$\mathcal{H}_B = \left\{ f = \sum_m c_m\varphi_m : c_m = 0 \text{ if } \lambda_m = 0, \quad \sum_{m:\lambda_m>0}\frac{|c_m|^2}{\lambda_m} < \infty \right\},$$

$$\|f\|_{\mathcal{H}_B}^2 = \sum_{m:\lambda_m>0}\frac{|c_m|^2}{\lambda_m}.$$

Indeed, $\sqrt{\lambda_m}\varphi_m$ supplies a feature basis, and $\sum_m \lambda_m < \infty$ ensures bounded point evaluation by Cauchy–Schwarz. The nullspace of T_B is the closed span of zero-eigenvalue characters. Finally, for $\mathcal{L}(\eta) = \{m : \lambda_m \geq \eta\}$,

$$|\mathcal{L}(\eta)| \leq A/\eta, \qquad \tfrac{1}{2}|\mathcal{L}(\eta)| \leq \mathcal{N}_B(\eta) := \sum_m \frac{\lambda_m}{\lambda_m+\eta} \leq A/\eta.$$

The first inequality sums η over the threshold set. For the second, each mode in that set contributes at least $1/2$, while every mode contributes at most λ_m/η.

C.2 Certified objective comparison

Proof of Proposition 6 For $x \geq 0$, $h_\eta(x) = x/(x+\eta)$ is increasing and has derivative at most $1/\eta$. Since $0 \leq \lambda_m - \lambda_{m,L} \leq R_L$, weighted summation gives the first bound in Equation (18). The map e^{-2tx} is decreasing with derivative magnitude at most $2t$, giving the second bound.

Let $\widehat{B}$ maximize $J_{\eta,L}$ and let $B^\star$ maximize J_η. With $\varepsilon = WR_L/\eta$,

$$J_\eta(B^\star) - J_\eta(\widehat{B}) \leq J_{\eta,L}(B^\star) + \varepsilon - J_{\eta,L}(\widehat{B}) \leq \varepsilon.$$

If the truncated gap exceeds ε, then for every other B, $J_\eta(B) \leq J_{\eta,L}(B) + \varepsilon < J_{\eta,L}(\widehat{B}) \leq J_\eta(\widehat{B})$. For minimizing E_t, use $E_{t,L} - \varepsilon_t \leq E_t \leq E_{t,L}$, with $\varepsilon_t = 2tWR_L$, and repeat the comparison with reversed signs. □

For squared ridge error the same derivative argument yields an error bound $2WR_L/\eta$, because $x \mapsto (\eta/(x+\eta))^2$ has derivative magnitude at most $2/\eta$. This is distinct from the retention-score bound used for the geometric certificate.

C.3 Two stability statements

Proposition 7 (Perturbed positive-kernel flow) *Let T be bounded, self-adjoint and nonnegative on a Hilbert space. Suppose A_t has the same properties, depends continuously on t, and $\|A_t - T\|_{\mathrm{op}} \leq \varepsilon$ on $[0,\tau]$. For $\dot{e}_t = -A_t e_t$ and $\bar{e}_t = e^{-tT} e_0$,*

$$\|e_t - \bar{e}_t\| \leq t\varepsilon\|e_0\|, \qquad 0 \leq t \leq \tau.$$

Proof Nonnegativity implies $\frac{\mathrm{d}}{\mathrm{d}t}\|e_t\|^2 = -2\langle A_t e_t, e_t\rangle \leq 0$. Variation of constants gives $e_t - \bar{e}_t = -\int_0^t e^{-(t-s)T}(A_s - T)e_s \,\mathrm{d}s$. The semigroup is a contraction, so the integral norm is at most $t\varepsilon\|e_0\|$. □

Applying this flow bound requires uniform operator-norm control throughout the trajectory. The experiments record terminal Frobenius drift as a diagnostic and compute discrete frozen-kernel predictions $(I - \alpha A_n^{\mathrm{init}})^k y$, where A_n^{init} is the initial empirical kernel matrix normalized by the number of training points.

Proposition 8 (Stability under feature-amplitude changes) *Fix B and two probability vectors p, p'. Replace g_B by $g_p(t) = \sum_j p_j \cos(\omega_j^{\mathsf{T}} t)$ and define $\lambda(p)$ through Theorem 1. With $\delta = \|p - p'\|_1/2$,*

$$\sum_m |\lambda_m(p) - \lambda_m(p')| \leq 2\big[\Psi(1) - \Psi(1-\delta)\big].$$

If $\Psi'(1) < \infty$, the right side is at most $\Psi'(1)\|p - p'\|_1$. For the ReLU profile it is $O(\sqrt{\delta})$ as $\delta \downarrow 0$.

