# SPECTRAL BIAS IN THE EVOLVING NEURAL TANGENT KERNEL AT FINITE WIDTH

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

This paper shows how spectral bias behaves in the finite sample size and finite network width setting, for regression on the unit circle using a shallow ReLU network. Finite samples and finite width are the setting of practical training, which theory mostly overlooks. Spectral bias is the tendency of gradient-based training to learn the low-frequency part of a target before its high-frequency part. In the infinite-width limit the Neural Tangent Kernel (NTK) under the uniform measure on the circle is a convolution, so its eigenfunctions are the Fourier modes, and low frequencies, corresponding to larger eigenvalues, are learned faster. In this specific case, the limiting kernel and its spectrum are known in closed form. For a network of finite width, the NTK is computed from its actual parameters, so it is random at initialization and changes during training. Using the limiting kernel as a reference, we can measure two things separately for a finite-width NTK: how close its eigenspaces are to the Fourier modes, through the cosine similarity between the two subspaces (eigenspace alignment), and how large its average eigenvalue is on each frequency (spectral strength). With the kernel held fixed at initialization, i.e. the lazy training regime, we prove that on the low frequencies its eigenvalues and eigenvectors stay close to the limiting ones, with a deviation of $O(n^{-1/2})$ in the number of training points $n$ and $O(m^{-1/2})$ in the width $m$, and the measured deviations decay at these rates. We observe that at smaller widths the neural network with evolving weights reaches a lower training loss than the same network trained in the lazy training regime, and it is correlated with an increase in the spectral strength of the lowest frequencies while eigenspace alignment degrades at higher frequencies. The increase persists even for a target from which these low frequency modes are absent, and we prove that the expected rate at which the finite-width NTK changes at initialization does not depend on the target, so we believe the strengthened directions are set by the network's input embedding and bias, architecture and parameter distribution. A formal theory of the evolving-kernel regime remains open.

## 1 INTRODUCTION

Overparameterized neural networks generalize well despite their capacity to fit random labels (Zhang et al., 2017; Belkin et al., 2019). One of the most robust and interpretable regularities that has emerged is *spectral bias*, also called the frequency principle: gradient-based training learns smooth, low-frequency structure in a target before it learns fine, high-frequency detail (Rahaman et al., 2019; Xu et al., 2019; Zhi-Qin et al., 2020). It is reported across architectures and data types, and the standard remedies work by reshaping the effective spectrum of the network (Section 2).

In the setting of least-squares regression, under gradient flow, the residual $r_t$, i.e. error between predictions and target $(f_\theta - y)$, obeys a linear differential equation $\partial_t r_t = -T_t r_t$, during training, with $T_t$ the operator that the model and the data induce on function space. When $T_t \equiv T$ is fixed the residual decomposes in its eigenbasis and each component decays at the rate of its own eigenvalue. What is learned first then depends on the spectrum of this operator: the direction in which the eigenvectors of $T$ point, and how large the matching eigenvalues are. For a model linear in its parameters, $T$ is the data covariance operator, so the fast directions come from the inputs describing most variance in the data. However, adding a second linear layer already makes the least-squares



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.

objective non-convex in the parameters (Saxe et al., 2014), and adding nonlinearity makes analyzing the spectrum of the operator even more difficult.

Jacot et al. (2020) showed that in an infinite-width setting, a neural network, $f_\theta$, also follows a linear differential equation during training governed by a kernel called the Neural Tangent Kernel (NTK). As width grows the tangent kernel becomes deterministic and stops moving during training, so the dynamics collapse to $\partial_t r_t = -T r_t$ with $T$ fixed and known (Jacot et al., 2020). The picture rests on two assumptions: $T$ is static only in the limit, and its eigenvectors are Fourier modes only when the input distribution is symmetric enough to make $T$ a convolution, needing the continuum of data. Once both hold, the spectrum of the NTK follows a frequency ordering and low frequencies, corresponding to the largest eigenvalues, are learned first (Section 2).

Neither assumption holds in practice because we have finitely sampled training points and the width of the network is finite, so the tangent kernel is random at initialization, and once training starts it moves. This is where the network learns features. We study how much of the limiting description survives finite sampling and finite width, and what does the kernel actually do to the spectral picture once it is allowed to move?

We answer both in a setting where the idealized dynamics can be written down completely: regression on the unit circle $S^1$ using least-squared error, with a one-hidden-layer ReLU network and a uniform input distribution. Here the limiting NTK is a convolution, the Fourier modes are exact eigenfunctions, and the per-frequency eigenvalues are computable in closed form. This makes the circle a fully solvable testbed, since we can *measure* how sampling, width, and kernel evolution each perturb the tangent kernel, helping us separate effects that are otherwise entangled.

**Contributions.**

- **An explicit $S^1$ operator as the measurement instrument (Section 3).** We give the limiting NTK on $S^1$ and its Fourier eigenvalues in closed form with the output weights trainable, unlike the closely related two-layer analyses of Ronen et al. (2019); Basri et al. (2020).
- **An initialization baseline via concentration (Section 4).** On a fixed low-frequency subspace we prove that the sampled operator ($O(n^{-1/2})$) and the frozen finite-width operator ($O(m^{-1/2})$) concentrate around the limiting operator, and the measured errors decay at these rates.
- **An alignment versus strength dissociation (Section 5).** At small width we observe that the evolving kernel raises the low-frequency spectral strength instead of improving Fourier alignment, on the constant mode and $k = 1$ even when the target excites neither.
- **An exact first-order result for the moving kernel (Appendix D).** The expected initial velocity of the kernel does not depend on the target (Theorem 13) and is negative of order $1/m$ on $k = 0$ and $k = 1$ (Theorem 15), the opposite sign to the growth we observe, so that growth is not a first-order effect.

## 2 Related work

**Spectral bias and the frequency principle.** Earlier understanding of spectral bias has been largely through empirical observations. Rahaman et al. (2019) use Fourier analysis to show that deep ReLU networks are biased towards low-frequency functions, and that the parameters must be finely tuned to express high frequencies, so the high-frequency components are the ones least robust to parameter perturbation. Literature on training dynamics show that neural networks fit a target from low to high frequency, and attributes it to the regularity of the commonly used activation functions (Xu et al., 2019; Zhi-Qin et al., 2020), with a stage-by-stage treatment for general architectures in Luo et al. (2021). Fridovich-Keil et al. (2022) measure the same preference in image classification networks on CIFAR-10 and ImageNet. These arguments use the parameterization and the properties of the activation function rather than considering a function-space operator.

**NTK and fixed-kernel spectral bias.** The NTK reduces wide-network training to a linear dynamical system governed by a fixed operator (Jacot et al., 2020; Lee et al., 2020), and convergence is

controlled by its spectrum (Du et al., 2019; Arora et al., 2019). Cao et al. (2020) make the spectral-bias link explicit by showing that residual components aligned with larger eigenvalues are learned faster. Under rotational symmetry the operator is diagonal in a harmonic basis, so the eigenvalue ordering is a frequency ordering on the circle and sphere (Ronen et al., 2019; Bietti & Mairal, 2019; Basri et al., 2020). The decay of those eigenvalues has been analysed directly, with the rate tied to the activation function (Murray et al., 2023), and a related ordering runs over sample size rather than training time, with kernel regression fitting successively higher spectral modes as the training set grows (Bordelon et al., 2020). We use exactly this setting, but keep the output layer trainable so the kernel and its odd-mode spectrum differ from the bias-free two-layer models, and we track how the Fourier structure is perturbed rather than only stating it in the limit.

**Finite width and kernel evolution.** Finite width makes the NTK random at initialization and non-stationary during training (Hanin & Nica, 2019). Bordelon & Pehlevan (2023) analyse finite-width kernel and prediction fluctuations, separating a lazy training regime from feature learning, and Vyas et al. (2022) show the empirical NTK keeps evolving and that the frozen kernel can miss finite-network generalization. Feature-learning and mean-field limits describe large kernel motion but through a parameter-distribution evolution rather than a fixed function-space operator (Mei et al., 2018; Chizat & Bach, 2018; Chizat et al., 2019). Much of the motion occurs in an early transient during which the kernel changes rapidly (Fort et al., 2020). A related line reads that motion as alignment: tangent features align along a small number of task-relevant directions (Baratin et al., 2021), and Canatar et al. (2021) use the target's alignment with the kernel eigenbasis, together with the spectrum, to explain generalization in kernel regression and wide networks.

**Spectral bias in practice.** Coordinate networks blur high-frequency detail and are fixed by Fourier-feature mappings or periodic activations that reshape the effective spectrum (Tancik et al., 2020; Sitzmann et al., 2020); PINNs fail on high-frequency PDE components and are corrected by eigenvalue-based rescaling of the loss terms (Wang et al., 2022); and preconditioned methods deliberately modify the NTK spectrum to accelerate training and reduce spectral bias (Geifman et al., 2024; Jiang et al., 2026; Yang, 2024). Appendix E.7 uses the same eigenvalue rescaling to check our finite-sample analysis (Section 4.1) by verifying that rescaling by the predicted diagonal flattens the per-mode rates, as the finite-sample result predicts.

## 3 Setup and the continuum reference model

### 3.1 Network, domain, and residual dynamics

We identify the unit circle $S^1$ with the angular domain $\Omega = [0, 2\pi)$ under the uniform probability measure $d\mu(\varphi) = d\varphi/(2\pi)$, and embed it as $x(\varphi) = (\cos\varphi, \sin\varphi) \in \mathbb{R}^2$. This embedding introduces rotational symmetry and provides the natural Fourier basis for $L^2(\mu)$.

We consider a one-hidden-layer ReLU network of width $m$ in NTK parameterization,

$$f_m(x;\theta) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \, \sigma(w_i^\top x + b_i), \tag{1}$$

with $\sigma(z) = \max\{z, 0\}$, and where the parameters are initialized as

$$a_i \sim \mathcal{N}(0,1), \qquad w_i \sim \mathcal{N}(0, I_2), \qquad b_i(0) = 0, \tag{2}$$

independently across $i$. The hidden biases start at zero but remain trainable. Unlike the closely related two-layer analyses of Ronen et al. (2019); Basri et al. (2020), which train only the input weights and biases, we keep the output weights $a_i$ trainable and include their contribution to the kernel.

Every target we train on is a finite Fourier mixture

$$y(\varphi) = \sum_k c_k \cos(k\varphi + \beta_k), \tag{3}$$

so a mode is one frequency $k$ and $c_k$ is its amplitude; the frequencies, amplitudes and phases are stated per experiment. For such a $y$ and $f_t(\varphi) := f_m(x(\varphi); \theta(t))$, we minimize the population

least-squares loss $\mathcal{L}(\theta) = \frac{1}{2}\|f_t - y\|^2_{L^2(\mu)}$ and write the residual $r_t = f_t - y$. The empirical, or finite-width, tangent kernel at time $t$ is $\Theta^{(m)}_t(\varphi, \varphi') = \langle \nabla_\theta f_m(\varphi; \theta(t)), \nabla_\theta f_m(\varphi'; \theta(t))\rangle$, with integral operator $T^{(m)}_t$. Under gradient flow the residual and loss obey

$$\partial_t r_t = -T^{(m)}_t r_t, \qquad \frac{d}{dt}\tfrac{1}{2}\|r_t\|^2_{L^2(\mu)} = -\langle r_t,\, T^{(m)}_t r_t\rangle_{L^2(\mu)}. \tag{4}$$

We compare two versions of these dynamics at the same width. In the *evolving* dynamics all parameters $a_i, w_i, b_i$ are trained, so $T^{(m)}_t$ changes with $t$. In the *frozen* dynamics the operator is held at its value at initialization, $\partial_t r_t = -T^{(m)}_0 r_t$, which is training the linearization of the network around $\theta(0)$, i.e. the lazy training regime (Chizat et al., 2019). We call $\partial_t T^{(m)}_t$ the *velocity* of the kernel, which measures how fast the kernel changes during training, and its value at $t = 0$ the initial velocity. Some experiments use the *centered* model $f_\theta - f_{\theta(0)}$ (Chizat et al., 2019), whose output is zero at initialization, so its initial residual is $-y$. In experiments $\mu$ is replaced by the $n$ training points, and gradient descent runs on $\frac{1}{2}\|r_t\|^2_2 = \frac{1}{2}\sum_{i=1}^n r_t(\varphi_i)^2$.

### 3.2 Infinite-width limit and Fourier diagonalization

In the infinite-width limit the tangent kernel converges almost surely to a deterministic limit $T_\infty = \lim_{m\to\infty} T^{(m)}_0$ and, in this regime, stays constant during training (Jacot et al., 2020; Lee et al., 2020). The residual then evolves under a fixed linear operator. Because the initialization is isotropic, the limiting kernel is rotation-invariant, so $\Theta(\varphi, \varphi')$ depends only on the angular difference and $T_\infty$ acts as a circular convolution. Its eigenfunctions are therefore the Fourier modes. We use the real Fourier basis of $L^2(\mu)$,

$$\phi_0(\varphi) = 1, \qquad \phi_{k,c}(\varphi) = \sqrt{2}\cos(k\varphi), \qquad \phi_{k,s}(\varphi) = \sqrt{2}\sin(k\varphi), \qquad k \geq 1, \tag{5}$$

and for an integer $K \geq 1$, let the truncated low-frequency subspace be $\mathcal{H}_K = \mathrm{span}\{\phi_0, \phi_{k,c}, \phi_{k,s} : 1 \leq k \leq K\}$ of dimension $d = 2K + 1$. Let $P_K : L^2 \to \mathcal{H}_K$ be the orthogonal projection onto $\mathcal{H}_K$. Throughout, all comparisons between the continuum, sampled, frozen, and evolving settings are made on $\mathcal{H}_K$ as it reduces the infinite-dimensional problem to a fixed low-frequency subspace that survives every perturbation and can be compared directly. Each frequency $k \geq 1$ contributes a two-dimensional plane $\mathcal{F}_k = \mathrm{span}\{\phi_{k,c}, \phi_{k,s}\}$, with $\mathcal{F}_0 = \mathrm{span}\{\phi_0\}$.

**Strength and alignment.** In experiments the operator $T^{(m)}_t$ is represented by the matrix $\widehat{T}_{ij} = \frac{1}{n}\Theta^{(m)}_t(\varphi_i, \varphi_j)$ on the $n$ training points, the $\frac{1}{n}$ making it approximate the integral operator. Let $U_k$ be the matrix whose orthonormal columns are the Fourier modes of frequency $k$ evaluated on these points, $d_k = \dim \mathcal{F}_k$, and $P_{\mathcal{F}_k} = U_k U_k^\top$ the projection onto their span. The *spectral strength* on $\mathcal{F}_k$ is the mean eigenvalue of $\widehat{T}$ restricted to that subspace,

$$\mu_{\mathcal{F}_k} = \tfrac{1}{d_k}\,\mathrm{tr}\big(P_{\mathcal{F}_k}\widehat{T}P_{\mathcal{F}_k}\big). \tag{6}$$

For *eigenspace alignment* we take the $d_k$ eigenvectors $v_j$ of $\widehat{T}$ with the largest overlap $\|U_k^\top v_j\|^2$, which does not assume that the eigenvalues keep the frequency ordering, and collect them as the columns of $V_k$. The singular values of $U_k^\top V_k$ are the cosines of the principal angles $\omega_1, \ldots, \omega_{d_k}$ between the two subspaces, and we report their root mean square,

$$s^{\cos}_k = \Big(\tfrac{1}{d_k}\textstyle\sum_{i=1}^{d_k}\cos^2\omega_i\Big)^{1/2} \in [0, 1], \tag{7}$$

which is 1 exactly when the matched eigenspace equals $\mathcal{F}_k$. The principal angles do not depend on the basis chosen inside a plane, where cos and sin can rotate into each other, and the root mean square uses both angles, so it tells us how well the plane is aligned as a whole.

### 3.3 Kernel and spectrum on the circle

The limiting kernel and its Fourier eigenvalues are available in closed form. Derivations are in Appendix A.

**Proposition 1** (Limiting NTK on $S^1$). *For the network (equation 1) and initialization (equation 2), the limiting kernel $\Theta(\varphi, \varphi') = \lim_{m\to\infty} \Theta_0^{(m)}(\varphi, \varphi')$ depends only on the angle $\gamma \in [0, \pi]$ between $x(\varphi)$ and $x(\varphi')$, and equals*

$$\Theta(\gamma) = \frac{1}{2\pi}\Big(\sin\gamma + 2(\pi - \gamma)\cos\gamma + (\pi - \gamma)\Big). \tag{8}$$

**Theorem 2** (Fourier eigenvalues of the limiting NTK). *Let $\lambda_k$ be the eigenvalue of $T_\infty$ at Fourier frequency $k$. Then*

$$\lambda_0 = \frac{1}{4} + \frac{3}{\pi^2}, \qquad \lambda_1 = \frac{1}{4} + \frac{1}{\pi^2}, \tag{9}$$

*and for $k \geq 2$,*

$$\lambda_k = \begin{cases} \dfrac{k^2+3}{\pi^2(k^2-1)^2}, & k \geq 2 \text{ even}, \\ \dfrac{1}{\pi^2 k^2}, & k \geq 3 \text{ odd}. \end{cases} \tag{10}$$

*Both branches satisfy $\lambda_k = O(k^{-2})$, so higher-frequency residual components decay more slowly under fixed-kernel gradient flow.*

This continuum model is the reference against which every finite effect below is measured.

## 4 Robustness under finite sampling and frozen finite width

We keep the kernel fixed and replace each idealization in turn, the uniform measure by $n$ sampled points and the infinite-width operator by a frozen finite-width one, and bound how far each takes the operator from the continuum Fourier block.

### 4.1 Finite-sample control

Let $\varphi_1, \ldots, \varphi_n \overset{\text{iid}}{\sim} \mathrm{Unif}[0, 2\pi)$. The sampled Fourier basis is no longer exactly orthonormal and the modes are no longer exact eigenvectors. We track three errors that measure how close the sampled basis is to orthonormal, how close the operator is to diagonal in that basis, and how close the sampled modes are to eigenvectors. They are built from the feature matrix $\Phi_{i,p} = \phi_p(\varphi_i) \in \mathbb{R}^{n\times d}$, where $p$ indexes the $d$ basis functions of $\mathcal{H}_K$, the sampled kernel matrix $A_{ij} = \frac{1}{n}\Theta(\varphi_i - \varphi_j)$, which depends on the difference alone by Proposition 1, the Gram matrix $G = \frac{1}{n}\Phi^\top\Phi$, and the matrix of $A$ in the sampled Fourier basis, $H = \frac{1}{n}\Phi^\top A\Phi$. The comparison is against the diagonal finite-sample prediction $\Lambda^{(n)} = \mathrm{diag}(\lambda_p^{(n)})$, whose entries correct the continuum eigenvalues by an $O(1/n)$ term.

**Proposition 3** (Expectation of the restricted sampled operator). *With $\lambda_p^{(n)} = (1 - \frac{1}{n})\lambda_p + \frac{\Theta(0)}{n}$ and $\Lambda^{(n)} = \mathrm{diag}(\lambda_p^{(n)})$, we have $\mathbb{E}[H] = \Lambda^{(n)}$.*

**Theorem 4** (Finite-sample concentration on $\mathcal{H}_K$). *Let $\kappa = \sup_\gamma |\Theta(\gamma)| < \infty$. For every $\varepsilon > 0$,*

$$\mathbb{P}(\|G - I_d\|_{\max} \geq \varepsilon) \leq d(d+1)\exp\Big(-\tfrac{n\varepsilon^2}{4+\frac{4}{3}\varepsilon}\Big), \tag{11}$$

$$\mathbb{P}\Big(\|H - \Lambda^{(n)}\|_{\max} \geq \varepsilon\Big) \leq 2d^2\exp\Big(-\tfrac{n\varepsilon^2}{32\kappa^2}\Big), \tag{12}$$

*and, for $n \geq 2$,*

$$\mathbb{P}\Big(\|A\Phi - \Phi\Lambda^{(n)}\|_{\max} \geq \varepsilon\Big) \leq 2nd\exp\Big(-\tfrac{n\varepsilon^2}{2\kappa^2+2\kappa\varepsilon}\Big). \tag{13}$$

*Consequently, with probability at least $1-\delta$, all three errors are $O\big(\sqrt{\log(nd/\delta)/n}\big)$ simultaneously.*

Proofs are in Appendix B. All three errors decrease with $n$ (Figure 5, Appendix E).