Proof The signed step laws have total variation distance at most δ. Couple each pair of steps so they disagree with probability at most δ, and use independent couplings over the ℓ steps. Endpoint distributions then differ in ℓ^1 norm by at most $2[1-(1-\delta)^\ell]$. Summing with weights a_ℓ proves the first bound. Since $1-(1-\delta)^\ell \le \ell\delta$, the finite-moment bound follows. The ReLU endpoint expansion in Appendix B.2 gives the final assertion. □

The amplitude bound compares probability weights on the same fixed integer frequency set.

Appendix D Complete experimental specification

D.1 Numerical implementation and checks

The accompanying code implements all training and spectral routines using NumPy and SciPy in float64. Each worker uses one Basic Linear Algebra Subprograms (BLAS) thread, with four independent workers for the recorded run. Analytic gradients are checked against centered finite differences with step 10^{-6} on a small independently seeded network; the maximum discrepancy is about 2.34×10^{-10}. An explicit concatenated parameter Jacobian verifies the empirical normalized kernel to below 10^{-15}.

For training inputs X with rows x_i, preactivations $Z = XW^\top$, $H = \mathrm{ReLU}(Z)$, and $D_{ij} = v_j \mathbf{1}\{Z_{ij} > 0\}$, the empirical normalized kernel is

$$A_n = \frac{2}{nN}\left[HH^\top + (DD^\top) \odot (XX^\top)\right].$$

The initial and final matrices provide drift $\|A_n^{\mathrm{final}} - A_n^{\mathrm{init}}\|_F / \|A_n^{\mathrm{init}}\|_F$. For multi-mode design runs, we also record the off-diagonal fraction of the initial kernel in the selected cosine basis. A separate spectral decomposition of A_n^{init} gives frozen-kernel residuals. These are independent calculations from the nonlinear parameter updates.

The geometric convolution check uses all reachable two-dimensional coefficients through degree 24. Its kernel grid has $N = 256$ per axis, exceeding $2L\Omega = 96$. The truncation sweep uses $L = 0, 2, 4, 8, 12, 16, 24, 32, 48$ and includes the origin. Its finite-rank moment uses the truncated order moment $\sum_{\ell \le L} \ell a_\ell$, rather than the untruncated $\Psi_q'(1)$. The ReLU tail check uses $N = 2^{20}$ with 2^{19} as a resolution comparison.

Table 1 summarizes the one-dimensional training conditions; the following subsections provide the target, initialization, and search details.

D.2 Targets, phases, and initialization

The single-mode target is $\sqrt{2}\cos(m\theta)$, which has unit population squared norm. Multi-mode targets are $\sum_{m \in S} \sqrt{2w_m}\cos(m\theta + \phi_m)$ with the target set in the main text. Their phases are drawn once using NumPy's default generator with seed 100, uniformly in $[0, 2\pi)$, and are shared by all methods and four energy patterns. The raw phases are recorded in every relevant training record. The unnormalized mixed-task weights, in

increasing order of S, are

$$(0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 1.00, 0.90, 0.70, 0.80,\\ 0.95, 0.75, 0.85, 1.10, 0.80, 0.70, 1.20, 0.75, 0.70, 0.65).$$

They sum to 16.2 and are normalized for the ReLU error objective. Geometric retention-score certification uses these unnormalized weights, so its bound uses $W = 16.2$.

Pair seeds are 0–7; support seeds 0–4; network design seeds 1000–1007; cross-dictionary seeds 300–304; and grid-sensitivity seeds 400–404. Every run draws W then v (denoted by a in the implementation) from independent standard Gaussians using its recorded generator seed. The draw order and NumPy version are recorded for reproducibility. Competing dictionaries use the same array shapes and seeds to pair their initialization randomness.

All tests use $\theta_j = 2\pi(j + 0.37)/1024$, $j = 0, \ldots, 1023$, offset from the training grid. The frozen initial function is subtracted at these points as well as at training points. The cosine residual is $r_m(t) = [2N^{-1} \sum_j e_j(t) \cos(m\theta_j + \phi_m)]/[2N^{-1} \sum_j y_j \cos(m\theta_j + \phi_m)]$. Single-mode figures plot its mean and standard deviation. Rates fit $\log |r_m(t)|$ at recorded times up to 10, excluding values below 10^{-10}.