### 4.2 FROZEN FINITE-WIDTH CONTROL

We now consider the situation where the width is finite, returning to the continuum input space and freezing the tangent kernel at initialization. The object is the low-frequency subspace $B_m^{(K)} = P_K T_0^{(m)} P_K$ of the frozen operator, where $T_0^{(m)} = \frac{1}{m}\sum_{i=1}^m T^{(i)}$ and $T^{(i)}$ is the operator of the kernel that neuron $i$ contributes on its own, so the frozen operator is an average of $m$ i.i.d. copies of it (Appendix C.1). The continuum reference block is $\Lambda_K = P_K T_\infty P_K = \mathrm{diag}(\lambda_0, \lambda_1, \lambda_1, \ldots, \lambda_K, \lambda_K)$.

**Theorem 5** (Finite-width concentration of the frozen low-mode operator)**.** *Fix $K \geq 1$, $d = 2K+1$. Then*

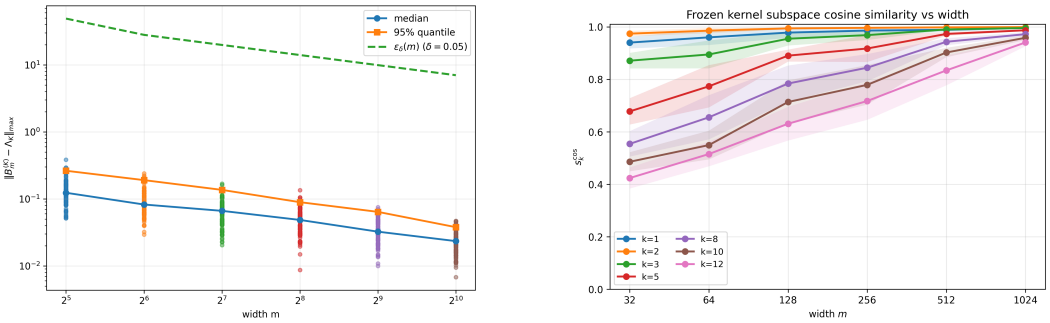
$$\mathbb{E}[B_m^{(K)}] = \Lambda_K, \tag{14}$$

*and there is a universal constant $c > 0$ such that for every $\varepsilon > 0$,*

$$\mathbb{P}\Big(\|B_m^{(K)} - \Lambda_K\|_{\max} \geq \varepsilon\Big) \leq 2d^2 \exp\Big(-cm \min\big(\tfrac{\varepsilon^2}{32^2}, \tfrac{\varepsilon}{32}\big)\Big). \tag{15}$$

*In particular, for fixed $K$, $\|B_m^{(K)} - \Lambda_K\|_{\max} = O\big(\sqrt{\log K/m}\big)$.*

The proof is in Appendix C. Figure 1 (a) shows the entrywise concentration trend. The bound itself is loose, since $c$ is the universal constant of Bernstein's inequality for sub-exponential variables, which we do not optimize, and the $\psi_1$ bound we feed it is crude, so we believe a better constant would bring the envelope much closer to the measured error. The same trend holds geometrically: measuring the alignment $s_k^{\cos}$ of equation 7 for the frozen kernel $T_0^{(m)}$, alignment improves with width, with higher frequencies needing larger widths (Figure 1 b; the projector-error view is in Appendix E.4).




(a) Block concentration.

(b) Subspace cosine similarity vs width.

Figure 1: Frozen finite-width control. (a) Median and 95% quantile of $\|B_m^{(K)} - \Lambda_K\|_{\max}$ over 100 initializations, with the envelope from Theorem 5 ($\delta = 0.05$, illustrative $c = 0.2$). (b) Eigenspace alignment between $\mathcal{F}_k$ and the matched empirical eigenspace improves with width; this is the initialization baseline for Section 5.

As long as the kernel is held fixed, the infinite-width Fourier picture explains the spectral bias we measure at finite sample size and finite width. On the low-frequency block, the error between the finite operator and the limiting one is bounded by Theorems 4 and 5, and it shrinks as $n$ and $m$ grow. The closest prior result, Bowman & Montufar (2022), bounds the whole trajectory in $L^2$ but keeps the two finite effects together and gives no per-mode corrected eigenvalues, which is what separating them on this block provides. What none of this covers is the kernel actually moving. Appendix D covers its motion at first order.

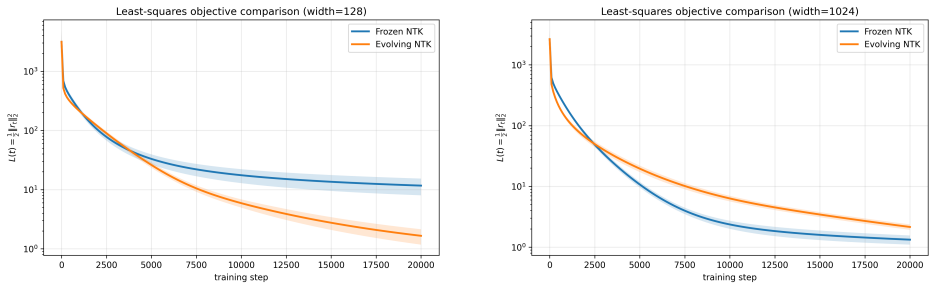
## 5 THE EVOLVING KERNEL

We now compare the frozen tangent operator with the operator that *evolves* during full training of the finite-width network. This part is empirical: at finite width the kernel has no closed form and drifts as the parameters move, and we measure what that drift does to the Fourier picture. Unless a caption says otherwise, every run in Sections 5.1 and 5.2 takes equation 3 over $k = 0, \ldots, 6$

with amplitudes $c_k = 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4$ and phases $\beta_k = \pi/2$ for $k = 2, 5, 6$ and 0 otherwise, on $n = 2048$ equally spaced points, for $20,000$ gradient-descent steps at learning rate $10^{-4}$, averaged over 5 seeds.

### 5.1 The evolving kernel wins at small width

Figure 2 compares the least-squares objective under frozen and evolving finite-width dynamics across widths. In the frozen regime, wider is uniformly better, exactly as the concentration result predicts. In the evolving regime the curves cluster: at widths 128 and 256 the evolving kernel reaches a *lower* final loss than the frozen kernel of the same width, the advantage shrinks at 512, and by width 1024 the frozen kernel is again better. Kernel evolution helps, but only at small width.




(a) Width m=128.

(b) Width m=1024.

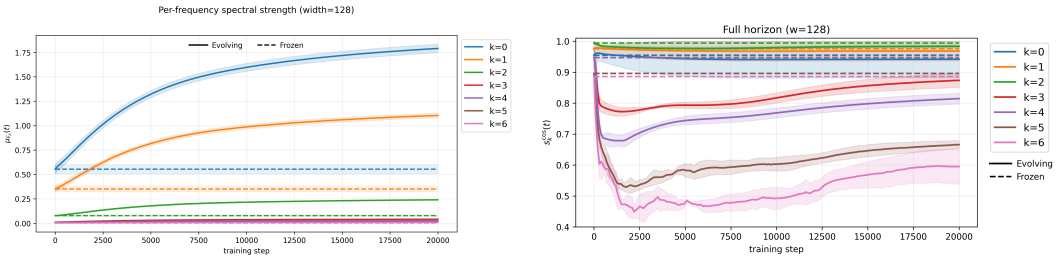
Figure 2: Least-squares objective $\frac{1}{2}\|r_t\|_2^2$ under frozen and evolving finite-width tangent-kernel dynamics, at the two ends of the width range. Widths 256 and 512 are in Figure 17, Appendix E.5. Curves are means over 5 seeds, with $\pm 1$ standard-deviation bands.

### 5.2 Spectral strength & alignment

By the loss-decay identity equation 4, faster decay can come from two different changes in the operator: better *alignment* of its eigenspaces with the residual (here, the Fourier subspaces), or larger *strength* in those subspaces. We measure both. For each Fourier subspace $\mathcal{F}_k$ we track the alignment $s_k^{\cos}(t)$ of equation 7 and the spectral strength $\mu_{\mathcal{F}_k}(t)$ of equation 6 for the current operator $T_t^{(m)}$. If the evolving kernel helped by preserving geometry, alignment should improve; if it helped by reweighting, low-frequency strength should grow even as alignment stays flat or worsens.

Figure 3 shows the joint diagnostics at width 128, where the effect is strongest; the same two panels at all four widths, and the mode-wise residual decay, are in Appendix E.5. The low-frequency-first *ordering* of the mode-wise decay is preserved; kernel evolution changes relative rates within the retained modes but does not overturn the ordering. Alignment does *not* improve. The lowest subspaces ($k = 0, 1$, and to a lesser extent $k = 2$) stay close to their Fourier planes, while several higher-frequency subspaces lose alignment early and only partly recover. The largest change in operator strength happens in the low-frequency blocks that were already well aligned. At width 128 the strength in the constant and first-frequency blocks grows by a large factor during training; the effect is present at 256, weaker at 512, and small at 1024.

We believe that the small-width benefit comes from added low-frequency strength on blocks that are already aligned, rather than from better Fourier geometry. The spectral explanation of the bias therefore seems to persist despite the kernel's motion: the moving kernel strengthens the same low-frequency eigenvalues that already decay fastest. The strength increase seems to correlate with lower loss. The effect fades as width grows, which is what lazy training predicts (Jacot et al., 2020; Lee et al., 2020; Chizat et al., 2019). Nor is it an artifact of the target amplitudes: the low-frequency strength still rises above the frozen baseline under equal-amplitude and reversed-amplitude targets (Appendix E.5).




(a) Spectral strength, width 128.

(b) Subspace alignment, width 128.

Figure 3: The mechanism at width 128. The evolving kernel (solid) sharply increases low-frequency spectral strength (left) while its Fourier alignment (right) does not improve and often degrades at higher frequencies. Curves are means over 5 seeds, with $\pm 1$ standard-deviation bands.

### 5.3 Effect of the input embedding and bias on strengthened directions

Why does the added strength concentrate on $k = 0$ and $k = 1$? The architecture gives a candidate reason. With inputs $x(\varphi) = (\cos\varphi, \sin\varphi)$, the pre-activation of hidden neuron $j$ is

$$w_j^\top x(\varphi) + b_j = \|w_j\| \cos(\varphi - \psi_j) + b_j, \tag{16}$$

where $\psi_j$ is the angle of $w_j$, so the bias and the input embedding carry frequencies $k = 0$ and $k = 1$ directly, while higher harmonics appear only after the ReLU acts on this frequency-one pre-activation. We hypothesize that the strengthened directions are set by the input embedding and bias. We take the time derivative of the kernel at initialization and ask whether it already predicts the growth in the low-frequency eigenvalues that we observe (Appendix D). Whatever frequencies the residual carries, the pre-activation of each neuron can only move inside $\mathrm{span}\{\phi_0, \phi_{1,c}, \phi_{1,s}\}$ (Lemma 11), since a bias and a two-dimensional input embedding are all it has. This constrains the pre-activations and not the kernel itself. The kernel depends on the pre-activations through the ReLU and its derivative, which contain all frequencies, so Lemma 11 does not show that the spectral strength grows on $k = 0$ and $k = 1$. It is consistent with what we observe, but it does not explain it.

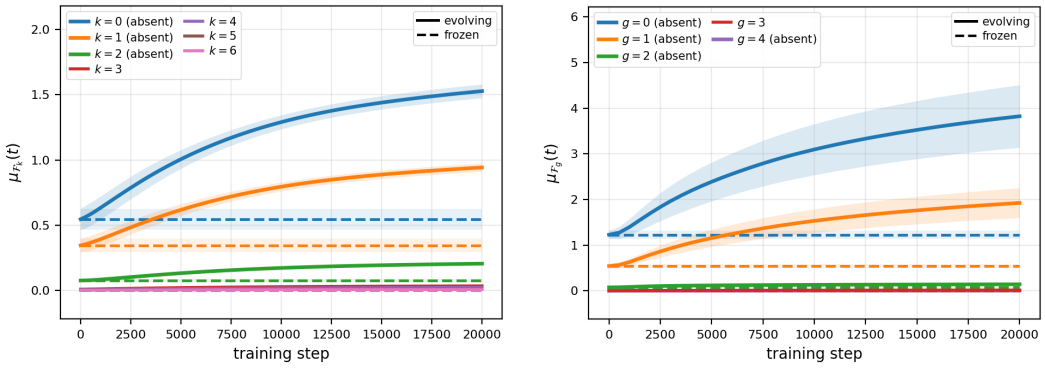
We also test the alternative that gradient descent amplifies wherever the residual is largest. Theorem 13 already rules it out at first order for any target, so what the experiment tests is the accumulated motion it leaves open. We train on a target built only from frequencies $k = 3, 4, 5, 6$, so the constant mode, $k = 1$ and $k = 2$ are absent and the target has zero empirical mean. The model is centered, so the step-0 residual is exactly $-y$ and those three blocks carry no target mass, $\|P_{\mathcal{F}_k} y\| = 0$; they acquire no appreciable residual mass $\|P_{\mathcal{F}_k} r_t\|$ at any later point either. We expect the amplification to happen in frequencies $k = 3$ through $k = 6$ but experiments show that most strength is accumulated in the first two blocks (Figure 4a), while the four blocks the target excites stay negligible throughout. The computation in Appendix D shows that in expectation the initial velocity of the kernel is independent of the target (Theorem 13), exactly zero for the centered model used here (Corollary 14), and negative on the low-frequency blocks otherwise (Theorem 15), so it does not predict the growth.

We also run the same experiment on a different domain and a different input distribution. We replace $S^1$ by $\mathbb{R}^5$ with $x \sim \mathcal{N}(0, I_5)$ and Fourier frequency by total Hermite degree (O'Donnell, 2014), and hold the rest of the design fixed, tracking degrees 0 through 4. The target is drawn from the degree-3 block alone,

$$y = \sum_{|\alpha|=3} c_\alpha h_\alpha, \qquad h_\alpha(x) = \prod_{i=1}^{5} \frac{\mathrm{He}_{\alpha_i}(x_i)}{\sqrt{\alpha_i!}}, \qquad \textstyle\sum_{|\alpha|=3} c_\alpha^2 = 1, \tag{17}$$

where $\mathrm{He}_\ell$ is the probabilists' Hermite polynomial of degree $\ell$ and the 35 coefficients $c_\alpha$ are a fixed Gaussian draw, so $\|y\|_{L^2} = 1$. The circle target spans four frequencies, this one is a *single* degree, so degree 3 is the only block carrying target mass and every other tracked degree is absent from it, degree 4 included. We observe the same behaviour (Figure 4b). Degrees 0 and 1, which are the bias and $w_j^\top x$, take almost all of the gain, while the absent degrees 2 and 4 barely move. Lemma 11 transfers here unchanged, since the pre-activation is still affine in $x$: the pre-activation motion stays

inside Hermite degree at most 1, which is where the strength increases, with the same caveat that this does not constrain the kernel itself. The results that the velocity is linear in the residual and odd in the output weights, which give its target independence, transfer as well (Section D.4).




(a) $S^1$, uniform inputs, frequency blocks. (b) $\mathbb{R}^5$, Gaussian inputs, Hermite degree blocks.

Figure 4: Mode-absent controls in two input geometries, showing the spectral strength $\mu_{\mathcal{F}_k}$ of equation 6. Solid is the evolving kernel, dashed the frozen one, bands are $\pm 1$ standard deviation over seeds. Blocks marked absent carry *exactly* zero target mass, and by Corollary 14 the expected initial velocity of every block is zero for this centered model. Both runs use width 128, the centered model $f_\theta - f_{\theta(0)}$, and 20000 gradient-descent steps at learning rate $10^{-4}$. **(a)** Uniform inputs on $S^1$, $n = 1024$, ten seeds, target $k = 3, 4, 5, 6$ with amplitudes $0.7, 0.6, 0.5, 0.4$. **(b)** Gaussian inputs $x \sim \mathcal{N}(0, I_5)$, $n = 2048$, five seeds, Hermite degree in place of frequency, block dimensions $1, 5, 15, 35, 70$, target equation 17.

## 6 Discussion and Conclusion

On the uniform circle, spectral bias has an exact description: the infinite-width NTK is diagonal in the Fourier basis, each frequency has its own decay rate, and low frequencies are learned faster because their eigenvalues are larger. Our results place two boundaries on this description. It is *robust* to finite sampling and frozen finite width, degrading only as controlled $O(n^{-1/2})$ and $O(m^{-1/2})$ perturbations on a fixed low-frequency block. And it is *incomplete* once the kernel evolves. At small width the lower loss is accompanied by added strength in the kernel's lowest-frequency components rather than by improved Fourier alignment, a reinforcement of the same low-frequency directions that fixed-kernel theory does not capture and that fades as width grows.

The Fourier picture relies on a rotation-invariant kernel and a uniform input measure. A non-uniform density breaks the convolution structure (Basri et al., 2020; Bowman & Montufar, 2022), and Gaussian inputs replace the Fourier basis by the Hermite polynomials, graded by total degree in place of frequency (O'Donnell, 2014). We ran the Gaussian case in Section 5.3, where the mode-absent control behaves as it does on the circle.

The concentration results are limited to fixed blocks and fixed $K$, and the entrywise union bounds are conservative, so tighter operator-norm or variance-sensitive tools (Tropp, 2015; Serfling, 1980) would close the gap to the observed errors. Our finite-width theorem controls only the frozen operator, and the observed strengthening effect is empirical, on a single architecture, and remains a correlation with lower loss rather than an established cause. A full account would require treating finite sampling, finite width, and training time together, over the time scale on which the slow modes are learned. The initial velocity does not account for the strengthening (Appendix D).

The most direct next step is to control how the low-frequency block drifts from initialization and what that drift does to the residual dynamics. These results hold on a single architecture and a single initialization model, so varying both is the natural follow-up.

### REPRODUCIBILITY STATEMENT

Every theoretical claim is stated with its assumptions in the main text and proved in full in the appendices: the limiting kernel and its Fourier spectrum on $S^1$ in Appendix A, the finite-sample concentration bounds in Appendix B, the frozen finite-width bounds in Appendix C, and the initial kernel velocity in Appendix D. The network, its NTK parameterization, the initialization, and the training objective are given in Section 3.1; all data is generated synthetically on the uniform circle, so there is no dataset collection or preprocessing to describe. Widths, sample sizes, target frequencies, and seed counts are stated in Section 5 or in the caption of the relevant figure. Averaged curves report means over 5 seeds with $\pm 1$ standard-deviation bands, except where a caption says otherwise; the appendix eigenvalue diagnostics average over 10 seeds, and the mode-absent controls of Section 5.3 use ten seeds on $S^1$ and five on Gaussian inputs. Appendix E reports the all-width diagnostics and the fixed-kernel preconditioning consistency check behind the claims in the main text.

### ETHICS STATEMENT

This work is a theoretical and empirical study of training dynamics on synthetic regression targets over the unit circle. It involves no human subjects, no personal or sensitive data, and no released dataset. The models are one-hidden-layer networks of width at most a few thousand, trained on generated data, so the compute footprint is small and the results carry no direct application to a deployed system. We see no foreseeable path from these findings to harm, and we declare no conflicts of interest or sponsorship that could have influenced the analysis.

### AI USE STATEMENT

We used generative AI tools as a writing and editing aid on the manuscript: tightening prose, shortening passages carried over from a longer thesis version, and checking LaTeX formatting. Every sentence so produced was read and revised by the authors, and no text was accepted without verification against the underlying result it describes.