D.3 Design search and all baselines

The exhaustive ReLU search scores E_{20} for every four-element subset of $\{1, \ldots, 27\}$. Spectra are computed in batches by direct Fourier transforms of Equation (12). Comparing grids 4,096 and 8,192 changes the objective by at most 1.30×10^{-5} over candidates and tasks; the minimizers coincide. The top 32 candidates per task are refined at 65,536 points to assess numerical stability.

The greedy heuristic adds the best unused frequency at each of four stages, recomputing the normalized kernel for the current dictionary size. It then evaluates all single-frequency replacements and takes the best improving exchange until none improves the objective by more than 10^{-12}. Addition ties are resolved by increasing frequency and exchange ties lexicographically. The procedure uses 195 or 287 spectral evaluations, yielding $(4, 6, 12, 19)$, $(3, 4, 15, 21)$, $(3, 7, 11, 13)$, and $(8, 12, 15, 25)$ for mixed, flat, low, and high targets. Its spectral quadrature size is 8,192.

Top-energy uses the four largest w_m. For the flat spectrum all entries tie, so eight independent, uniformly selected four-element subsets of S are used, with seeds 30000–30007. The fixed low-frequency and spread dictionaries are $(1, 2, 3, 4)$ and $(1, 9, 18, 27)$. Uniform random selection samples four distinct integers from 1–27, with seeds 20000–20007.

The Gaussian-scale baseline samples four independent $Z_j \sim \mathcal{N}(0, \sigma^2)$ and sets $b_j = \min(27, \max(1, \text{round}\,|Z_j|))$. Repeated entries remain valid labeled features. For each $\sigma \in \{2, 4, 8, 16, 32\}$, 64 pilot draws with seeds 7000–7063 are scored by the same target-aware population-error approximation as the spectral designs. The best average scale is 16 for mixed, flat and high tasks and 8 for the low task. Network evaluation uses fresh dictionary seeds 10000–10007, with dispersion across the eight runs reflecting both dictionary and initialization variability.

The geometric certification uses the degree-24 polynomial, $q = 0.55$, $\eta = 0.04$, the same candidate family, and the unnormalized mixed weights. Grid size 2,048 exceeds $2L\Omega = 1,296$, so the truncated coefficients are alias-free. The winner is $(7, 8, 11, 12)$, the runner-up is $(4, 11, 12, 19)$, and the winner's score is 2.9734617763. Its gap 0.0234279286 exceeds $WR_{24}/\eta = 0.0001307743$, so Proposition 6 certifies the full-profile winner over this family. The bound controls mathematical truncation error; the implementation uses float64 rather than interval arithmetic.

Appendix E Additional measurements and uncertainty

E.1 Finite width and kernel drift

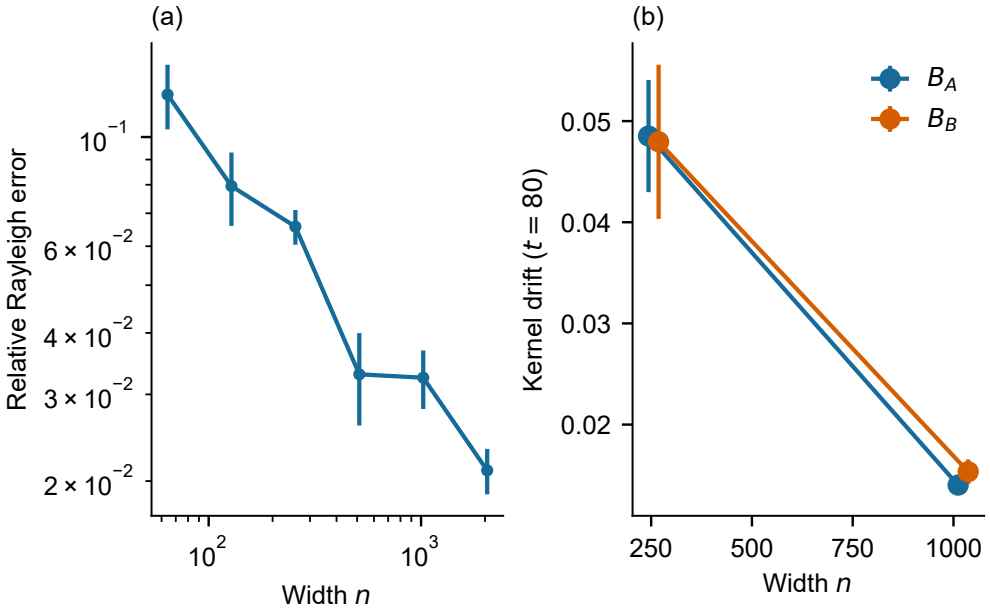


Fig. 7 Finite-width diagnostics. (a) Mean relative initial-kernel Rayleigh error over seven modes; error bars are the standard error of the mean (SEM) across 12 seed-level means. (b) Terminal relative Frobenius drift for the matched pair, mean ± SD over eight seeds