AI tools were occasionally used to suggest proof ideas and to clean up the write-up of the derivations in Appendix D. Every statement and every step was checked by the authors.

Experiment code was written with an AI assistant under author supervision, and its outputs were checked against the closed-form predictions derived in the paper.

We did not use generative AI tools to generate research ideas or hypotheses, or to retrieve and summarize related work; the literature discussion in Section 2 was written by the authors. We take responsibility for the final content of this work, including all text, claims, and figures.

## REFERENCES

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks, 2019. URL https://arxiv.org/abs/1901.08584.

Aristide Baratin, Thomas George, César Laurent, R. Devon Hjelm, Guillaume Lajoie, Pascal Vincent, and Simon Lacoste-Julien. Implicit regularization via neural feature alignment. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2269–2277, 2021.

Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. Frequency bias in neural networks for input of non-uniform density. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 685–694. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/basri20a.html.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, July 2019. ISSN 1091-6490. doi: 10.1073/pnas.1903070116. URL http://dx.doi.org/10.1073/pnas.1903070116.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels, 2019. URL https://arxiv.org/abs/1905.12173.

Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks, 2023. URL https://arxiv.org/abs/2304.03408.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1024–1034, 2020.

Benjamin Bowman and Guido Montufar. Spectral bias outside the training set for deep networks in the kernel regime, 2022. URL https://arxiv.org/abs/2206.02927.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):2914, 2021.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning, 2020. URL https://arxiv.org/abs/1912.01198.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport, 2018. URL https://arxiv.org/abs/1805.09545.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, and Surya Ganguli. Deep learning versus kernel learning: An empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5850–5861, 2020.

Sara Fridovich-Keil, Raphael Gontijo Lopes, and Rebecca Roelofs. Spectral bias in practice: The role of function frequency in generalization. *Advances in Neural Information Processing Systems*, 35:7368–7382, 2022.

Amnon Geifman, Daniel Barzilai, Ronen Basri, and Meirav Galun. Controlling the inductive bias of wide neural networks by modifying the kernel's spectrum. *Transactions on Machine Learning Research*, 2024.

Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel, 2019. URL https://arxiv.org/abs/1909.05989.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020. URL https://arxiv.org/abs/1806.07572.

Shuai Jiang, Eric C. Cyr, Ben S. Southworth, and Alexey Voronin. On the convergence behavior of preconditioned gradient descent toward the rich learning regime. In *International Conference on Learning Representations*, 2026.

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent $*$. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12): 124002, December 2020. ISSN 1742-5468. doi: 10.1088/1742-5468/abc62b. URL http://dx.doi.org/10.1088/1742-5468/abc62b.

Tao Luo, Zheng Ma, Zhi-Qin John Xu, and Yaoyu Zhang. Theory of the frequency principle for general deep neural networks. *CSIAM Transactions on Applied Mathematics*, 2(3):484–507, 2021.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33), July 2018. ISSN 1091-6490. doi: 10.1073/pnas.1806579115. URL http://dx.doi.org/10.1073/pnas.1806579115.

Michael Murray, Hui Jin, Benjamin Bowman, and Guido Montufar. Characterizing the spectrum of the ntk via a power series expansion, 2023. URL https://arxiv.org/abs/2211.07844.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. ISBN 9781107038325. doi: 10.1017/CBO9781139814782.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks, 2019. URL https://arxiv.org/abs/1806.08734.

Basri Ronen, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5ac8bb8a7d745102a978c5f8ccdb61b8-Paper.pdf.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, 2014. URL https://arxiv.org/abs/1312.6120.

Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. Wiley, 1980. doi: 10.1002/9780470316481.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1–2):1–230, May 2015. ISSN 1935-8237. doi: 10.1561/2200000048. URL https://doi.org/10.1561/2200000048.

Roman Vershynin. *Concentration of Sums of Independent Random Variables*, pp. 11–37. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018a.

Roman Vershynin. *Concentration Without Independence*, pp. 98–126. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018b.

Nikhil Vyas, Yamini Bansal, and Preetum Nakkiran. Limitations of the ntk for understanding generalization in deep learning, 2022. URL https://arxiv.org/abs/2206.10012.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022.

Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *International Conference on Neural Information Processing*, pp. 264–274. Springer, 2019.

Yingzhen Yang. Sharp generalization for nonparametric regression in interpolation space by over-parameterized neural networks trained with preconditioned gradient descent and early stopping. *arXiv preprint arXiv:2407.11353*, 2024.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017. URL https://arxiv.org/abs/1611.03530.

John Zhi-Qin, John Xu, Luo Tao, Xiao Yanyang, and Ma Zheng. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5): 1746–1767, 2020.

## APPENDIX CONTENTS

The appendices collect the derivations and additional experiments deferred from the main text: the derivation of the limiting kernel and its Fourier spectrum on $S^1$ (Appendix A), the finite-sample concentration proofs (Appendix B), the frozen finite-width concentration proofs (Appendix C), the closed form and the expectation of the initial velocity of the evolving kernel (Appendix D), and additional experimental results, including the fixed-kernel preconditioning consistency check and all-width diagnostics (Appendix E).

## A DERIVATION OF THE LIMITING KERNEL ON $S^1$

This appendix derives the explicit form of the limiting neural tangent kernel on the circle for the model introduced in Section 3.1, and computes the corresponding Fourier eigenvalues of the integral operator $T$. These derivations justify Proposition 1 and Theorem 2.

### A.1 DECOMPOSITION OF THE EMPIRICAL NEURAL TANGENT KERNEL

Recall the network

$$f(x;\theta) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \sigma(w_i^\top x + b_i),$$

where $x \in S^1 \subset \mathbb{R}^2$, $\sigma(t) = t_+$, and $\theta = \{(a_i, w_i, b_i)\}_{i=1}^m$. For $x, y \in S^1$, the empirical neural tangent kernel is

$$\Theta_m(x,y) = \big\langle \nabla_\theta f(x;\theta), \nabla_\theta f(y;\theta) \big\rangle. \tag{18}$$

For each hidden unit $i$, define

$$u_i := w_i^\top x + b_i, \qquad v_i := w_i^\top y + b_i.$$

Then

$$\frac{\partial f(x;\theta)}{\partial a_i} = \frac{1}{\sqrt{m}} \sigma(u_i), \qquad \frac{\partial f(x;\theta)}{\partial w_i} = \frac{1}{\sqrt{m}} a_i \sigma'(u_i)\, x, \qquad \frac{\partial f(x;\theta)}{\partial b_i} = \frac{1}{\sqrt{m}} a_i \sigma'(u_i).$$

Substituting into equation 18 yields

$$\Theta_m(x,y) = \Theta_m^{(a)}(x,y) + \Theta_m^{(w)}(x,y) + \Theta_m^{(b)}(x,y), \tag{19}$$

where

$$\Theta_m^{(a)}(x,y) = \frac{1}{m} \sum_{i=1}^{m} \sigma(u_i)\sigma(v_i),$$

$$\Theta_m^{(w)}(x,y) = \frac{1}{m} \sum_{i=1}^{m} a_i^2 \sigma'(u_i)\sigma'(v_i)\, x^\top y,$$

$$\Theta_m^{(b)}(x,y) = \frac{1}{m} \sum_{i=1}^{m} a_i^2 \sigma'(u_i)\sigma'(v_i).$$

### A.2 INFINITE-WIDTH LIMIT

At initialization, the parameters satisfy

$$a_i \sim \mathcal{N}(0,1), \qquad w_i \sim \mathcal{N}(0, I_2), \qquad b_i(0) = 0,$$

independently across $i$. Since the hidden units are independent and identically distributed, the law of large numbers implies that $\Theta_m(x,y)$ converges almost surely, as $m \to \infty$, to a deterministic kernel $\Theta(x,y)$.

Because $b_i(0) = 0$, the limiting kernel is

$$\Theta(x,y) = \mathbb{E}\big[\sigma(u)\sigma(v)\big] + \big(x^\top y + 1\big)\mathbb{E}\big[\sigma'(u)\sigma'(v)\big], \tag{20}$$

where

$$u = w^\top x, \qquad v = w^\top y, \qquad w \sim \mathcal{N}(0, I_2).$$

We write

$$K_0(x,y) := \mathbb{E}\big[\sigma(u)\sigma(v)\big], \qquad K_1(x,y) := \mathbb{E}\big[\sigma'(u)\sigma'(v)\big],$$

so that

$$\Theta(x,y) = K_0(x,y) + \big(x^\top y + 1\big)K_1(x,y).$$

### A.3 RESTRICTION TO THE CIRCLE

Write

$$x(\varphi) = (\cos\varphi, \sin\varphi), \qquad x(\psi) = (\cos\psi, \sin\psi),$$

and let $\delta \in [0,\pi]$ denote the angle between $x(\varphi)$ and $x(\psi)$, so that

$$x(\varphi)^\top x(\psi) = \cos\delta.$$

By rotational invariance of the Gaussian measure, the kernel depends only on $\delta$. We therefore write $\Theta(\delta)$, $K_0(\delta)$, and $K_1(\delta)$.

To compute these quantities, we rotate coordinates so that

$$x = (1,0), \qquad y = (\cos\delta, \sin\delta).$$

If $w = (Z_1, Z_2)$ with $Z_1, Z_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, then

$$u = Z_1, \qquad v = Z_1\cos\delta + Z_2\sin\delta.$$

**Lemma 6.** *For $\delta \in [0,\pi]$,*

$$K_1(\delta) = \mathbb{P}(u > 0, v > 0) = \frac{\pi - \delta}{2\pi}. \tag{21}$$

*Proof.* Since $(u,v)$ is a centered bivariate Gaussian with correlation $\cos\delta$, the probability that both coordinates are positive is the standard Gaussian quadrant probability, which equals $(\pi-\delta)/(2\pi)$. □

**Lemma 7.** *For $\delta \in [0,\pi]$,*

$$K_0(\delta) = \mathbb{E}[\sigma(u)\sigma(v)] = \frac{1}{2\pi}\Big(\sin\delta + (\pi-\delta)\cos\delta\Big). \tag{22}$$

*Proof.* This is the standard arc-cosine covariance formula for ReLU features in two dimensions. □

*Proof of Proposition 1.* Using equation 20, equation 21, and equation 22, we obtain

$$\begin{aligned}\Theta(\delta) &= \frac{1}{2\pi}\Big(\sin\delta + (\pi-\delta)\cos\delta\Big) + \frac{1+\cos\delta}{2\pi}(\pi-\delta) \\ &= \frac{1}{2\pi}\Big(\sin\delta + 2(\pi-\delta)\cos\delta + (\pi-\delta)\Big).\end{aligned} \tag{23}$$

This is exactly equation 8. The convolution form of the operator follows from the fact that $\Theta$ depends only on the angular difference. □

### A.4 FOURIER COEFFICIENTS OF THE LIMITING KERNEL

Because the operator $T$ is a convolution operator, it is diagonalized by the real Fourier basis. If $\lambda_k$ denotes the eigenvalue associated with frequency $k$, then

$$\lambda_k = \frac{1}{\pi}\int_0^\pi \Theta(\delta)\cos(k\delta)\,d\delta, \qquad k \geq 0. \tag{24}$$

Substituting equation 23 gives

$$\lambda_k = \frac{1}{2\pi^2}\int_0^\pi \Big(\sin\delta + 2(\pi-\delta)\cos\delta + (\pi-\delta)\Big)\cos(k\delta)\,d\delta.$$

We first record a basic integration lemma.

**Lemma 8.** *For every integer $m \geq 1$,*

$$I_m := \int_0^\pi (\pi-\delta)\cos(m\delta)\,d\delta = \frac{1-(-1)^m}{m^2}. \tag{25}$$

*Proof.* Integrating by parts gives

$$I_m = \left[(\pi-\delta)\frac{\sin(m\delta)}{m}\right]_0^\pi + \frac{1}{m}\int_0^\pi \sin(m\delta)\,d\delta.$$

The boundary term vanishes, and

$$\frac{1}{m}\int_0^\pi \sin(m\delta)\,d\delta = \frac{1}{m}\left[-\frac{\cos(m\delta)}{m}\right]_0^\pi = \frac{1-(-1)^m}{m^2}.$$

□

*Proof of Theorem 2.* For $k \geq 2$, write

$$\lambda_k = \frac{1}{2\pi^2}(A_k + B_k + C_k),$$

where

$$\begin{aligned} A_k &:= \int_0^\pi \sin\delta\,\cos(k\delta)\,d\delta, \\ B_k &:= \int_0^\pi 2(\pi-\delta)\cos\delta\cos(k\delta)\,d\delta, \\ C_k &:= \int_0^\pi (\pi-\delta)\cos(k\delta)\,d\delta. \end{aligned}$$

Using

$$\sin\delta\cos(k\delta) = \frac{1}{2}\big(\sin((k+1)\delta) - \sin((k-1)\delta)\big),$$

we obtain

$$A_k = \frac{1}{2}\left(\frac{1-(-1)^{k+1}}{k+1} - \frac{1-(-1)^{k-1}}{k-1}\right).$$

Using

$$2\cos\delta\cos(k\delta) = \cos((k+1)\delta) + \cos((k-1)\delta),$$

together with Lemma 8, we obtain

$$B_k = \frac{1-(-1)^{k+1}}{(k+1)^2} + \frac{1-(-1)^{k-1}}{(k-1)^2}, \qquad C_k = \frac{1-(-1)^k}{k^2}.$$

If $k \geq 3$ is odd, then $k \pm 1$ are even, so $A_k = B_k = 0$, while $C_k = 2/k^2$. Hence

$$\lambda_k = \frac{1}{\pi^2 k^2}, \qquad k \geq 3 \text{ odd}. \tag{26}$$

If $k \geq 2$ is even, then $k \pm 1$ are odd, so $C_k = 0$, and

$$A_k = -\frac{2}{k^2-1}, \qquad B_k = \frac{4(k^2+1)}{(k^2-1)^2}.$$

Therefore

$$A_k + B_k = \frac{2(k^2+3)}{(k^2-1)^2},$$

and so

$$\lambda_k = \frac{k^2+3}{\pi^2(k^2-1)^2}, \qquad k \geq 2 \text{ even.} \tag{27}$$

It remains to compute $k = 0$ and $k = 1$ directly. For $k = 0$,

$$\lambda_0 = \frac{1}{2\pi^2}\left(\int_0^\pi \sin\delta\, d\delta + 2\int_0^\pi (\pi-\delta)\cos\delta\, d\delta + \int_0^\pi (\pi-\delta)\, d\delta\right) = \frac{1}{4} + \frac{3}{\pi^2}.$$

For $k = 1$,

$$\lambda_1 = \frac{1}{2\pi^2}\left(\int_0^\pi \sin\delta\cos\delta\, d\delta + 2\int_0^\pi (\pi-\delta)\cos^2\delta\, d\delta + \int_0^\pi (\pi-\delta)\cos\delta\, d\delta\right) = \frac{1}{4} + \frac{1}{\pi^2}.$$

This proves equation 9 and equation 10. □

# B Proofs of the finite-sample estimates

This appendix proves the finite-sample results from Section 4.1. We work with the truncated Fourier subspace $\mathcal{H}_K$, the feature matrix $\Phi$, the sampled kernel matrix $A$, the Gram matrix $G$, and the matrix $H$ defined in Section 4.1.

## B.1 Preliminaries

For each basis function in equation 5,

$$|\phi_p(\varphi)| \leq \sqrt{2} \qquad \text{for all } \varphi \in [0, 2\pi), \quad 1 \leq p \leq d. \tag{28}$$

We also write

$$\kappa := \sup_{\theta \in [0,2\pi)} |\Theta(\theta)| < \infty. \tag{29}$$

For each basis index $p$, let $\lambda_p$ denote the eigenvalue of the operator $T$ associated with $\phi_p$, so that

$$T\phi_p = \lambda_p \phi_p.$$

Define the finite-sample corrected eigenvalues by

$$\lambda_p^{(n)} = \left(1 - \frac{1}{n}\right)\lambda_p + \frac{\Theta(0)}{n}, \tag{30}$$

and let

$$\Lambda^{(n)} = \operatorname{diag}(\lambda_p^{(n)}).$$

## B.2 Expectation of $H$

*Proof of Proposition 3.* By definition,

$$H_{pq} = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \phi_p(\varphi_i)\Theta(\varphi_i - \varphi_j)\phi_q(\varphi_j). \tag{31}$$

Split the sum into the cases $i \neq j$ and $i = j$:

$$H_{pq} = \frac{1}{n^2}\sum_{i \neq j} \phi_p(\varphi_i)\Theta(\varphi_i - \varphi_j)\phi_q(\varphi_j) + \frac{1}{n^2}\sum_{i=1}^n \phi_p(\varphi_i)\Theta(0)\phi_q(\varphi_i).$$

For the diagonal terms,

$$\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\phi_p(\varphi_i)\Theta(0)\phi_q(\varphi_i)\right] = \frac{\Theta(0)}{n}\delta_{pq}.$$

For the off-diagonal terms, independence of $\varphi_i$ and $\varphi_j$ gives

$$\begin{aligned}\mathbb{E}[\phi_p(\varphi_i)\Theta(\varphi_i-\varphi_j)\phi_q(\varphi_j)] &= \int_0^{2\pi}\phi_p(\varphi)\left(\int_0^{2\pi}\Theta(\varphi-\psi)\phi_q(\psi)\,d\mu(\psi)\right)d\mu(\varphi)\\ &= \int_0^{2\pi}\phi_p(\varphi)(T\phi_q)(\varphi)\,d\mu(\varphi) = \lambda_q\delta_{pq}.\end{aligned}$$

Since there are $n(n-1)$ off-diagonal pairs, we conclude that

$$\mathbb{E}[H_{pq}] = \left(1-\frac{1}{n}\right)\lambda_q\delta_{pq} + \frac{\Theta(0)}{n}\delta_{pq} = \lambda_q^{(n)}\delta_{pq}.$$

Therefore $\mathbb{E}[H] = \Lambda^{(n)}$. □

### B.3 Concentration of the Gram matrix

*Proof of equation 11.* For fixed $p, q$,

$$G_{pq} = \frac{1}{n}\sum_{i=1}^{n} Z_i, \qquad Z_i := \phi_p(\varphi_i)\phi_q(\varphi_i).$$

The variables $Z_i$ are independent and identically distributed, with

$$\mathbb{E}[Z_i] = \delta_{pq}$$

by orthonormality of the Fourier basis.

Set

$$Y_i := Z_i - \delta_{pq}.$$

Then $\mathbb{E}[Y_i] = 0$ and

$$G_{pq} - \delta_{pq} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

By equation 28,

$$|Z_i| \le 2.$$

If $p \neq q$, then $\delta_{pq} = 0$, so $|Y_i| = |Z_i| \le 2$. If $p = q$, then

$$Y_i = \phi_p(\varphi_i)^2 - 1,$$

and since $0 \le \phi_p(\varphi_i)^2 \le 2$, we have $|Y_i| \le 1 \le 2$. Thus

$$|Y_i| \le 2$$

for all $i$.

Next,

$$\mathrm{Var}(Y_i) \le \mathbb{E}[Y_i^2] \le \mathbb{E}[Z_i^2].$$

Using again equation 28,

$$Z_i^2 = \phi_p(\varphi_i)^2\phi_q(\varphi_i)^2 \le 2\,\phi_p(\varphi_i)^2.$$

Taking expectations and using orthonormality gives

$$\mathbb{E}[Z_i^2] \le 2\int \phi_p(\varphi)^2\,d\mu(\varphi) = 2.$$

Hence

$$\mathrm{Var}(Y_i) \le 2.$$

We now apply Bernstein's inequality (Vershynin, 2018a) in the following form: if $Y_1, \dots, Y_n$ are independent, centered random variables such that

$$|Y_i| \le M, \qquad \mathrm{Var}(Y_i) \le \sigma^2$$

for every $i$, then for every $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| \ge \varepsilon\right) \le 2\exp\left(-\frac{n\varepsilon^2}{2\sigma^2 + \frac{2}{3}M\varepsilon}\right).$$

In our case, we have already shown that $\mathbb{E}[Y_i] = 0$, $|Y_i| \le 2$, and $\mathrm{Var}(Y_i) \le 2$. Thus we may take

$$M = 2, \qquad \sigma^2 = 2.$$

Since

$$G_{pq} - \delta_{pq} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

Bernstein's inequality for the sample mean yields

$$\mathbb{P}(|G_{pq} - \delta_{pq}| \ge \varepsilon) \le 2\exp\left(-\frac{n\varepsilon^2}{4 + \frac{4}{3}\varepsilon}\right).$$

Since $G - I_d$ is symmetric, it suffices to union bound over the $d(d+1)/2$ entries in the upper triangle. Therefore

$$\mathbb{P}(\|G - I_d\|_{\max} \ge \varepsilon) \le d(d+1)\exp\left(-\frac{n\varepsilon^2}{4 + \frac{4}{3}\varepsilon}\right),$$

which proves equation 11. □

### B.4 Concentration of $H$

*Proof of equation 12.* Fix $p, q$, and write $H_{pq} = F(\varphi_1, \dots, \varphi_n)$, where

$$F(\varphi_1, \dots, \varphi_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \phi_p(\varphi_i)\Theta(\varphi_i - \varphi_j)\phi_q(\varphi_j).$$

We apply McDiarmid's inequality (Vershynin, 2018b) in the following form: if $X_1, \dots, X_n$ are independent and $F(X_1, \dots, X_n)$ satisfies the bounded-differences condition

$$|F(x_1, \dots, x_s, \dots, x_n) - F(x_1, \dots, x_s', \dots, x_n)| \le c_s$$

for each $s$, then for every $\varepsilon > 0$,

$$\mathbb{P}(|F - \mathbb{E}F| \ge \varepsilon) \le 2\exp\left(-\frac{2\varepsilon^2}{\sum_{s=1}^{n} c_s^2}\right).$$

It therefore remains to estimate the effect of changing a single sample point, say $\varphi_s \mapsto \varphi_s'$, while keeping all others fixed. Only those summands in the double sum with $i = s$ or $j = s$ can change. There are at most $2n$ such summands.

By equation 28 and equation 29, each summand satisfies

$$|\phi_p(\cdot)\Theta(\cdot)\phi_q(\cdot)| \le \sqrt{2}\,\kappa\,\sqrt{2} = 2\kappa.$$

Hence, when $\varphi_s$ is replaced by $\varphi_s'$, a single summand can change by at most $4\kappa$.

Because at most $2n$ summands are affected, the total change in the double sum is bounded by

$$(2n)\cdot(4\kappa) = 8\kappa n.$$

Finally, since $F$ contains the prefactor $1/n^2$, the change in $F$ itself is bounded by

$$\frac{8\kappa n}{n^2} = \frac{8\kappa}{n}.$$

Thus we may take

$$c_s = \frac{8\kappa}{n} \qquad \text{for all } s = 1, \dots, n.$$

Therefore

$$\sum_{s=1}^{n} c_s^2 = n \left( \frac{8\kappa}{n} \right)^2 = \frac{64\kappa^2}{n}.$$

McDiarmid's inequality yields

$$\mathbb{P}(|H_{pq} - \mathbb{E}[H_{pq}]| \geq \varepsilon) \leq 2 \exp\left( -\frac{2\varepsilon^2}{64\kappa^2/n} \right) = 2 \exp\left( -\frac{n\varepsilon^2}{32\kappa^2} \right).$$

A union bound over all $d^2$ entries gives

$$\mathbb{P}\Big( \|H - \Lambda^{(n)}\|_{\max} \geq \varepsilon \Big) \leq 2d^2 \exp\left( -\frac{n\varepsilon^2}{32\kappa^2} \right),$$

as claimed. □

*Remark* 9. The entry $H_{pq}$ is a bounded second-order V-statistic, so Bernstein-type refinements are in principle available through concentration inequalities for bounded U/V-statistics. Since the McDiarmid bound already captures the correct $n^{-1/2}$ concentration scale and is sufficient for the results stated here, we do not pursue this refinement here.

### B.5 Control of the difference $A\Phi - \Phi\Lambda^{(n)}$

*Proof of equation 13.* Fix $i, p$. By definition,

$$(A\Phi)_{i,p} = \frac{1}{n} \sum_{j=1}^{n} \Theta(\varphi_i - \varphi_j) \phi_p(\varphi_j).$$

Split off the diagonal term $j = i$:

$$(A\Phi)_{i,p} = \frac{\Theta(0)}{n} \phi_p(\varphi_i) + \frac{1}{n} \sum_{j \neq i} \Theta(\varphi_i - \varphi_j) \phi_p(\varphi_j).$$

Condition on $\varphi_i$. For $j \neq i$, define

$$X_j := \Theta(\varphi_i - \varphi_j) \phi_p(\varphi_j).$$

Then $\{X_j\}_{j \neq i}$ are conditionally independent and identically distributed.

Their conditional mean is

$$\mathbb{E}[X_j \mid \varphi_i] = \int_0^{2\pi} \Theta(\varphi_i - \psi) \phi_p(\psi) \, d\mu(\psi) = (T\phi_p)(\varphi_i) = \lambda_p \phi_p(\varphi_i).$$

Therefore

$$\mathbb{E}[(A\Phi)_{i,p} \mid \varphi_i] = \frac{\Theta(0)}{n} \phi_p(\varphi_i) + \frac{n-1}{n} \lambda_p \phi_p(\varphi_i) = \lambda_p^{(n)} \phi_p(\varphi_i).$$

Now set

$$Y_j := X_j - \mathbb{E}[X_j \mid \varphi_i].$$

Conditional on $\varphi_i$, the variables $\{Y_j\}_{j \neq i}$ are independent, centered, and identically distributed.

We apply Bernstein's inequality (Vershynin, 2018a) in the following form: if $Y_1, \dots, Y_m$ are independent, centered random variables such that

$$|Y_j| \leq M, \qquad \mathrm{Var}(Y_j) \leq \sigma^2$$

for all $j$, then for every $\eta > 0$,

$$\mathbb{P}\left( \left| \frac{1}{m} \sum_{j=1}^{m} Y_j \right| \geq \eta \right) \leq 2 \exp\left( -\frac{m\eta^2}{2\sigma^2 + \frac{2}{3} M\eta} \right).$$

We now verify the assumptions conditionally on $\varphi_i$.

First, by equation 28 and equation 29,

$$|X_j| \le \kappa\sqrt{2} \le 2\kappa.$$

Also,

$$|\mathbb{E}[X_j \mid \varphi_i]| = \left| \int_0^{2\pi} \Theta(\varphi_i - \psi)\phi_p(\psi)\, d\mu(\psi) \right| \le \kappa \int_0^{2\pi} |\phi_p(\psi)|\, d\mu(\psi) \le \kappa \|\phi_p\|_{L^2(\mu)} = \kappa.$$

Hence

$$|Y_j| = |X_j - \mathbb{E}[X_j \mid \varphi_i]| \le 2\kappa + \kappa = 3\kappa.$$

Next,

$$\mathrm{Var}(Y_j \mid \varphi_i) = \mathrm{Var}(X_j \mid \varphi_i) \le \mathbb{E}[X_j^2 \mid \varphi_i].$$

Using again equation 28 and equation 29,

$$\mathbb{E}[X_j^2 \mid \varphi_i] = \int_0^{2\pi} \Theta(\varphi_i - \psi)^2 \phi_p(\psi)^2\, d\mu(\psi) \le \kappa^2 \int_0^{2\pi} \phi_p(\psi)^2\, d\mu(\psi) = \kappa^2.$$

Thus, conditionally on $\varphi_i$, Bernstein's inequality applies with

$$M = 3\kappa, \qquad \sigma^2 = \kappa^2.$$

Let $m := n - 1$. Applying Bernstein to the conditional sample mean

$$\frac{1}{m}\sum_{j\neq i} Y_j = \frac{1}{m}\sum_{j\neq i} X_j - \lambda_p \phi_p(\varphi_i),$$

we obtain

$$\mathbb{P}\left( \left| \frac{1}{m}\sum_{j\neq i} X_j - \lambda_p \phi_p(\varphi_i) \right| \ge \eta \;\middle|\; \varphi_i \right) \le 2\exp\left( -\frac{m\eta^2}{2\kappa^2 + 2\kappa\eta} \right).$$

Now

$$(A\Phi)_{i,p} - \lambda_p^{(n)} \phi_p(\varphi_i) = \frac{m}{n} \left( \frac{1}{m}\sum_{j\neq i} X_j - \lambda_p \phi_p(\varphi_i) \right).$$

Thus the event

$$\left| (A\Phi)_{i,p} - \lambda_p^{(n)} \phi_p(\varphi_i) \right| \ge \varepsilon$$

implies

$$\left| \frac{1}{m}\sum_{j\neq i} X_j - \lambda_p \phi_p(\varphi_i) \right| \ge \frac{n}{m}\varepsilon.$$

Hence

$$\mathbb{P}\Big( \left| (A\Phi)_{i,p} - \lambda_p^{(n)} \phi_p(\varphi_i) \right| \ge \varepsilon \;\Big|\; \varphi_i \Big) \le 2\exp\left( -\frac{m\left(\frac{n}{m}\varepsilon\right)^2}{2\kappa^2 + 2\kappa\left(\frac{n}{m}\varepsilon\right)} \right).$$

Removing the conditioning and simplifying gives

$$\mathbb{P}\Big( \left| (A\Phi)_{i,p} - \lambda_p^{(n)} \phi_p(\varphi_i) \right| \ge \varepsilon \Big) \le 2\exp\left( -\frac{n^2\varepsilon^2}{2\kappa^2(n-1) + 2\kappa n\varepsilon} \right).$$

Since $n - 1 \le n$, we obtain

$$\mathbb{P}\Big( \left| (A\Phi)_{i,p} - \lambda_p^{(n)} \phi_p(\varphi_i) \right| \ge \varepsilon \Big) \le 2\exp\left( -\frac{n\varepsilon^2}{2\kappa^2 + 2\kappa\varepsilon} \right).$$

Finally, taking a union bound over all $nd$ pairs $(i, p)$ yields

$$\mathbb{P}\Big( \|A\Phi - \Phi\Lambda^{(n)}\|_{\max} \ge \varepsilon \Big) \le 2nd\exp\left( -\frac{n\varepsilon^2}{2\kappa^2 + 2\kappa\varepsilon} \right),$$

which proves equation 13. □

### B.6 Proof of the simultaneous high-probability bound

**Corollary 10** (High-probability bounds)**.** *Let $\delta \in (0,1)$, and define*

$$u^{(G)}_{n,d,\delta} := \log\left(\frac{3d(d+1)}{\delta}\right), \tag{32}$$

$$u^{(H)}_{n,d,\delta} := \log\left(\frac{6d^2}{\delta}\right), \tag{33}$$

*and*

$$u^{(A)}_{n,d,\delta} := \log\left(\frac{6nd}{\delta}\right). \tag{34}$$

*Now set*

$$\alpha_{n,d,\delta} := 2\sqrt{\frac{u^{(G)}_{n,d,\delta}}{n}} + \frac{2}{3}\frac{u^{(G)}_{n,d,\delta}}{n}, \tag{35}$$

$$\beta_{n,d,\delta} := \kappa\sqrt{\frac{32}{n}u^{(H)}_{n,d,\delta}}, \tag{36}$$

*and*

$$\gamma_{n,d,\delta} := \kappa\sqrt{\frac{2}{n}u^{(A)}_{n,d,\delta}} + \kappa\frac{u^{(A)}_{n,d,\delta}}{n}. \tag{37}$$

*Then, with probability at least $1-\delta$,*

$$\|G - I_d\|_{\max} \le \alpha_{n,d,\delta}, \qquad \|H - \Lambda^{(n)}\|_{\max} \le \beta_{n,d,\delta}, \qquad \|A\Phi - \Phi\Lambda^{(n)}\|_{\max} \le \gamma_{n,d,\delta}. \tag{38}$$

*Proof of Corollary 10.* Set

$$u^{(G)}_{n,d,\delta} = \log\left(\frac{3d(d+1)}{\delta}\right), \qquad u^{(H)}_{n,d,\delta} = \log\left(\frac{6d^2}{\delta}\right), \qquad u^{(A)}_{n,d,\delta} = \log\left(\frac{6nd}{\delta}\right).$$

Define

$$\alpha_{n,d,\delta} = 2\sqrt{\frac{u^{(G)}_{n,d,\delta}}{n}} + \frac{2}{3}\frac{u^{(G)}_{n,d,\delta}}{n},$$

$$\beta_{n,d,\delta} = \kappa\sqrt{\frac{32}{n}u^{(H)}_{n,d,\delta}}, \qquad \gamma_{n,d,\delta} = \kappa\sqrt{\frac{2}{n}u^{(A)}_{n,d,\delta}} + \kappa\frac{u^{(A)}_{n,d,\delta}}{n}.$$

We first recall the standard Bernstein inversion: if

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| \ge t\right) \le 2\exp\left(-\frac{nt^2}{2\sigma^2 + \frac{2}{3}Mt}\right),$$

then choosing

$$t = \sqrt{\frac{2\sigma^2 u}{n}} + \frac{Mu}{3n}$$

ensures

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| \ge t\right) \le 2e^{-u}.$$

For the Gram matrix bound equation 11, the Bernstein proof used

$$\sigma^2 = 2, \qquad M = 2.$$

Thus, with

$$u = u^{(G)}_{n,d,\delta},$$

the choice

$$\alpha_{n,d,\delta} = \sqrt{\frac{2\cdot 2\,u}{n}} + \frac{2u}{3n} = 2\sqrt{\frac{u}{n}} + \frac{2u}{3n}$$

yields

$$\mathbb{P}(\|G - I_d\|_{\max} \geq \alpha_{n,d,\delta}) \leq d(d+1)e^{-u^{(G)}_{n,d,\delta}} = \frac{\delta}{3}.$$

For the compressed operator, the McDiarmid estimate equation 12 gives

$$\mathbb{P}\Big(\|H - \Lambda^{(n)}\|_{\max} \geq \varepsilon\Big) \leq 2d^2 \exp\left(-\frac{n\varepsilon^2}{32\kappa^2}\right).$$

Substituting $\varepsilon = \beta_{n,d,\delta}$ gives

$$\frac{n\beta^2_{n,d,\delta}}{32\kappa^2} = u^{(H)}_{n,d,\delta},$$

and therefore

$$\mathbb{P}\Big(\|H - \Lambda^{(n)}\|_{\max} \geq \beta_{n,d,\delta}\Big) \leq 2d^2 e^{-u^{(H)}_{n,d,\delta}} = \frac{\delta}{3}.$$

For the approximate eigen-action bound equation 13, the Bernstein proof used

$$\sigma^2 = \kappa^2, \qquad M = 3\kappa.$$

Hence, with

$$u = u^{(A)}_{n,d,\delta},$$

the standard Bernstein inversion gives the threshold

$$\sqrt{\frac{2\kappa^2 u}{n}} + \frac{3\kappa u}{3n} = \kappa\sqrt{\frac{2u}{n}} + \kappa\frac{u}{n},$$

which is precisely $\gamma_{n,d,\delta}$. Therefore

$$\mathbb{P}\Big(\|A\Phi - \Phi\Lambda^{(n)}\|_{\max} \geq \gamma_{n,d,\delta}\Big) \leq 2nd\, e^{-u^{(A)}_{n,d,\delta}} = \frac{\delta}{3}.$$

A final union bound gives

$$\mathbb{P}\Big(\|G - I_d\|_{\max} \leq \alpha_{n,d,\delta},\ \|H - \Lambda^{(n)}\|_{\max} \leq \beta_{n,d,\delta},\ \|A\Phi - \Phi\Lambda^{(n)}\|_{\max} \leq \gamma_{n,d,\delta}\Big) \geq 1 - \delta,$$

which is exactly equation 38. □

# C Proofs of the finite-width frozen-kernel estimates

This appendix proves the finite-width frozen-kernel results stated in Section 4.2. We use the notation introduced in Section 3.1. In particular, $T_0^{(m)}$ denotes the frozen finite-width tangent operator at initialization, $T_\infty$ the infinite-width limiting operator, $\mathcal{H}_K$ the truncated Fourier subspace, $P_K$ the orthogonal projection onto $\mathcal{H}_K$, and

$$B_m^{(K)} = P_K T_0^{(m)} P_K, \qquad \Lambda_K = P_K T_\infty P_K.$$

Since the input space is the continuum $S^1$, there is no sampling randomness in this appendix. The only randomness comes from the finite number of initialized neurons.

## C.1 One-neuron tangent feature and kernel

Fix one neuron $r$, and write

$$u_r(\theta) := w_r^\top x(\theta) + b_r.$$

At initialization, $b_r = 0$, so $u_r(\theta) = w_r^\top x(\theta)$.

We compute the tangent feature of the $r$-th neuron contribution

$$a_r \sigma(u_r(\theta))$$

with respect to the parameter block $(a_r, w_r, b_r)$. First,

$$\partial_{a_r}\big(a_r\sigma(u_r(\theta))\big) = \sigma(u_r(\theta)).$$

Next, by the chain rule,

$$\nabla_{w_r}\big(a_r\sigma(u_r(\theta))\big) = a_r\sigma'(u_r(\theta))\nabla_{w_r}u_r(\theta).$$

Since

$$u_r(\theta) = w_r^\top x(\theta) + b_r,$$

we have

$$\nabla_{w_r}u_r(\theta) = x(\theta).$$

For ReLU,

$$\sigma'(z) = \mathbf{1}_{\{z>0\}}$$

away from $z = 0$, which is sufficient almost surely under the Gaussian initialization. Hence

$$\nabla_{w_r}\big(a_r\sigma(u_r(\theta))\big) = a_r\mathbf{1}_{\{u_r(\theta)>0\}}x(\theta).$$

Similarly,

$$\partial_{b_r}\big(a_r\sigma(u_r(\theta))\big) = a_r\sigma'(u_r(\theta))\partial_{b_r}u_r(\theta) = a_r\mathbf{1}_{\{u_r(\theta)>0\}}.$$

Thus the one-neuron tangent feature is

$$\psi_r(\theta) = \begin{pmatrix} \sigma(u_r(\theta)) \\ a_r\mathbf{1}_{\{u_r(\theta)>0\}}\cos\theta \\ a_r\mathbf{1}_{\{u_r(\theta)>0\}}\sin\theta \\ a_r\mathbf{1}_{\{u_r(\theta)>0\}} \end{pmatrix} \in \mathbb{R}^4. \tag{39}$$

The corresponding one-neuron kernel is

$$\Theta^{(r)}(\theta,\eta) = \psi_r(\theta)^\top\psi_r(\eta). \tag{40}$$

Substituting equation 39 gives

$$\begin{aligned}\Theta^{(r)}(\theta,\eta) &= \sigma(u_r(\theta))\sigma(u_r(\eta)) \\ &\quad + a_r^2\mathbf{1}_{\{u_r(\theta)>0\}}\mathbf{1}_{\{u_r(\eta)>0\}}\big(\cos\theta\cos\eta + \sin\theta\sin\eta + 1\big).\end{aligned} \tag{41}$$

Using

$$\cos\theta\cos\eta + \sin\theta\sin\eta = \cos(\theta-\eta),$$

we obtain

$$\Theta^{(r)}(\theta,\eta) = \sigma(u_r(\theta))\sigma(u_r(\eta)) + a_r^2\mathbf{1}_{\{u_r(\theta)>0\}}\mathbf{1}_{\{u_r(\eta)>0\}}\big(\cos(\theta-\eta)+1\big). \tag{42}$$

Let $T^{(r)}$ denote the associated one-neuron integral operator,

$$(T^{(r)}g)(\theta) := \int_0^{2\pi}\Theta^{(r)}(\theta,\eta)g(\eta)\,d\mu(\eta). \tag{43}$$

Then, by the one-neuron decomposition of Section 4.2,

$$T_0^{(m)} = \frac{1}{m}\sum_{r=1}^m T^{(r)}.$$

### C.2 Projected one-neuron matrices

Using the ordered basis

$$(\phi_0, \phi_{1,c}, \phi_{1,s}, \ldots, \phi_{K,c}, \phi_{K,s})$$

of $\mathcal{H}_K$, the entries of $B_m^{(K)}$ are

$$(B_m^{(K)})_{pq} = \langle\phi_p, T_0^{(m)}\phi_q\rangle_{L^2(\mu)}. \tag{44}$$

Since $T_0^{(m)} = \frac{1}{m}\sum_{r=1}^m T^{(r)}$,

$$(B_m^{(K)})_{pq} = \frac{1}{m}\sum_{r=1}^m\langle\phi_p, T^{(r)}\phi_q\rangle.$$

Define

$$\xi_{r,pq} := \langle \phi_p, T^{(r)} \phi_q \rangle. \tag{45}$$

Then

$$(B_m^{(K)})_{pq} = \frac{1}{m} \sum_{r=1}^{m} \xi_{r,pq}. \tag{46}$$

Equivalently, if $Z_r$ denotes the $d \times d$ random matrix defined by

$$(Z_r)_{pq} := \xi_{r,pq},$$

then

$$B_m^{(K)} = \frac{1}{m} \sum_{r=1}^{m} Z_r. \tag{47}$$

Since the neurons are initialized independently and identically, the matrices $Z_1, \ldots, Z_m$ are i.i.d.