Figure 7 shows the finite-width Rayleigh errors and matched-pair kernel drift. For $B = (1, 4)$, the width experiment uses modes $0, 1, 2, 3, 4, 5, 8$ on a grid of size 512 and widths $64, 128, 256, 512, 1024, 2048$. Twelve seeds per width are numbered 200–211. For each nonzero mode, the reported value averages cosine and sine Rayleigh quotients. These measure kernel action in fixed Fourier directions; the finite-width empirical kernel can also couple different modes. The mean relative error against the population spectrum decreases from approximately 12.20% at width 64 to 2.10% at width 2,048.

E.2 Independent ridge solve and grid effects

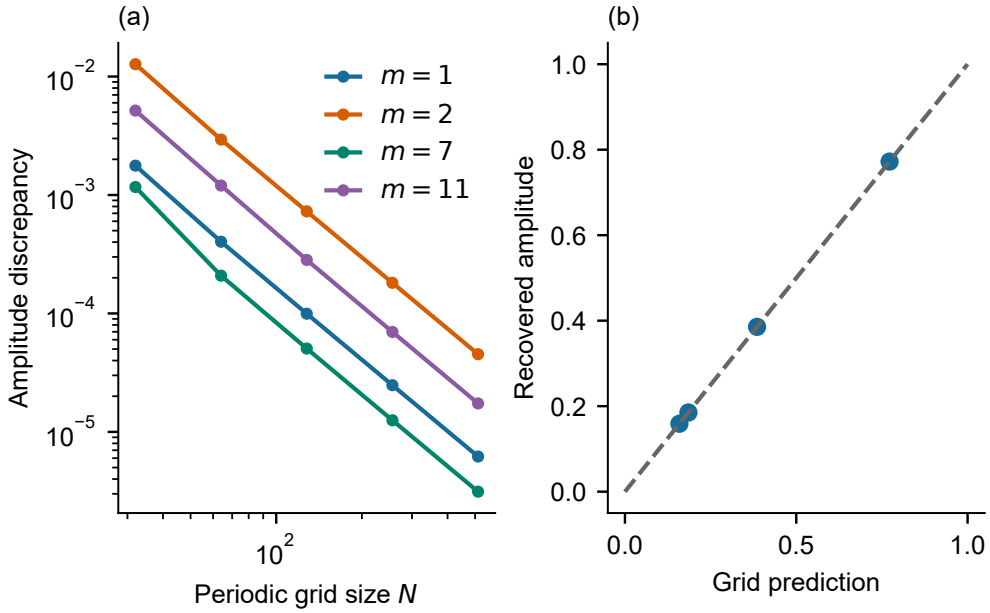


Fig. 8 Ridge calculation with $B = (1, 4, 7)$ and $\eta = 0.05$. The prediction uses a Fourier eigenvalue; the recovered amplitude is extracted from the dense solve $K(K+N\eta I)^{-1}y$. (a) Continuum predictions differ from grid solutions through aliasing. (b) Predictions using the grid eigenvalue agree with the independent solve

Figure 8 compares Fourier predictions with the independent dense ridge solve. The target modes are $(1, 2, 7, 11)$ with cosine amplitudes $(1, 0.7, 0.5, 0.4)$. The dense Gram matrix is constructed directly from embedded inputs, and SciPy solves the linear system independently of Fourier filtering. Grid sizes are 32, 64, 128, 256 and 512. Population reference eigenvalues use a grid of size 262,144. Both grid and population predictions are recorded for comparison with the dense solutions.

The matched-seed grid-sensitivity experiment gives average early rates for B_A of 0.06328, 0.06316, and 0.06315 for $N = 128, 256, 512$; corresponding B_B rates are 0.003456, 0.003361, and 0.003342. These runs use different seeds and width from the main pair experiment and are not pooled with it. The population-to-grid discrepancy is proportionally more noticeable for a small eigenvalue.

E.3 Predictor comparisons and interval construction

The additional prospective predictors have Spearman correlations 0.876, 0.842, and 0.801 for the order-two contribution, and 0.917, 0.905, and 0.924 for negative first active order, at targets $m = 5, 9, 17$, respectively. Table 3 reports the corresponding spectrum and RMS comparisons, bootstrap intervals, and calibration errors.

For each target, Spearman correlation compares the dictionary's mean fitted rate over five seeds with each prospective predictor. First active order is computed from the positive ReLU coefficients and walk probabilities through order 20; all enumerated cases arrive within that range. Order-two contribution is $a_2\mu_B^{*2}(m)$, with many tied values. Bootstrap intervals use 2,000 dictionary resamples with replacement, generator seed 111, and percentile endpoints. They characterize reweighting of the enumerated dictionary family. RMS correlations retain their signs.

For design comparisons, the saved analysis includes paired mean-MSE differences and Student-t intervals over eight paired seeds. The four target energy patterns share phases and are not independent task samples. The high-weighted task is particularly informative about computational tradeoffs: the exhaustive and greedy mean errors are 0.548989 and 0.549086, nearly indistinguishable at this resolution. A strong spectral heuristic can therefore match the exhaustive result with far fewer candidate evaluations. Figure 9 gives the complete training curves for all four energy patterns and seven dictionary rules.

Appendix F Natural-image protocol and complete results

F.1 Data and preprocessing

We use BSDS500 test images 100007, 100039, 100099, 10081, and 101027, selected by lexicographic filename order before training. RGB bytes are divided by 255 and mapped to grayscale by $0.2126R + 0.7152G + 0.0722B$, applied to the encoded red, green, and blue (RGB) channel values without gamma linearization. A 64×64 center crop is extracted without resizing, smoothing, or windowing. Subtracting its pixel mean and dividing by its root-mean-square (RMS) value gives step-zero normalized MSE one. Segmentation annotations are unused. Image hashes, crop origins, means, and RMS values are recorded in the provenance data in Online Resource 1.

F.2 Coordinates, training, and discrete spectrum

The grid is $(\theta_{\mathrm{row}}, \theta_{\mathrm{col}}) = 2\pi(i, j)/64$, with $0 \le i, j < 64$. We use eight scalar input features, the network normalization of Equation (11), and frozen initial output subtraction. Both hidden weights and output coefficients are initialized independently from standard normal distributions using NumPy's default generator with seeds 20260923–20260927; each seed is reused across images and dictionaries. Unregularized full-batch gradient descent minimizes half the mean squared residual using simultaneous parameter updates, step size 0.1, width 1,024, and 200 iterations. The implementation uses CPU float64 arithmetic and one BLAS thread. Trajectories are saved every five steps, including zero and 200. The normalized FFT of $\Psi_R(g_B)$ gives a finite-grid spectrum with trace two. On a small grid, its action agrees with an independently constructed dense kernel; centered finite differences check all analytic parameter gradients on a small network.