### C.3 Expectation of the projected operator

*Proof of equation 14.* From equation 47,

$$\mathbb{E}[B_m^{(K)}] = \frac{1}{m} \sum_{r=1}^{m} \mathbb{E}[Z_r] = \mathbb{E}[Z_1].$$

The $(p, q)$-entry is

$$\mathbb{E}[(Z_1)_{pq}] = \mathbb{E}[\xi_{1,pq}] = \mathbb{E}[\langle \phi_p, T^{(1)} \phi_q \rangle].$$

By linearity of expectation,

$$\mathbb{E}[\xi_{1,pq}] = \langle \phi_p, \mathbb{E}[T^{(1)}] \phi_q \rangle = \langle \phi_p, T_\infty \phi_q \rangle.$$

Since $T_\infty$ is diagonal in the Fourier basis,

$$\langle \phi_p, T_\infty \phi_q \rangle = \lambda_q \delta_{pq}.$$

Therefore

$$\mathbb{E}[B_m^{(K)}] = \Lambda_K.$$

□

### C.4 Sub-exponential bound for one projected entry

We next control $\xi_{r,pq}$. By definition,

$$\xi_{r,pq} = \int_0^{2\pi} \int_0^{2\pi} \phi_p(\theta) \Theta^{(r)}(\theta, \eta) \phi_q(\eta) \, d\mu(\eta) \, d\mu(\theta). \tag{48}$$

Taking absolute values gives

$$|\xi_{r,pq}| \leq \int_0^{2\pi} \int_0^{2\pi} |\phi_p(\theta)| |\Theta^{(r)}(\theta, \eta)| |\phi_q(\eta)| \, d\mu(\eta) \, d\mu(\theta).$$

Since every retained Fourier basis function satisfies

$$\|\phi_p\|_\infty \leq \sqrt{2},$$

we obtain

$$|\xi_{r,pq}| \leq 2 \int_0^{2\pi} \int_0^{2\pi} |\Theta^{(r)}(\theta, \eta)| \, d\mu(\eta) \, d\mu(\theta).$$

We now bound the kernel. Since

$$u_r(\theta) = w_r^\top x(\theta), \qquad \|x(\theta)\| = 1,$$

Cauchy–Schwarz gives

$$|u_r(\theta)| \leq \|w_r\|.$$

Therefore

$$0 \le \sigma(u_r(\theta)) \le |u_r(\theta)| \le \|w_r\|,$$

and hence

$$0 \le \sigma(u_r(\theta))\sigma(u_r(\eta)) \le \|w_r\|^2.$$

For the derivative and bias term, we use

$$\mathbf{1}_{\{u_r(\theta)>0\}}\mathbf{1}_{\{u_r(\eta)>0\}} \le 1$$

and

$$0 \le \cos(\theta-\eta)+1 \le 2.$$

Thus

$$a_r^2\mathbf{1}_{\{u_r(\theta)>0\}}\mathbf{1}_{\{u_r(\eta)>0\}}(\cos(\theta-\eta)+1) \le 2a_r^2.$$

Using equation 42, we conclude that

$$|\Theta^{(r)}(\theta,\eta)| \le \|w_r\|^2 + 2a_r^2.$$

Since $\mu$ is a probability measure,

$$\iint (\|w_r\|^2 + 2a_r^2)\, d\mu(\eta)d\mu(\theta) = \|w_r\|^2 + 2a_r^2.$$

Therefore

$$|\xi_{r,pq}| \le 2\|w_r\|^2 + 4a_r^2. \tag{49}$$

Define

$$Y_r := 2\|w_r\|^2 + 4a_r^2.$$

Then

$$|\xi_{r,pq}| \le Y_r \qquad \text{for all } p,q.$$

We now show that $Y_r$ is sub-exponential with an explicit parameter. Since $w_r \sim \mathcal{N}(0, I_2)$, its squared Euclidean norm is the sum of two independent standard Gaussian squares, and therefore $\|w_r\|^2$ follows a chi-squared distribution with two degrees of freedom $\|w_r\|^2 \sim \chi_2^2$.

Similarly, since $a_r \sim \mathcal{N}(0,1)$, its square follows a chi-squared distribution with one degree of freedom $a_r^2 \sim \chi_1^2$ and these variables are independent, for $t < 1/8$,

$$\mathbb{E}[e^{tY_r}] = \mathbb{E}[e^{2t\|w_r\|^2}]\mathbb{E}[e^{4ta_r^2}].$$

If $X \sim \chi_\nu^2$, then

$$\mathbb{E}[e^{sX}] = (1-2s)^{-\nu/2}, \qquad s < 1/2.$$

Hence

$$\mathbb{E}[e^{2t\|w_r\|^2}] = (1-4t)^{-1}, \qquad \mathbb{E}[e^{4ta_r^2}] = (1-8t)^{-1/2}.$$

Thus

$$\mathbb{E}[e^{tY_r}] = (1-4t)^{-1}(1-8t)^{-1/2}, \qquad t < 1/8. \tag{50}$$

Choosing $t = 1/16$, we obtain

$$\mathbb{E}[e^{Y_r/16}] = \left(\frac{3}{4}\right)^{-1}\left(\frac{1}{2}\right)^{-1/2} = \frac{4}{3}\sqrt{2} < 2.$$

Therefore, by the standard definition of the $\psi_1$-norm,

$$\|Y_r\|_{\psi_1} \le 16. \tag{51}$$

Since $|\xi_{r,pq}| \le Y_r$,

$$\|\xi_{r,pq}\|_{\psi_1} \le 16.$$

Finally,

$$\|\xi_{r,pq} - \mathbb{E}[\xi_{r,pq}]\|_{\psi_1} \le \|\xi_{r,pq}\|_{\psi_1} + |\mathbb{E}[\xi_{r,pq}]|.$$

Using $|\mathbb{E}[\xi_{r,pq}]| \le \mathbb{E}|\xi_{r,pq}|$ and $\mathbb{E}|X| \le \|X\|_{\psi_1}$, we obtain

$$\|\xi_{r,pq} - \mathbb{E}[\xi_{r,pq}]\|_{\psi_1} \le 32. \tag{52}$$

### C.5 Entrywise and blockwise concentration

*Proof of equation 15.* Fix $p, q$, and define

$$X_{r,pq} := \xi_{r,pq} - \mathbb{E}[\xi_{r,pq}].$$

Then $X_{1,pq}, \dots, X_{m,pq}$ are i.i.d., centered, and by equation 52,

$$\|X_{r,pq}\|_{\psi_1} \le 32.$$

Moreover,

$$(B_m^{(K)})_{pq} - (\Lambda_K)_{pq} = \frac{1}{m}\sum_{r=1}^{m} X_{r,pq}.$$

We apply the standard Bernstein inequality for centered i.i.d. sub-exponential random variables: if $X_1, \dots, X_m$ satisfy $\|X_i\|_{\psi_1} \le L$, then there exists a universal constant $c > 0$ such that

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{r=1}^{m} X_r\right| \ge \varepsilon\right) \le 2\exp\left(-cm\min\left(\frac{\varepsilon^2}{L^2}, \frac{\varepsilon}{L}\right)\right).$$

Applying this with $L = 32$, we obtain

$$\mathbb{P}\Big(\Big|(B_m^{(K)})_{pq} - (\Lambda_K)_{pq}\Big| \ge \varepsilon\Big) \le 2\exp\left(-cm\min\left(\frac{\varepsilon^2}{32^2}, \frac{\varepsilon}{32}\right)\right). \tag{53}$$

There are $d^2 = (2K+1)^2$ entries. By the union bound,

$$\mathbb{P}\Big(\|B_m^{(K)} - \Lambda_K\|_{\max} \ge \varepsilon\Big) \le \sum_{p,q=1}^{d} \mathbb{P}\Big(\Big|(B_m^{(K)})_{pq} - (\Lambda_K)_{pq}\Big| \ge \varepsilon\Big).$$

Using equation 53,

$$\mathbb{P}\Big(\|B_m^{(K)} - \Lambda_K\|_{\max} \ge \varepsilon\Big) \le 2d^2\exp\left(-cm\min\left(\frac{\varepsilon^2}{32^2}, \frac{\varepsilon}{32}\right)\right).$$

Since $d = 2K+1$, this proves equation 15. □

## D Initial velocity of the evolving kernel

### D.1 Motivation

Section 5.3 reports spectral strength growing on subspaces that carry exactly zero target mass, and growing most on the two the parameterization represents directly. There are two possible explanations: the kernel grows wherever there is error left to fit, or it grows on the directions the architecture can move most easily. The mode-absent controls separate them empirically. This appendix tests whether the kernel's equation of motion separates them as well.

We compute $\partial_t \Theta_t^{(m)}$ and ask what it depends on. At $t = 0$ the weights are the Gaussian draw equation 2 and every bias is zero, while at any later time $\theta(t)$ has no tractable law, so the initial velocity is the only point at which the calculation is available. The question is whether the first-order term already shows the low-frequency preference the experiments show. Section D.3 bounds what the parameterization can move, Section D.4 determines what the velocity depends on, and Section D.5 computes its size and sign.

### D.2 Setup

Throughout, $\langle \cdot, \cdot \rangle$ is the $L^2(\mu)$ inner product of Section 3.1. For neuron $i$ write

$$u_i = w_i^\top x + b_i, \qquad s_i = \sigma(u_i), \qquad h_i = \mathbf{1}\{u_i > 0\}, \qquad \rho_i = \|w_i\|, \qquad w_i = \rho_i \hat{w}_i,$$

so that $u_i(\varphi) = \rho_i \cos(\varphi - \psi_i) + b_i$ with $\hat{w}_i = (\cos\psi_i, \sin\psi_i)$, as in equation 16. Let $c(\varphi, \varphi') = 1 + x(\varphi)^\top x(\varphi')$, which is nonnegative. The empirical kernel of Appendix A.1 is then

$$\Theta_t^{(m)}(\varphi, \varphi') = \frac{1}{m} \sum_{i=1}^{m} \Big[ s_i(\varphi)\, s_i(\varphi') + a_i^2\, h_i(\varphi)\, h_i(\varphi')\, c(\varphi, \varphi') \Big]. \tag{54}$$

We use one piece of notation throughout. For a function $f \in L^2(\mu)$ and a neuron $i$, let

$$V_i[f](\varphi) = \int f(\eta)\, h_i(\eta)\, c(\eta, \varphi)\, d\mu(\eta) = \langle f, h_i \rangle + \langle f, h_i x \rangle^\top x(\varphi), \tag{55}$$

where $\langle f, h_i x \rangle \in \mathbb{R}^2$ is the pair $(\langle f, h_i \cos\rangle, \langle f, h_i \sin\rangle)$. By the second expression $V_i[f]$ is an affine function of $x$, so it always lies in $\operatorname{span}\{\phi_0, \phi_{1,c}, \phi_{1,s}\} = \mathcal{H}_1$, whatever $f$ is. This is what Section D.3 uses.

Differentiating $\mathcal{L}(\theta) = \frac{1}{2}\|f_\theta - y\|^2_{L^2(\mu)}$ with the tangent features of Appendix C.1 gives, for a residual $r = f_\theta - y$, the gradient-flow equations

$$\begin{gathered} \dot{a}_i = -m^{-1/2} \langle r, s_i \rangle, \qquad \dot{w}_i = -a_i\, m^{-1/2} \langle r, h_i x \rangle, \qquad \dot{b}_i = -a_i\, m^{-1/2} \langle r, h_i \rangle, \\ \text{hence} \qquad \dot{u}_i = -a_i\, m^{-1/2}\, V_i[r]. \end{gathered} \tag{56}$$

For a continuous $g$ on the circle write $q_g(\Theta) = \iint g\, \Theta\, g\, d\mu\, d\mu$, so that the strength equation 6 is $\mu_{\mathcal{F}_k}(t) = d_k^{-1} \sum_g q_g(\Theta_t^{(m)})$ over the orthonormal basis of $\mathcal{F}_k$. Substituting equation 54 and using equation 55 with $f = g$,

$$q_g\big(\Theta_\theta^{(m)}\big) = \frac{1}{m} \sum_{i=1}^{m} \Big[ \langle g, s_i \rangle^2 + a_i^2\, \langle g, h_i V_i[g] \rangle \Big]. \tag{57}$$

From Section D.5 onwards everything is evaluated at initialization, where $b_i = 0$. There $h_i$ is the indicator of the half circle $|\varphi - \psi_i| < \pi/2$, its two endpoints are $\varphi_i^\varepsilon = \psi_i + \varepsilon\pi/2$ for $\varepsilon \in \{+1, -1\}$, and $s_i = \rho_i \cos(\varphi - \psi_i) h_i$.

For ReLU the Hessian of $f$ contains $\sigma'' = \delta$, so $\partial_t \Theta_t^{(m)}$ is a distribution in $\varphi$ and cannot be evaluated pointwise. The quadratic forms $q_g$ are nevertheless differentiable, and the reason is geometric: $\langle g, h_i \rangle$ and $\langle g, s_i \rangle$ are integrals over the arc on which $h_i = 1$, and the delta function is the motion of that arc's two endpoints. We record this once, since both theorems below need it.

**Differentiability and domination.** Fix a continuous $g$. At $t = 0$ the map $\theta \mapsto q_g(\Theta_\theta^{(m)})$ is differentiable. By equation 57 it suffices that $\langle g, h_i \rangle$, $\langle g, h_i x \rangle$ and $\langle g, s_i \rangle$ be differentiable, and each is an integral of a continuous function over $[\varphi_i^-, \varphi_i^+]$, whose endpoints are smooth in $(\rho_i, \psi_i, b_i)$ as long as $|b_i| < \rho_i$; at $t = 0$ we have $b_i = 0 < \rho_i$ almost surely. Differentiating $u_i(\varphi_i^\varepsilon(t), t) = 0$ gives $\partial_\varphi u_i(\varphi_i^\varepsilon)\dot{\varphi}_i^\varepsilon + \dot{u}_i(\varphi_i^\varepsilon) = 0$, and $\partial_\varphi u_i(\varphi_i^\varepsilon) = -\varepsilon\rho_i$ at $b_i = 0$, so with equation 56 the arc endpoints move at

$$\dot{\varphi}_i^\varepsilon = -\varepsilon\, a_i\, m^{-1/2}\, V_i[r](\varphi_i^\varepsilon)/\rho_i. \tag{58}$$

Bounding the partial derivatives of equation 57 gives a constant $C_g$ depending only on $\|g\|_\infty$ with

$$\Big| \frac{d}{dt} q_g\big(\Theta_t^{(m)}\big)\Big|_{t=0} \Big| \le \frac{C_g\, \|r\|}{m^{3/2}} \sum_{i=1}^{m} \big(1 + |a_i|^3\big)\big(\rho_i + \rho_i^{-1}\big). \tag{59}$$

Indeed $0 \le s_i \le \rho_i$, $0 \le h_i \le 1$ and $\|x\| = 1$ give $|\langle g, s_i \rangle| \le \|g\|_\infty \rho_i$ and $|V_i[g]| \le 2\|g\|_\infty$, so $|\partial q_g / \partial a_i| \le C_g |a_i|/m$; in $\partial q_g/\partial w_i$ and $\partial q_g/\partial b_i$ the $\langle g, s_i \rangle^2$ part contributes at most $C_g \rho_i / m$, while the $a_i^2 \langle g, h_i V_i[g] \rangle$ part moves only through the endpoints and picks up the factor $\rho_i^{-1}$ of equation 58, contributing at most $C_g a_i^2/(m\rho_i)$. By equation 56 the velocities obey $|\dot{a}_i| \le m^{-1/2}\|r\|\rho_i$ and $\|\dot{w}_i\|, |\dot{b}_i| \le m^{-1/2}|a_i|\, \|r\|$; multiplying and summing gives equation 59.

### D.3 Confinement of the parameter motion

A neuron has a bias and a two-dimensional input embedding, so its pre-activation is an affine function of $x(\varphi) = (\cos\varphi, \sin\varphi)$ and nothing else. Gradient descent moves the coefficients of that affine function, and cannot give the pre-activation a shape the parameterization does not have.

**Lemma 11** (Frequency compression of the parameter motion)**.** *For every residual $r \in L^2(\mu)$, every time $t$ and every neuron, $\dot{u}_i \in \mathcal{H}_1 = \operatorname{span}\{\phi_0, \phi_{1,c}, \phi_{1,s}\}$.*

*Proof.* By equation 56, $\dot{u}_i = -a_i m^{-1/2} V_i[r]$, and $V_i[r]$ is affine in $x$ by equation 55. □

The confinement comes from the architecture, not from the loss, the target or the training time. This is the exact form of the claim made informally in Section 5.3, and it is why $k = 0$ and $k = 1$ are the distinguished frequencies there.

The lemma constrains $\dot{u}_i$, not $\partial_t \Theta_t^{(m)}$. Both $\sigma$ and $\mathbf{1}\{\cdot > 0\}$ are nonlinear, so a change of $u_i$ inside $\mathcal{H}_1$ changes $s_i$ and $h_i$ at every frequency, and through equation 54 the kernel moves in every block. The motion enters through a frequency-one channel but does not stay there.

### D.4 Independence from the target

The kernel equation 54 determines what the velocity can depend on. The output weights enter it only through $a_i^2$, so flipping the sign of every $a_i$ leaves the kernel unchanged while reversing the flow that moves it. This needs no computation of the velocity itself.

**Proposition 12** (Symmetry in the output weights)**.** *Fix $W = (w_i)$ and $b = (b_i)$, and fix a residual $r$ that does not depend on $a$. Let $\mathcal{V}_g(a; r)$ denote the value of $\frac{d}{dt} q_g(\Theta_t^{(m)})$ produced by equation 56 at that configuration. Then $\mathcal{V}_g$ is linear in $r$ and odd in $a$.*

*Proof.* Linearity is immediate, since $r$ enters equation 56 only through the inner products $\langle r, s_i \rangle$, $\langle r, h_i \rangle$ and $\langle r, h_i x \rangle$.