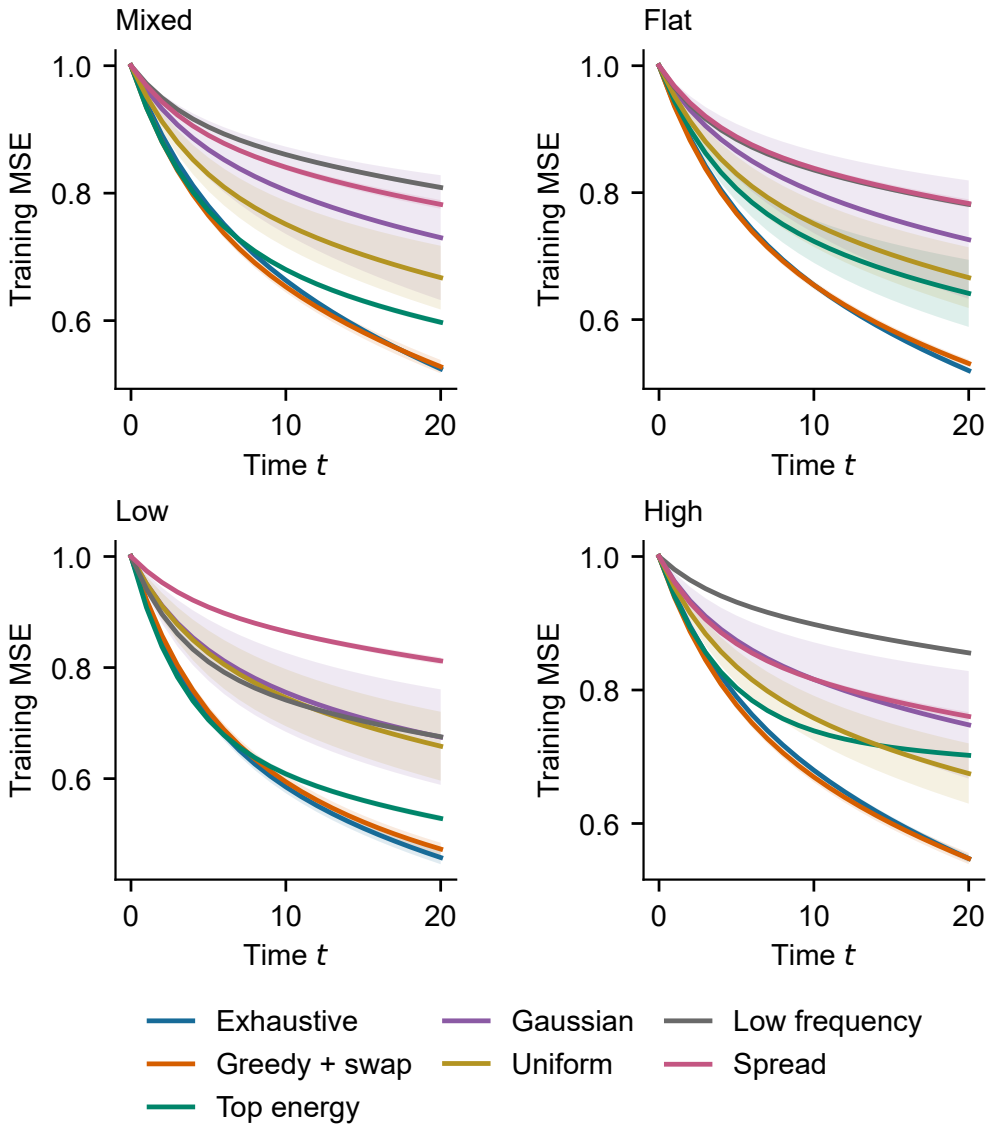


Fig. 9 Complete training curves for all four energy patterns and seven dictionary rules. Curves show means with one-SD bands over eight runs. The rules are selected before training and evaluated by finite-network MSE

F.3 Rate fitting and uncertainty

Before training, we select the 20 largest target-energy conjugate pairs excluding the zero-frequency (DC) component separately for each image. Each pair is included once, including Nyquist/self-conjugate cases. We fit minus the slope of log relative residual amplitude against iteration number at recorded steps $0, 5, \ldots, 50$. The predicted

amplitude rate is $-\log|1-0.1\lambda_m^{[64]}|$; the corresponding energy rate is twice this value. All selected modes and negative fitted rates are retained. Five fitted seed rates are averaged for each image, dictionary, and pair before computing Spearman correlations across the 20 pairs. Pooled correlations across the 100 seed-averaged pairs per dictionary are 0.908 and 0.941. They summarize mode ordering descriptively: pairs share image targets, and all images reuse the same seed list.

Table 4 Complete five-image results at step 200. Observed MSE is mean ± sample SD over five initialization seeds. The rate correlation uses the 20 seed-averaged amplitude rates fitted over steps 0–50. All errors are measured on fitted pixels after target centering and RMS normalization

Image	Dictionary	Predicted MSE	Observed MSE	Rate ρ
100007	A	0.5155	0.5137 ± 0.0022	0.761
100039	A	0.7337	0.7334 ± 0.0037	0.865
100099	A	0.4506	0.4502 ± 0.0003	0.976
10081	A	0.8353	0.8342 ± 0.0018	0.955
101027	A	0.6292	0.6289 ± 0.0027	0.867
100007	B	0.7222	0.7225 ± 0.0008	0.924
100039	B	0.8862	0.8866 ± 0.0012	0.970
100099	B	0.6479	0.6481 ± 0.0010	0.944
10081	B	0.8677	0.8686 ± 0.0015	0.953
101027	B	0.7829	0.7831 ± 0.0001	0.930

Table 4 reports the predicted and observed MSEs and rate correlations for each image and dictionary. Aggregate MSE bands use the sample SD over five seed-level means, each averaged over the five fixed crops. They measure initialization variability within this image set. During steps 0–50, the largest relative discrepancy between an image's seed-averaged MSE and its NTK prediction is 0.58% for B_A and 0.17% for B_B. At finite width, the empirical kernel can couple Fourier modes and transfer residual energy between them, producing larger deviations in individual mode rates. Figure 10 shows the five crops and their complete training curves.

Absolute modewise calibration is weaker than aggregate MSE agreement. The median relative error of the 100 seed-averaged rates per dictionary is 25.8% for B_A and 24.3% for B_B. Of these rates, 9 and 10, respectively, are negative despite positive infinite-width eigenvalues. The rank correlations summarize relative mode ordering, while these errors quantify departures from the predicted absolute dynamics.

F.4 Boundary convention

Identifying opposite crop boundaries defines the discrete torus used in Equation (19). The prediction applies to the sampled pixel array even when its edges are discontinuous. The dictionaries serve as fixed matched-scale controls chosen before training.

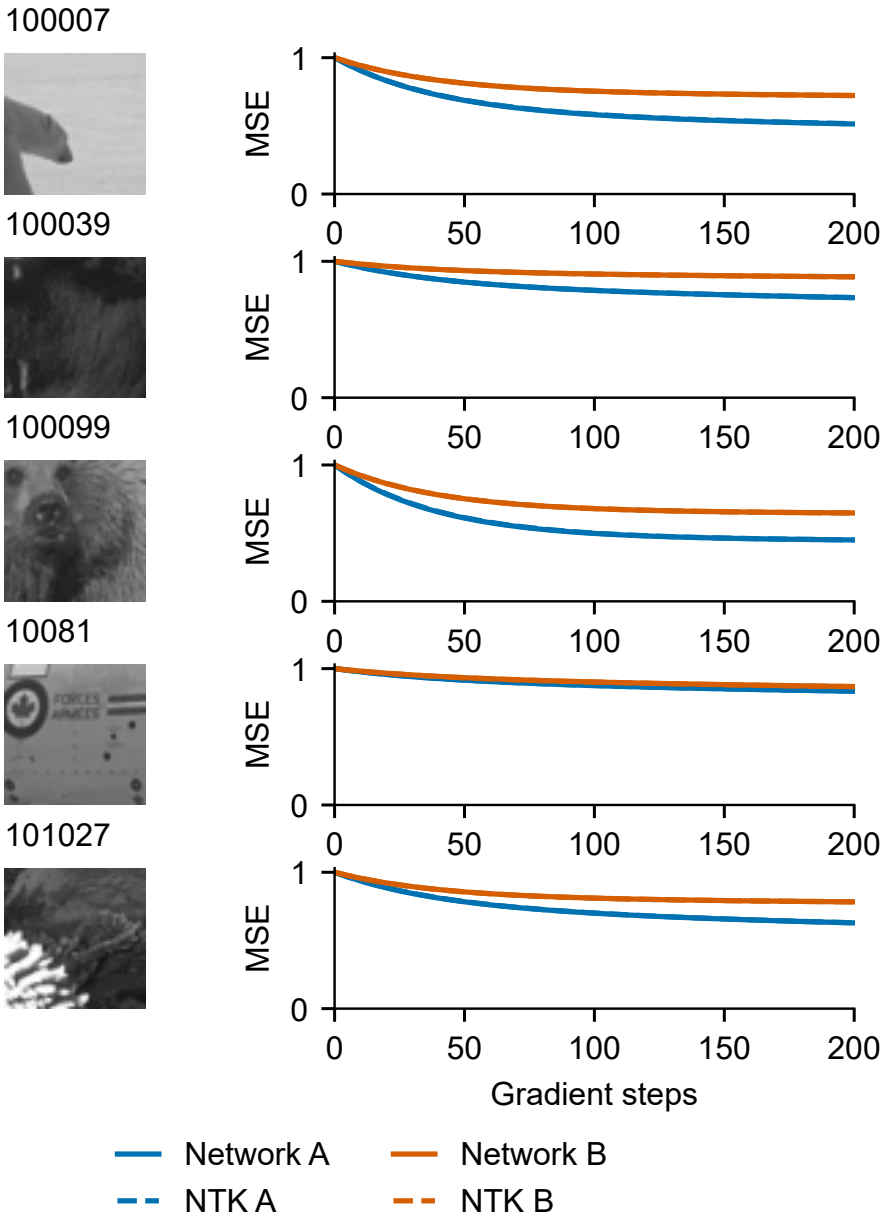


Fig. 10 All five selected center crops and their training curves. Solid lines and bands show mean ± one SD over five seeds within each image; dashed lines are exact infinite-width grid predictions. The common vertical scale permits comparison across images

References

[1] Tancik, M., Srinivasan, P.P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J.T., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. In: Advances in Neural Information Processing Systems, vol. 33, pp. 7537–7547 (2020)

[2] Benbarka, N., Höfer, T., Riaz, H.u.-M., Zell, A.: Seeing implicit neural representations as Fourier series. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2041–2050 (2022)