For the parity, $s_i$ and $h_i$ depend on $(w_i, b_i)$ only, so by equation 57 the function $q_g$ is *even* in $a$. Hence $\partial q_g / \partial a_i$ is odd in $a$, while $\partial q_g / \partial w_i$ and $\partial q_g / \partial b_i$ are even. By equation 56, $\dot{a}_i$ does not involve $a$ and is even, while $\dot{w}_i$ and $\dot{b}_i$ carry a factor $a_i$ and are odd. In

$$\mathcal{V}_g(a; r) = \sum_{i=1}^{m} \Big[ \frac{\partial q_g}{\partial a_i} \dot{a}_i + \Big( \frac{\partial q_g}{\partial w_i} \Big)^{\top} \dot{w}_i + \frac{\partial q_g}{\partial b_i} \dot{b}_i \Big]$$

every term pairs an odd factor with an even one, so the sum is odd in $a$. □

**Theorem 13** (The expected initial velocity does not depend on the target)**.** *Let $y \in L^2(\mu)$ be deterministic, let $\theta(0)$ be drawn from equation 2, and let $r_0 = f_0 - y$ with $f_0 = f_m(\,\cdot\,; \theta(0))$. For every continuous $g$ the expectation $\mathbb{E}\big[\frac{d}{dt} q_g(\Theta_t^{(m)})\big|_{t=0}\big]$ exists and equals $\mathbb{E}[\mathcal{V}_g(a; f_0)]$, the value obtained when the target is replaced by zero. In particular it does not depend on $y$.*

*Proof.* Existence follows from the domination bound equation 59, which involves $y$ only through $\|r_0\| \le \|y\| + m^{-1/2} \sum_j |a_j| \rho_j$, all Gaussian moments are finite, and $\mathbb{E}[\rho^{-1}] = \int_0^\infty e^{-\rho^2/2} d\rho = \sqrt{\pi/2}$, so Cauchy–Schwarz gives an integrable envelope.

For the cancellation, condition on $W$ and pair the draw $a$ with $-a$, which is legitimate because the $a_i$ are i.i.d. symmetric and independent of $W$. The residual is not fixed under the flip: $f_0 = m^{-1/2} \sum_i a_i s_i$ is odd in $a$, so at $-a$ the initial residual is $-f_0 - y$. Using Proposition 12, first its oddness at fixed residual and then its linearity in the residual,

$$\mathcal{V}_g(-a; -f_0 - y) = -\mathcal{V}_g(a; -f_0 - y) = \mathcal{V}_g(a; f_0) + \mathcal{V}_g(a; y),$$

while at $a$ the velocity is $\mathcal{V}_g(a; f_0) - \mathcal{V}_g(a; y)$. Averaging the pair leaves $\mathcal{V}_g(a; f_0)$. The target contribution cancels exactly, at every $m$ and for every $W$. □

**Corollary 14** (Centered model)**.** *If the trained function is the centered network $f_\theta - f_{\theta(0)}$ (Chizat et al., 2019), whose tangent kernel coincides with $\Theta_\theta^{(m)}$, then $\mathbb{E}\big[\frac{d}{dt} q_g(\Theta_t^{(m)})\big|_{t=0}\big] = 0$ for every continuous $g$ and every target $y$.*

*Proof.* Centering does not change $\nabla_\theta f$, since $\theta(0)$ is frozen, so equation 56 applies verbatim with $r_t = f_\theta - f_{\theta(0)} - y$. At $t = 0$ this residual is $-y$, which does not depend on $a$, so Proposition 12 gives $\mathcal{V}_g(-a; -y) = -\mathcal{V}_g(a; -y)$, and $a \overset{d}{=} -a$. □

Neither result uses the circle. The argument needs only that the kernel depend on the output weights through $a_i^2$, which is a property of the architecture, so both hold verbatim for the Gaussian-input experiment of Section 5.3 with $L^2(\mu)$ read on $\mathbb{R}^d$.

### D.5 Size and sign

The size and the sign require computing the derivative. The motion of the activation-arc endpoints equation 58 is already in place, and the rest is direct calculation.

**Theorem 15** (Expected initial velocity)**.** *In the setting of Theorem 13,*

$$\begin{aligned} \mathbb{E}\Big[\frac{d}{dt} q_g\big(\Theta_t^{(m)}\big)\Big|_{t=0}\Big] &= -\frac{1}{m}\, D(g), \\ D(g) = \mathbb{E}_w\Big[2\big(\langle g, s\rangle\langle g, hV[s]\rangle + \langle s, s\rangle\langle g, hV[g]\rangle\big) + \frac{3}{\pi\rho}\sum_{\varepsilon} g(\varphi^\varepsilon)\, V[s](\varphi^\varepsilon)\, V[g](\varphi^\varepsilon)\Big], \end{aligned} \tag{60}$$

*the expectation being over a single neuron $w \sim \mathcal{N}(0, I_2)$ with $b = 0$. Moreover $D(g) > 0$ whenever $g \geq 0$ is not identically zero, and*

$$\mathbb{E}\Big[\frac{d}{dt}\mu_{\mathcal{F}_0}(0)\Big] = -\frac{1}{m}\Big(\frac{1}{4} + \frac{7}{\pi^2}\Big) \approx -\frac{0.959}{m}, \qquad \mathbb{E}\Big[\frac{d}{dt}\mu_{\mathcal{F}_1}(0)\Big] = -\frac{1}{m}\Big(\frac{3}{16} + \frac{7}{2\pi^2}\Big) \approx -\frac{0.542}{m}. \tag{61}$$

*The sign condition covers $\phi_0$ but not $\phi_{1,c}$ and $\phi_{1,s}$, which take negative values, so for $\mathcal{F}_1$ positivity follows from equation 61.*

*Proof. Step 0: the closed form.* By equation 58 and the Leibniz rule,

$$\frac{d}{dt}\langle g, h_i\rangle = \frac{1}{2\pi}\sum_{\varepsilon} \varepsilon\, g(\varphi_i^\varepsilon)\, \dot\varphi_i^\varepsilon = -\frac{a_i\, m^{-1/2}}{2\pi\rho_i}\sum_{\varepsilon} g(\varphi_i^\varepsilon)\, V_i[r](\varphi_i^\varepsilon),$$

and the same computation with $g$ replaced by $gx$ gives $\frac{d}{dt}\langle g, h_i x\rangle$ with an extra factor $x(\varphi_i^\varepsilon)$ in each term; together these two are $\frac{d}{dt}\langle g, h_i V_i[g]\rangle$ up to the factor 2. For $s_i$ the endpoint terms vanish because $s_i(\varphi_i^\varepsilon) = 0$, so $\frac{d}{dt}\langle g, s_i\rangle = \langle g, h_i \dot u_i\rangle = -a_i m^{-1/2}\langle g, h_i V_i[r]\rangle$. Finally $\frac{d}{dt}a_i^2 = -2a_i m^{-1/2}\langle r, s_i\rangle$. Substituting these three into the derivative of equation 57 gives, for every continuous $g$,

$$\begin{aligned} \frac{d}{dt} q_g\big(\Theta_t^{(m)}\big)\Big|_{t=0} = -\frac{1}{m^{3/2}}\sum_{i=1}^{m}\Big[ 2a_i\Big(\langle g, s_i\rangle\, \langle g, h_i V_i[r]\rangle + \langle r, s_i\rangle\, \langle g, h_i V_i[g]\rangle\Big) \\ + \frac{a_i^3}{\pi\rho_i}\sum_{\varepsilon} g(\varphi_i^\varepsilon)\, V_i[r](\varphi_i^\varepsilon)\, V_i[g](\varphi_i^\varepsilon)\Big], \end{aligned} \tag{62}$$

in which the first group is the $[a, w]$ and $[a, b]$ blocks of the Hessian, the second the $[w, a]$ and $[b, a]$ blocks, and the third the $[w, w]$, $[w, b]$ and $[b, b]$ blocks, where $\sigma'' = \delta$ appears as the endpoint motion.

*Step 1: only the self part survives.* By Theorem 13 the $-y$ part of $r_0$ contributes nothing, so take $r = f_0 = m^{-1/2}\sum_j a_j s_j$. Then $V_i[r] = m^{-1/2}\sum_j a_j V_i[s_j]$ and $\langle r, s_i\rangle = m^{-1/2}\sum_j a_j\langle s_j, s_i\rangle$, so the $i$-th summand of equation 62 is $m^{-1/2}\sum_j a_j[a_i F_1(w_i, w_j) + a_i^3 F_3(w_i, w_j)]$. For $j \neq i$ the factor $a_j$ is symmetric and independent of everything else, so its expectation vanishes. For $j = i$, use $s_i h_i = s_i$, which gives $V_i[s_i] = V[s]$ and $\langle s_i, s_i\rangle = \langle s, s\rangle$ in the single-neuron notation. Taking expectations with $\mathbb{E}a^2 = 1$ and $\mathbb{E}a^4 = 3$, and collecting the prefactor $-m^{-3/2}\cdot m^{-1/2}\cdot m = -1/m$, gives equation 60. The cancellation is exact at every $m$; no large-width approximation is used.

*Step 2: positivity and the constants.* The single-neuron integrals are collected in Section D.7; they give $\langle s, s\rangle = \rho^2/4$ and $V[s](\varphi) = \rho/\pi + (\rho/4)\hat w^\top x(\varphi)$, so $V[s] > 0$ everywhere. For $g \geq 0$

the first two terms of equation 60 are then nonnegative, and so is the third, because $V[g](\varphi^\varepsilon) = \int g\, h\, c(\cdot, \varphi^\varepsilon)\, d\mu \geq 0$ by $c \geq 0$; moreover $\langle g, s\rangle\langle g, hV[s]\rangle > 0$ unless $g$ vanishes almost everywhere on the arc, and the arc is a uniformly random half circle, so $D(g) > 0$. The evaluations $D(\phi_0) = \frac{1}{4} + \frac{7}{\pi^2}$ and $D(\phi_{1,c}) = D(\phi_{1,s}) = \frac{3}{16} + \frac{7}{2\pi^2}$ are carried out in Section D.7, and averaging the two directions of $\mathcal{F}_1$ gives equation 61. □

### D.6 Consequences

Collecting the three results: the parameter motion is confined to $\mathcal{H}_1$ whatever the residual (Lemma 11); the expected initial velocity is blind to the target (Theorem 13) and exactly zero for the centered model used in the controls (Corollary 14); where it is not zero it is *negative* on both $\mathcal{F}_0$ and $\mathcal{F}_1$, and of order $1/m$ (Theorem 15).

The calculation therefore does not explain the growth. At first order the low-frequency strength moves down, by an amount that vanishes as width grows. It rules out two explanations: no initial-velocity argument can account for the increase of Section 5, and no residual-driven argument can account for it at first order. The effect must come from the accumulated motion, at second and higher order in $t$, whose sign this computation does not fix. Lemma 11 says which channel the motion enters through, not that the outcome concentrates there, so the parameterization account of Section 5.3 rests on the experiments rather than on this calculation.

### D.7 Deferred computations

This section collects the single-neuron integrals used in Step 2 of the proof of Theorem 15. Throughout $b = 0$, so $h = 1$ on $|\varphi - \psi| < \pi/2$ and $s = \rho\cos(\varphi - \psi)$ there. Substitute $\varphi = \psi + \tau$ with $\tau \in (-\pi/2, \pi/2)$ and write $\hat{w} = (\cos\psi, \sin\psi)$, $\hat{w}^\perp = (-\sin\psi, \cos\psi)$, so that $x(\psi + \tau) = \cos\tau\, \hat{w} + \sin\tau\, \hat{w}^\perp$. Then

$$\begin{aligned} \langle h\rangle &= \frac{1}{2}, \qquad \langle s\rangle = \frac{\rho}{2\pi}\int_{-\pi/2}^{\pi/2} \cos\tau\, d\tau = \frac{\rho}{\pi}, \qquad \langle s, s\rangle = \frac{\rho^2}{2\pi}\int_{-\pi/2}^{\pi/2} \cos^2\tau\, d\tau = \frac{\rho^2}{4}, \\ \langle hx\rangle &= \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} x(\psi+\tau)\, d\tau = \frac{\hat{w}}{\pi}, \qquad \langle sx\rangle = \frac{\rho}{2\pi}\int_{-\pi/2}^{\pi/2} \cos\tau\, x(\psi+\tau)\, d\tau = \frac{\rho}{4}\hat{w}, \end{aligned} \tag{63}$$

the odd-in-$\tau$ parts vanishing in each case. Hence $V[s](\varphi) = \rho/\pi + (\rho/4)\hat{w}^\top x(\varphi)$, and $V[s](\varphi^\varepsilon) = \rho/\pi$ because $\hat{w}^\top x(\varphi^\varepsilon) = 0$.

**Constant mode.** Take $g = \phi_0 = 1$. Then $\langle g, h\rangle = \frac{1}{2}$, $\langle g, hx\rangle = \hat{w}/\pi$, $\langle g, s\rangle = \rho/\pi$, and

$$\langle g, hV[g]\rangle = \langle g, h\rangle^2 + \|\langle g, hx\rangle\|^2 = \frac{1}{4} + \frac{1}{\pi^2}, \qquad \langle g, hV[s]\rangle = \frac{1}{2}\cdot\frac{\rho}{\pi} + \frac{\hat{w}^\top}{\pi}\cdot\frac{\rho\hat{w}}{4} = \frac{3\rho}{4\pi},$$

while $g(\varphi^\varepsilon) = 1$ and $V[g](\varphi^\varepsilon) = \frac{1}{2}$. The three groups of equation 60 are therefore

$$2\cdot\frac{\rho}{\pi}\cdot\frac{3\rho}{4\pi} = \frac{3\rho^2}{2\pi^2}, \qquad 2\cdot\frac{\rho^2}{4}\Big(\frac{1}{4} + \frac{1}{\pi^2}\Big) = \frac{\rho^2}{8} + \frac{\rho^2}{2\pi^2}, \qquad \frac{3}{\pi\rho}\cdot 2\cdot\frac{\rho}{\pi}\cdot\frac{1}{2} = \frac{3}{\pi^2},$$

and with $\mathbb{E}\rho^2 = 2$, $D(\phi_0) = 2\big(\frac{2}{\pi^2} + \frac{1}{8}\big) + \frac{3}{\pi^2} = \frac{1}{4} + \frac{7}{\pi^2}$.

**First harmonic.** Take $g = \phi_{1,c} = \sqrt{2}\cos\varphi = \sqrt{2}(\cos\psi\cos\tau - \sin\psi\sin\tau)$. The same substitution gives

$$\langle g, h\rangle = \frac{\sqrt{2}\cos\psi}{\pi}, \qquad \langle g, s\rangle = \frac{\sqrt{2}\,\rho\cos\psi}{4},$$
$$\langle g, hx\rangle = \frac{\sqrt{2}}{4}\big(\cos\psi\,\hat{w} - \sin\psi\,\hat{w}^\perp\big) = \frac{\sqrt{2}}{4}(1, 0),$$

so $\langle g, hV[g]\rangle = \frac{2\cos^2\psi}{\pi^2} + \frac{1}{8}$ and $\langle g, hV[s]\rangle = \sqrt{2}\rho\cos\psi\big(\frac{1}{\pi^2} + \frac{1}{16}\big)$. At $\varphi^\varepsilon = \psi + \varepsilon\pi/2$ one has $g(\varphi^\varepsilon) = -\varepsilon\sqrt{2}\sin\psi$ and $V[g](\varphi^\varepsilon) = \frac{\sqrt{2}\cos\psi}{\pi} - \varepsilon\frac{\sqrt{2}}{4}\sin\psi$, so the third group of equation 60 is

$$\frac{3}{\pi\rho}\cdot\frac{\rho}{\pi}\sum_{\varepsilon}(-\varepsilon\sqrt{2}\sin\psi)\Big(\frac{\sqrt{2}\cos\psi}{\pi} - \varepsilon\frac{\sqrt{2}}{4}\sin\psi\Big) = \frac{3}{\pi^2}\sum_{\varepsilon}\Big(-\varepsilon\frac{2\sin\psi\cos\psi}{\pi} + \frac{\sin^2\psi}{2}\Big)$$
$$= \frac{3\sin^2\psi}{\pi^2},$$

the cross terms cancelling between the two endpoints. Collecting, the integrand of equation 60 is

$$\rho^2\cos^2\psi\Big(\frac{1}{\pi^2} + \frac{1}{16}\Big) + \frac{\rho^2}{2}\Big(\frac{2\cos^2\psi}{\pi^2} + \frac{1}{8}\Big) + \frac{3\sin^2\psi}{\pi^2}.$$

Under $\mathcal{N}(0, I_2)$ the radius and the angle are independent, $\psi$ is uniform and $\mathbb{E}\rho^2 = 2$, so $\mathbb{E}\cos^2\psi = \mathbb{E}\sin^2\psi = \frac{1}{2}$ and

$$D(\phi_{1,c}) = \Big(\frac{1}{\pi^2} + \frac{1}{16}\Big) + \Big(\frac{1}{\pi^2} + \frac{1}{8}\Big) + \frac{3}{2\pi^2} = \frac{3}{16} + \frac{7}{2\pi^2}.$$

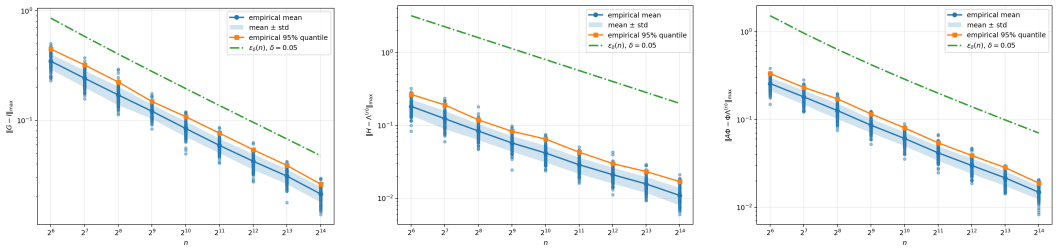
The same value holds for $\phi_{1,s}$ by rotation invariance of the law of $w$.

# E Additional Experimental Results

This appendix collects supporting figures that complement the main results but would interrupt the flow of the main text.

## E.1 Finite-sample: additional evidence

The three finite-sample errors of Theorem 4 all decrease with $n$ (Figure 5), and the corrected finite-sample eigenvalues $\lambda_p^{(n)}$ predict the early-time residual decay, with deviations appearing only at late times once a mode has decayed (Figure 6).

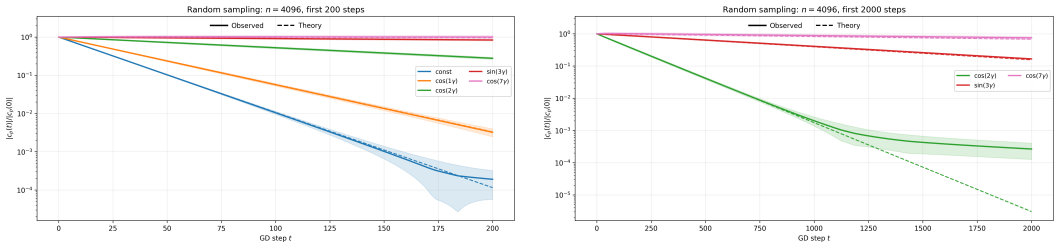



(a) $\|G - I_d\|_{\max}$ (b) $\|H - \Lambda^{(n)}\|_{\max}$ (c) $\|A\Phi - \Phi\Lambda^{(n)}\|_{\max}$

Figure 5: Finite-sample diagnostics. The three errors of Theorem 4 decrease with $n$; dashed curves are the high-probability envelopes.

Figure 7 shows how the normalised mode-1 residual $|\alpha_1(t)|/|\alpha_1(0)|$ evolves across a wide range of sample sizes. The convergence rate improves steadily with $n$ and is already well-behaved at $n = 1024$, consistent with the finite-sample bounds.