[3] Yüce, G., Ortiz-Jiménez, G., Besbinar, B., Frossard, P.: A structured dictionary perspective on implicit neural representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19228–19238 (2022). https://doi.org/10.1109/CVPR52688.2022.01863

[4] Ma, M., Zhu, Q., Zhan, Y., Yin, Z., Wang, H., Zheng, Y.: Robustifying Fourier features embeddings for implicit neural representations. arXiv preprint arXiv:2502.05482 (2025)

[5] Jacot, A., Gabriel, F., Hongler, C.: Neural tangent kernel: Convergence and generalization in neural networks. In: Advances in Neural Information Processing Systems, vol. 31 (2018)

[6] Lee, J., Xiao, L., Schoenholz, S.S., Bahri, Y., Novak, R., Sohl-Dickstein, J., Pennington, J.: Wide neural networks of any depth evolve as linear models under gradient descent. In: Advances in Neural Information Processing Systems, vol. 32 (2019)

[7] Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F.A., Bengio, Y., Courville, A.: On the spectral bias of neural networks. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 5301–5310 (2019). https://proceedings.mlr.press/v97/rahaman19a.html

[8] Cao, Y., Fang, Z., Wu, Y., Zhou, D.-X., Gu, Q.: Towards understanding the spectral bias of deep learning. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, pp. 2205–2211 (2021). https://doi.org/10.24963/ijcai.2021/304

[9] Arora, S., Du, S.S., Hu, W., Li, Z., Salakhutdinov, R., Wang, R.: On exact computation with an infinitely wide neural net. In: Advances in Neural Information Processing Systems, vol. 32 (2019)

[10] Murray, M., Jin, H., Bowman, B., Montúfar, G.: Characterizing the spectrum of the NTK via a power series expansion. In: International Conference on Learning Representations (2023)

[11] Belfer, Y., Geifman, A., Galun, M., Basri, R.: Spectral analysis of the neural tangent kernel for deep residual networks. Journal of Machine Learning Research **25**(184), 1–49 (2024)

[12] Li, Y., Yu, Z., Chen, G., Lin, Q.: On the eigenvalue decay rates of a class of neural-network related kernel functions defined on general domains. Journal of Machine Learning Research **25**(82), 1–47 (2024)

[13] Rahimi, A., Recht, B.: Random features for large-scale kernel machines. In: Advances in Neural Information Processing Systems, vol. 20 (2007)

[14] Rahimi, A., Recht, B.: Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In: Advances in Neural Information Processing Systems, vol. 21 (2008)

[15] Wang, S., Wang, H., Perdikaris, P.: On the eigenvector bias of Fourier feature networks: From regression to solving multi-scale PDEs with physics-informed neural networks. Computer Methods in Applied Mechanics and Engineering **384**, 113938 (2021) https://doi.org/10.1016/j.cma.2021.113938

[16] Kania, A., Mihajlovic, M., Prokudin, S., Tabor, J., Spurek, P.: FreSh: Frequency shifting for accelerated neural representation learning. In: International Conference on Learning Representations (2025). https://openreview.net/forum?id=zMjjzXxS64

[17] Molina, J., Petrache, M., Costabal, F.S., Courdurier, M.: Optimizing the dynamics of the frequency bias in Fourier features neural networks. Neurocomputing **697**, 134167 (2026) https://doi.org/10.1016/j.neucom.2026.134167

[18] Ke, J., Xu, Y., Wang, C., Wen, Y.-W.: Content-aware frequency encoding for implicit neural representations with Fourier-Chebyshev features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3646–3655 (2026). https://openaccess.thecvf.com/content/CVPR2026/html/Ke_Content-Aware_Frequency_Encoding_for_Implicit_Neural_Representations_with_Fourier-Chebyshev_Features_CVPR_2026_paper.html

[19] Morsali, A., Vaez, M., Soltani, M., Kazerouni, A., Taati, B., Mohammad-Noori, M.: A unified theory of sinusoidal activation families for implicit neural representations. Transactions on Machine Learning Research (2026)

[20] Canatar, A., Bordelon, B., Pehlevan, C.: Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. Nature Communications **12**(1), 2914 (2021) https://doi.org/10.1038/s41467-021-23103-1

[21] Caponnetto, A., De Vito, E.: Optimal rates for the regularized least-squares algorithm. Foundations of Computational Mathematics **7**(3), 331–368 (2007)

https://doi.org/10.1007/s10208-006-0196-8

[22] Arbeláez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(5), 898–916 (2011) https://doi.org/10.1109/TPAMI.2010.161

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supportmaterials.zip