Figure 8 compares, for a single sample size, the empirical eigenvalues of the compressed operator $C = G^{-1}H$ against the predicted finite-sample eigenvalues $\lambda_p^{(n)}$. The predicted values carry the correct Fourier multiplicity: the constant mode contributes a single eigenvalue, while each frequency $k \geq 1$ contributes a degenerate pair, visible as the flat steps in the figure. The empirical eigenvalues sit on these steps and split only slightly within each pair, confirming that the sampled operator is close to Fourier-diagonal on the retained block. The agreement is tightest for the leading (low-frequency) eigenvalues and loosens for the smallest eigenvalues, which is where the finite-sample action error is relatively largest.




(a) First 200 steps.

(b) Higher-frequency modes, 2000 steps.

Figure 6: Random sampling with $n = 4096$. Solid: observed normalized residual amplitudes; dashed: diagonal prediction from $\lambda_p^{(n)}$. The prediction captures early decay; late-time deviations appear once components are small.

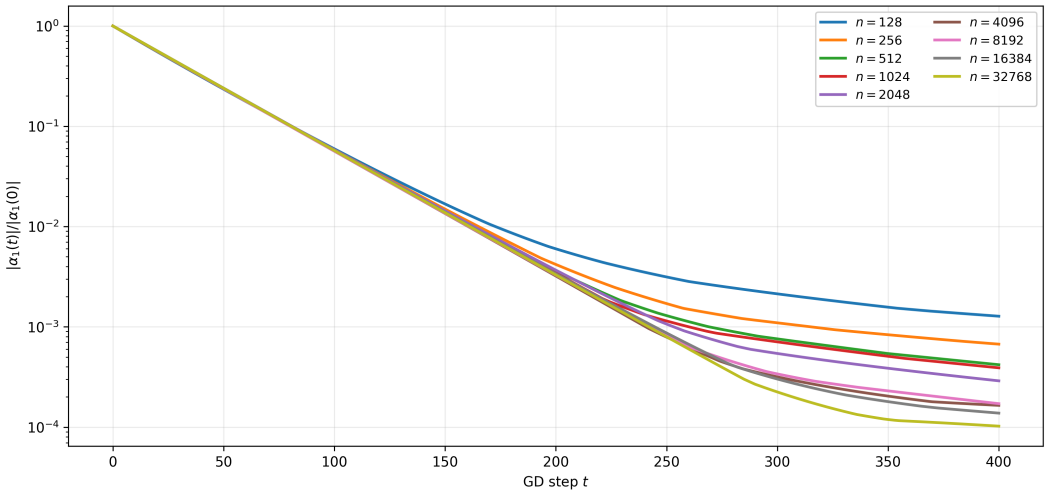



Figure 7: Normalised residual for mode $k = 1$ over the first 400 gradient steps, for sample sizes $n \in \{128, \ldots, 32768\}$. Larger $n$ gives faster effective decay, converging toward the continuum rate.

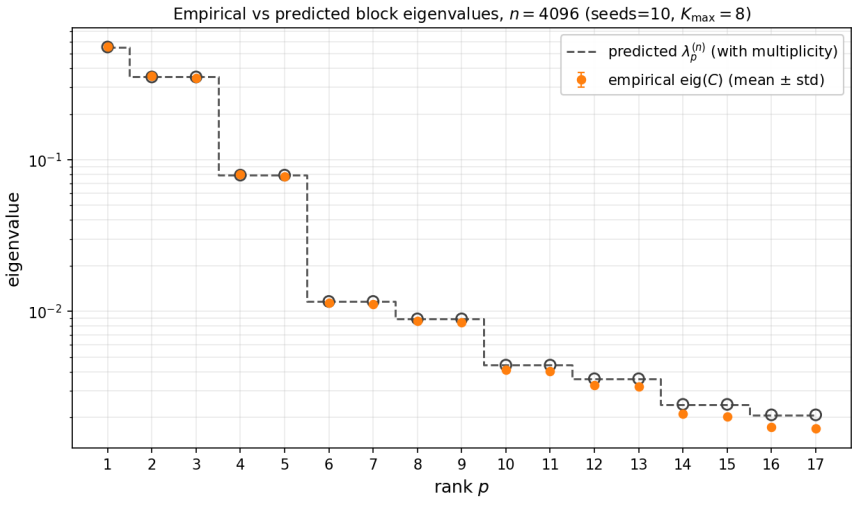



Figure 8: Empirical versus predicted block eigenvalues at $n = 4096$, on a logarithmic scale, averaged over 10 seeds. The dashed step shows the predicted eigenvalues $\lambda_p^{(n)}$ with their correct multiplicity (single for $k = 0$, degenerate pairs for $k \geq 1$); markers show the empirical eigenvalues $\mathrm{eig}(C)$ (mean $\pm$ standard deviation).

Figure 9 shows the effect of increasing $n$ on the block-diagonal structure of the sampled Gram matrix. The left panel is the baseline spectrum; the centre and right panels show how the theory- and empirical-preconditioned spectra flatten toward the identity as $n$ grows, validating the preconditioner construction.

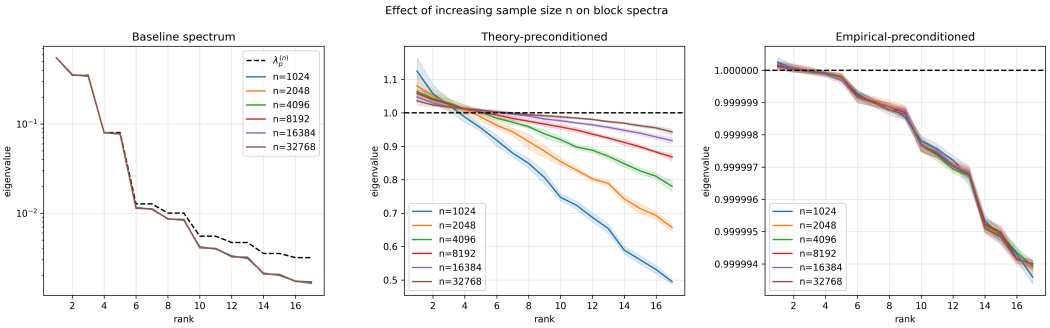



Figure 9: Block spectra of the compressed Gram matrix for $n \in \{1024, \ldots, 32768\}$. Left: baseline (decaying) spectrum. Centre: theory-preconditioned spectrum approaching 1. Right: empirical-preconditioned spectrum, closely tracking the theory.

Figure 10 shows the same three spectra for a single sample size $n = 4096$, annotated with the condition number of each block. The baseline block is badly conditioned ($\kappa \approx 1.4 \times 10^3$); the theory preconditioner reduces this to $\kappa \approx 2.1$, and the empirical preconditioner flattens the block to $\kappa \approx 1$.

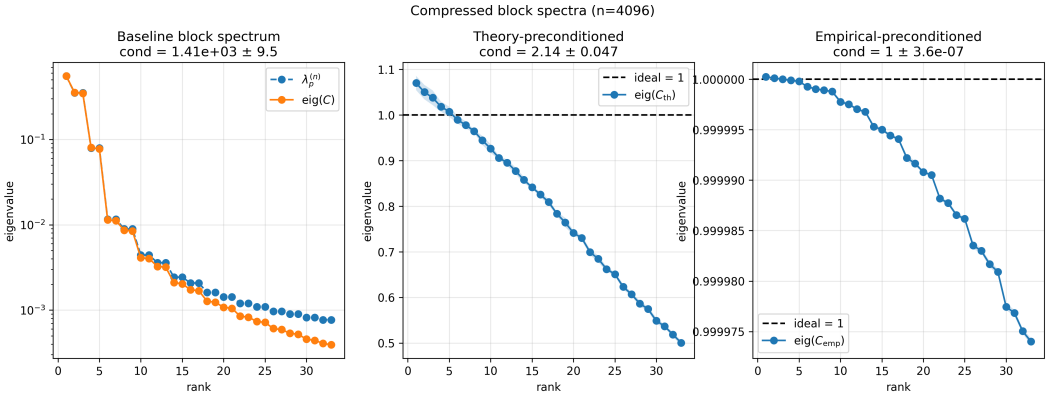



Figure 10: Compressed block spectra at $n = 4096$ with all active modes inside the preconditioning block. Left: baseline spectrum with a wide spread. Centre: theory-preconditioned spectrum, reduced spread but a residual slope. Right: empirical-preconditioned spectrum, essentially flat at 1. Titles report the per-block condition number.

Figures 11–12 show the normalised residual decay for individual Fourier modes under the baseline and both preconditioned learning rules. The baseline decay is negligible over 2000 steps for both modes; the preconditioned rules reduce the residual by several orders of magnitude within the same budget.

### E.2 Finite-sample mode decay across sample sizes

The main text shows the observed-versus-predicted mode decay for a single sample size $n = 4096$ (Figures 6a and 6b). Here we repeat the same experiment at two representative sample sizes, the smallest and largest in the sweep, $n = 128$ and $n = 32768$. For each $n$, the left panel shows all retained modes over the first 200 gradient-descent steps, and the right panel isolates the higher-frequency modes over 2000 steps. Solid curves are the observed normalised residual amplitudes

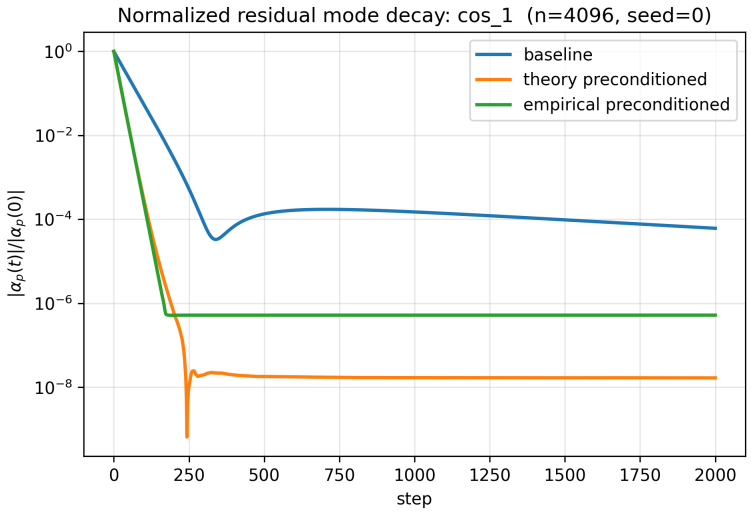



Figure 11: Per-mode residual decay for $\cos_1$ ($n = 4096$, seed 0). Baseline (blue) is effectively flat; theory-preconditioned (orange) and empirical-preconditioned (green) converge rapidly.

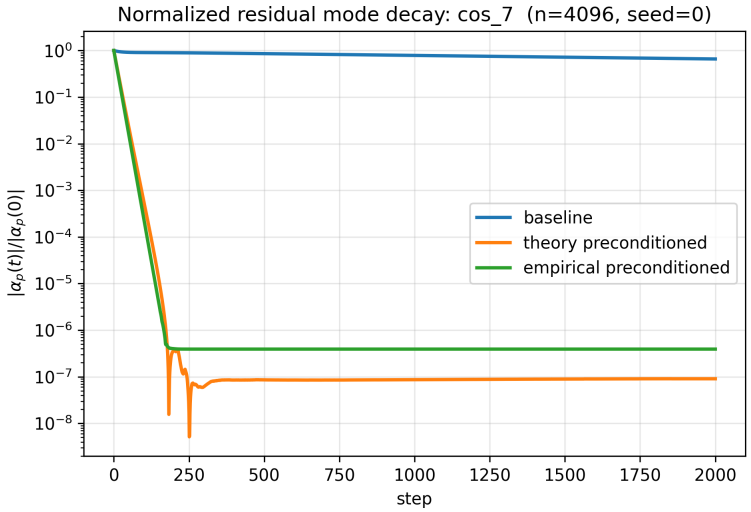


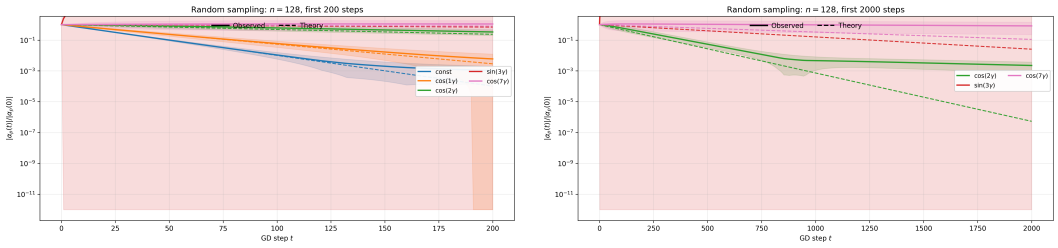

Figure 12: Per-mode residual decay for $\cos_7$ ($n = 4096$, seed 0). The spectral gap for this higher mode makes the baseline decay even slower; preconditioning recovers roughly six orders of magnitude of relative residual within 2000 steps.

$|\alpha_p(t)|/|\alpha_p(0)|$; dashed curves are the diagonal prediction from the finite-sample corrected eigenvalues $\lambda_p^{(n)}$.

The trend is consistent with the finite-sample theory. At small $n$, the sampled operator is a rough approximation, the prediction bands are wide, and the observed curves flatten early as the finite-sample action error becomes comparable to the remaining motion. As $n$ grows, the bands tighten and the observed decay tracks the diagonal prediction over an increasingly long horizon before flattening.

### E.3 Frozen-kernel Fourier eigenspace alignment

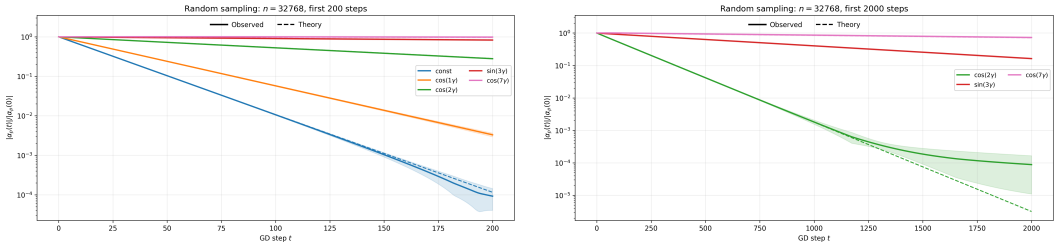
The frozen-kernel analysis assumes that the initial finite-width operator already reproduces the Fourier spectral structure of the infinite-width theory. This section examines how well that assump-




(a) All modes, first 200 steps.

(b) Higher-frequency modes, 2000 steps.

Figure 13: Random sampling with $n = 128$: observed (solid) versus finite-sample eigenvalue prediction (dashed). At the smallest sample size the prediction bands are wide and the observed curves flatten early.




(a) All modes, first 200 steps.

(b) Higher-frequency modes, 2000 steps.

Figure 14: Random sampling with $n = 32768$: observed (solid) versus finite-sample eigenvalue prediction (dashed). At the largest sample size the bands are tight and the agreement with the diagonal prediction is closest.

tion holds frequency by frequency. For each $k \geq 1$, the ideal eigenspace is the two-dimensional Fourier plane $\mathcal{F}_k = \mathrm{span}\{\phi_{k,c}, \phi_{k,s}\}$, so the comparison must be made at the level of subspaces rather than individual eigenvectors: even a perfectly recovered frequency block has an arbitrary orthonormal basis inside $\mathcal{F}_k$. We therefore use two basis-free diagnostics, the principal angles between the empirical eigenspace and $\mathcal{F}_k$, and the projector error $\|\widehat{P}_k - P_k\|_F$.

An important caveat is that eigenvalue agreement alone does not certify eigenspace recovery. The eigenvalue alignment error and the empirical splitting of the degenerate $k$-pair both decrease quickly with width, yet both can be small at high frequencies while the corresponding eigenspaces are still poorly aligned.

The eigenspace diagnostics show a clear frequency hierarchy. Comparing the mean empirical eigenvectors with the Fourier targets at width 256, the low frequencies are tracked closely while phase and amplitude distortions grow with $k$, and the same is visible in the raw Fourier plane, where the low-frequency clouds lie near the unit circle and higher frequencies drift inward. Figure 15 summarises this as a projector error that grows rapidly with $k$: only the lowest modes are recovered accurately at the widths considered here.

When the kernel is allowed to evolve, the empirical eigenspaces move *away* from the Fourier planes rather than toward them, even as the network optimises better than the frozen baseline. The principal angles increase during training at width 256, with the largest drift at $k = 5$ and $k = 6$, and the same drift shows in the raw Fourier plane. This is consistent with the interpretation that feature learning helps through spectral reweighting rather than by preserving the Fourier geometry.

### E.4 Finite-width frozen: projector stability

The frozen-NTK analysis relies on the eigenspaces of $C_0$ remaining close to those of $C_t$ throughout training. Figure 16 confirms this directly: the absolute projection error $\|\hat{P}_k(t) - \hat{P}_k(0)\|_F$ stays close

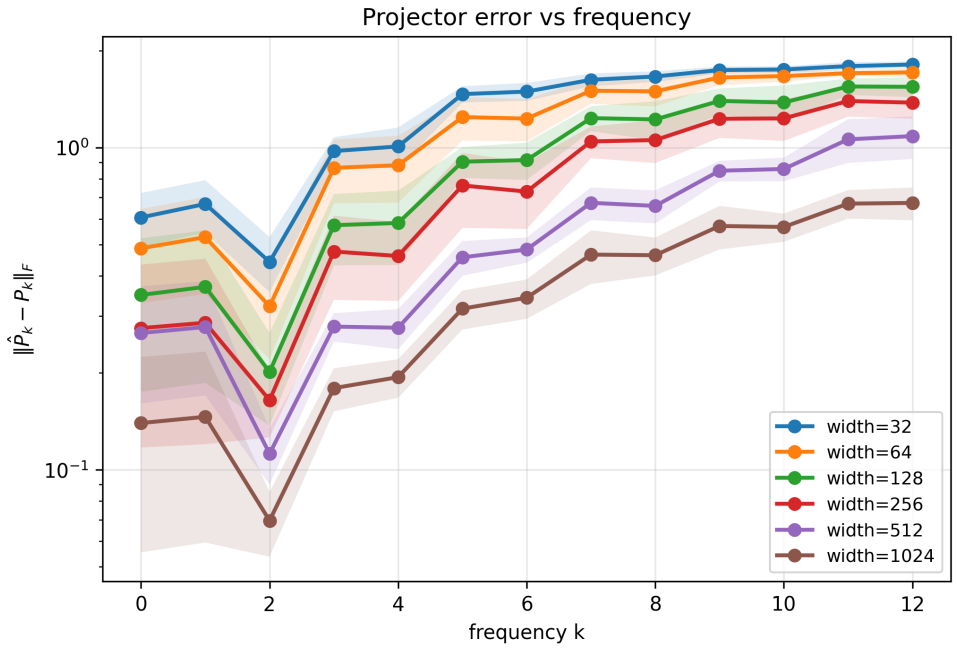
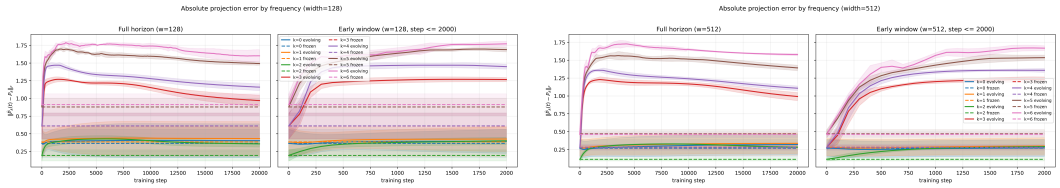



Figure 15: Projector error $\|\widehat{P}_k - P_k\|_F$ versus frequency for several widths. The error grows rapidly with $k$: subspace recovery deteriorates at higher frequencies even when the lowest modes are well aligned.

to zero for the frozen network at both widths, while it grows for the evolving network, especially at the smallest width.



(a) $w = 128$ (b) $w = 512$

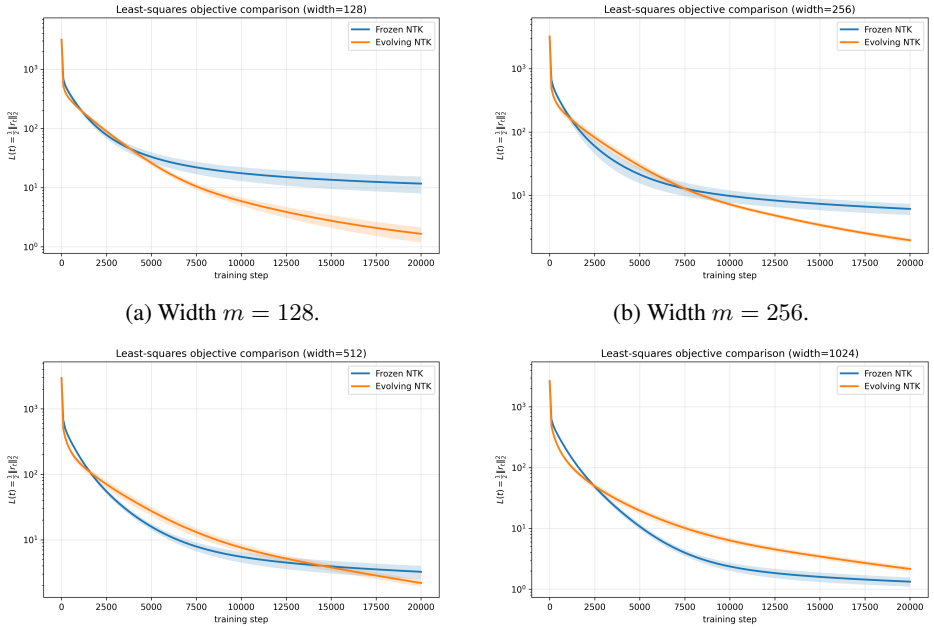
Figure 16: Absolute projection error $\|\hat{P}_k(t) - \hat{P}_k(0)\|_F$ by frequency $k$, comparing frozen (dashed) and evolving (solid) networks at widths 128 and 512. Frozen projectors are stable; evolving ones drift, increasingly so at low width.

### E.5 Finite-width evolving: cross-width summary and robustness

Figure 17 gives the least-squares objective of Figure 2 at every width. The evolving kernel reaches a lower final loss at widths 128 and 256, the advantage shrinks at 512, and at 1024 the frozen kernel is again better.

Figure 18 gives the two panels of Figure 3 at every width. The pattern is consistent: the spectral-strength increase over the frozen baseline is largest at width 128 and shrinks with width, while alignment does not improve with training at any width.

Figure 19 shows the two robustness diagnostics referenced from Section 5.2: the low-frequency block drift, which is largest at small width and matches where the evolving kernel helps, and the cumulative low-frequency strength under the three amplitude-variant targets, which rises above the frozen baseline in every case, ruling out target-amplitude bias.




(a) Width $m = 128$.

(b) Width $m = 256$.

(c) Width $m = 512$.

(d) Width $m = 1024$.

Figure 17: Least-squares objective $\frac{1}{2}\|r_t\|_2^2$ under frozen (dashed) and evolving (solid) finite-width tangent-kernel dynamics, one panel per width. Curves are means over 5 seeds, with $\pm 1$ standard-deviation bands.

Figure 20 shows the mode-wise residual decay behind the loss curves of Figure 2, at width 128. The low-frequency-first ordering is preserved under the evolving kernel: the constant mode and $k = 1$ still fall first, and the target-present high frequencies still fall last.

Figure 21 gives the per-frequency spectral strength of Figure 18 under the two amplitude-variant targets. The low-frequency blocks rise above the frozen baseline in both, and the rise shrinks with width as it does under the original target.

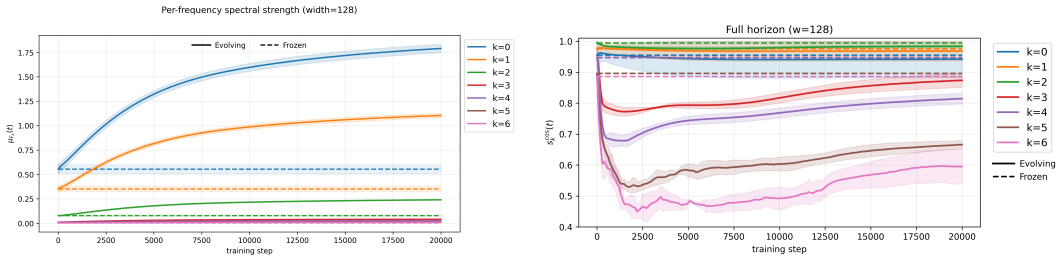
Figure 22 reports the two geometry diagnostics accumulated over the nested subspaces $\mathcal{F}_{\leq k}$ rather than per frequency. The cumulative RMS-cosine alignment and the cumulative projection error tell the same story as the per-frequency plots: the evolving kernel (solid) drifts away from the Fourier reference relative to the frozen baseline (dashed), with the gap widening as more high-frequency blocks are included.

### E.6 Gaussian inputs: least-squares objective

Section 5.3 reports the spectral strength for the Gaussian mode-absent control but not the loss. Figure 23 gives the least-squares objective for those runs, frozen against evolving, at each width in the sweep. The evolving kernel reaches a lower loss at every width, so the degree-3 target is being fitted while the absent degrees gain strength.

### E.7 Consistency check: Fourier-block preconditioning

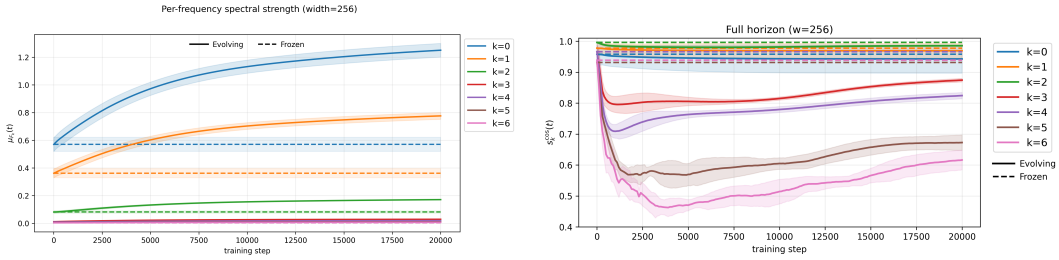
This subsection is a consistency check on the finite-sample analysis of Appendix B, not a proposed method. The finite-sample result predicts that, on the retained block, the sampled operator acts almost diagonally in the Fourier basis with the predicted eigenvalues $\Lambda^{(n)}$. If that is the reason the retained modes separate by rate, then rescaling each Fourier coefficient by its inverse predicted eigenvalue should remove most of the separation. We verify that it does. The same eigenvalue rescaling is already used in practice by the PINN and Fourier-feature work cited in the main text; nothing here is new as a technique, and it does not transfer to a finite-width trained network, whose rate-setting kernel is implicit and drifts.




(a) Spectral strength, width 128.



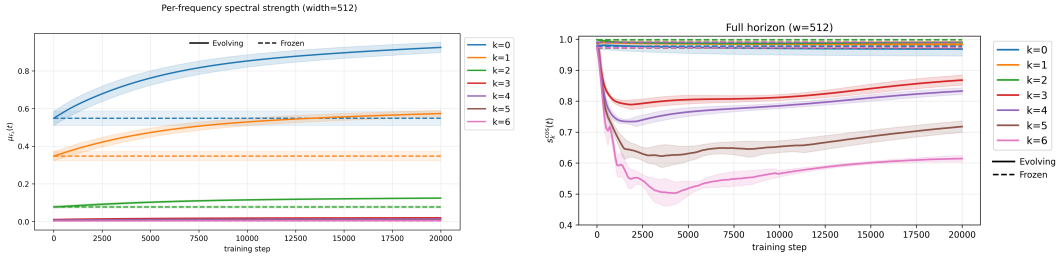

(b) Subspace alignment, width 128.




(c) Spectral strength, width 256.



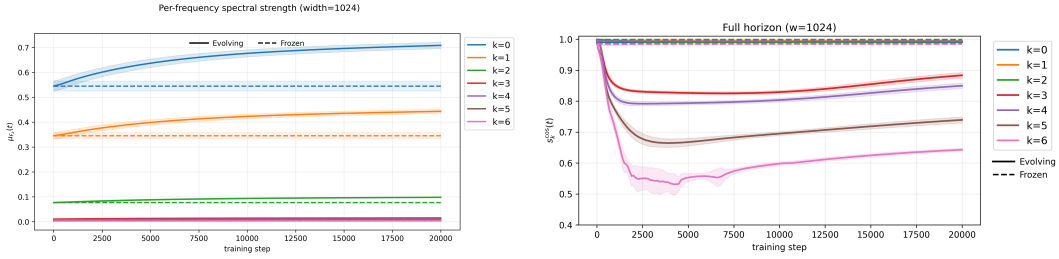

(d) Subspace alignment, width 256.




(e) Spectral strength, width 512.




(f) Subspace alignment, width 512.




(g) Spectral strength, width 1024.




(h) Subspace alignment, width 1024.

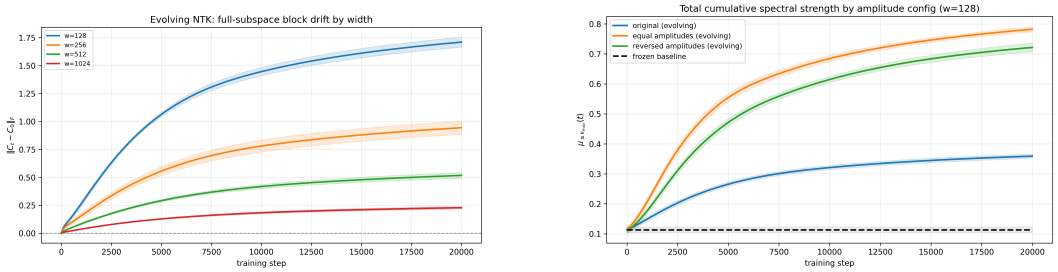
Figure 18: Per-frequency spectral strength (left) and subspace alignment (right) at widths $128, 256, 512$ and $1024$, for frozen (dashed) and evolving (solid) networks. Curves are means over 5 seeds, with $\pm 1$ standard-deviation bands. The width-128 row is Figure 3.

Let $A \in \mathbb{R}^{n\times n}$ be the normalized sampled kernel operator, $\Phi$ the sampled Fourier feature matrix, $G = \frac{1}{n}\Phi^\top\Phi$, and $C_n^{(K)} = G^{-1}H$ the observed coefficient-space block. Define the diagonal (theory) and observed (empirical) rescalings

$$P_{\mathrm{th}} = \tfrac{1}{n}\Phi\,(\Lambda^{(n)})^{-1}G^{-1}\Phi^\top, \qquad P_{\mathrm{emp}} = \tfrac{1}{n}\Phi\,(C_n^{(K)})^{-1}G^{-1}\Phi^\top,$$

and compare the baseline dynamics $B_{\mathrm{base}} = A$ with $B_{\mathrm{th}} = P_{\mathrm{th}}A$ and $B_{\mathrm{emp}} = P_{\mathrm{emp}}A$. The theory rescaling uses only the closed-form eigenvalues $\Lambda^{(n)}$; the empirical one uses the observed block and is exact on the retained modes by construction.

Figure 24 shows the outcome at $n = 4096$. The baseline dynamics separate the retained modes by frequency, and both rescalings flatten that separation, with the empirical one flattest because it




(a) Low-frequency block drift $\|C_t^{(K)} - C_0^{(K)}\|_F$.

(b) Cumulative low-frequency strength, width 128.

Figure 19: (a) Kernel drift is largest at small width, matching where the evolving kernel helps. (b) The low-frequency strength increase persists across original, equal-amplitude, and reversed-amplitude targets (dashed: frozen baseline), so it is not an artifact of the target amplitudes.

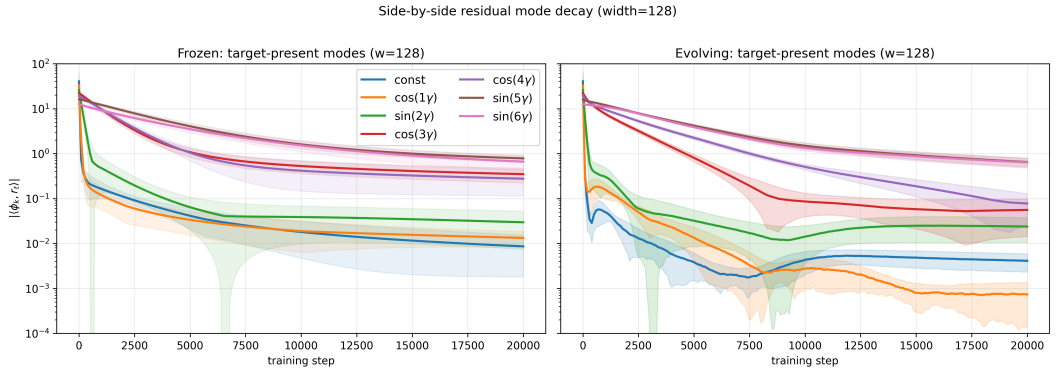
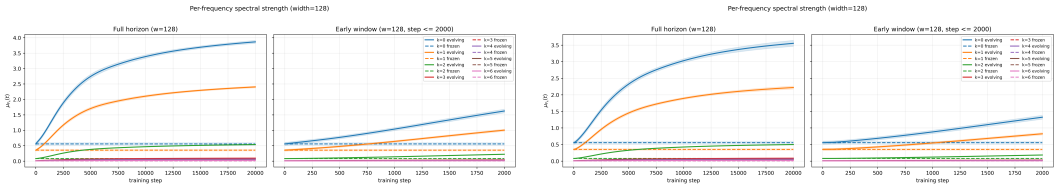



Figure 20: Per-mode residual components $|\langle \phi_k, r_t \rangle|$ at width 128 under the frozen (left) and evolving (right) kernel, means over 5 seeds with $\pm 1$ standard-deviation bands.

also removes the sample-dependent mixing inside the block. As $n$ grows, the theory-rescaled block spectrum approaches the ideal value 1, which is the same statement as $C_n^{(K)} \to \Lambda^{(n)}$ from the finite-sample analysis. The residual plateau is numerical: the retained block has condition number $\approx 330$, set by the spread of the retained eigenvalues, and after the theory rescaling the condition number drops toward 1, so the achievable relative accuracy $\kappa u$ falls from single to double precision (Table 1).

Table 1: Conditioning of the retained Fourier block and the resulting finite-precision floors, over 10 runs. $\kappa(C)$: uncorrected block; $\kappa(C_{\text{th}})$: after the theory rescaling; $\kappa u$: the relative-accuracy floor at each precision.

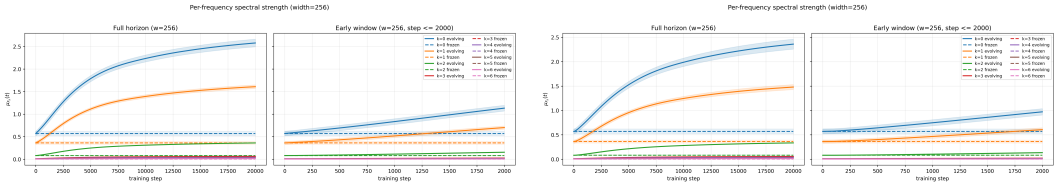
| $n$ | $\kappa(C)$ | $\kappa(C_{\text{th}})$ | $\kappa u$ (fp32) | $\kappa u$ (fp64) |
|---|---|---|---|---|
| 1024 | 340.8 | 2.273 | $2.7 \times 10^{-7}$ | $5.0 \times 10^{-16}$ |
| 4096 | 329.5 | 1.364 | $1.6 \times 10^{-7}$ | $3.0 \times 10^{-16}$ |
| 16384 | 327.1 | 1.142 | $1.4 \times 10^{-7}$ | $2.5 \times 10^{-16}$ |
| 32768 | 325.7 | 1.099 | $1.3 \times 10^{-7}$ | $2.4 \times 10^{-16}$ |



(a) Equal amplitude, width 128.



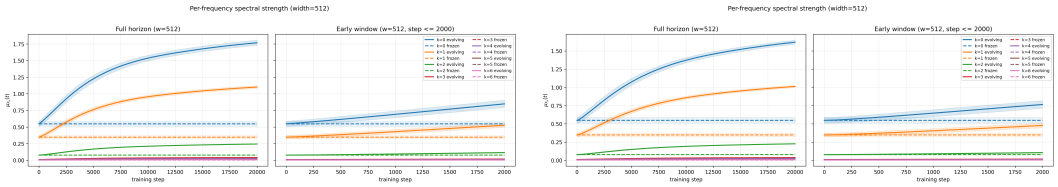
(b) Reversed amplitude, width 128.



(c) Equal amplitude, width 256.



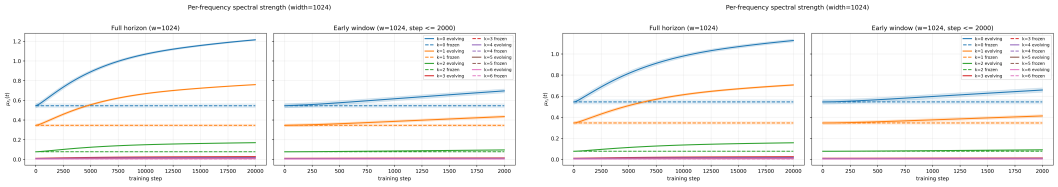
(d) Reversed amplitude, width 256.



(e) Equal amplitude, width 512.



(f) Reversed amplitude, width 512.

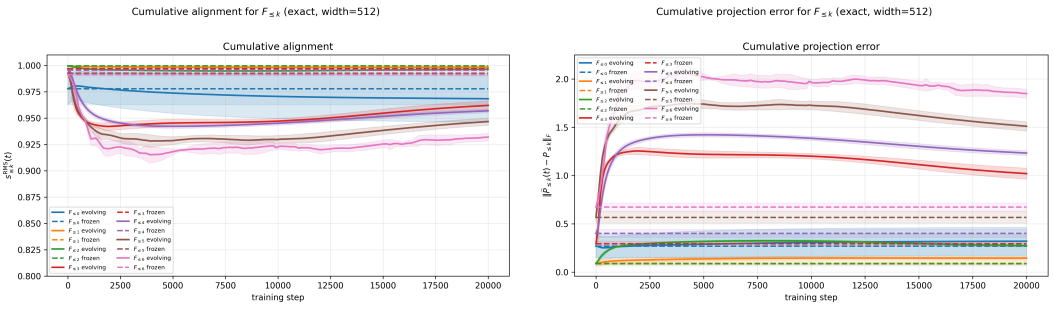


(g) Equal amplitude, width 1024.



(h) Reversed amplitude, width 1024.

Figure 21: Per-frequency spectral strength under the equal-amplitude (left) and reversed-amplitude (right) targets, at widths $128, 256, 512$ and $1024$, for frozen (dashed) and evolving (solid) networks. Curves are means over 5 seeds, with $\pm 1$ standard-deviation bands.

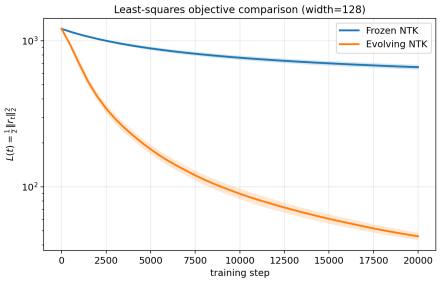


(a) Cumulative RMS-cosine alignment.

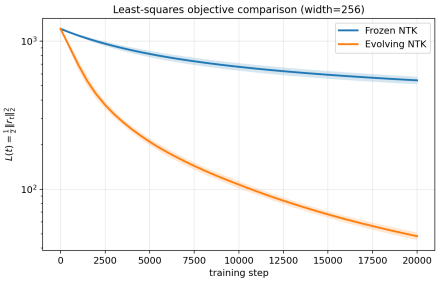


(b) Cumulative projection error.

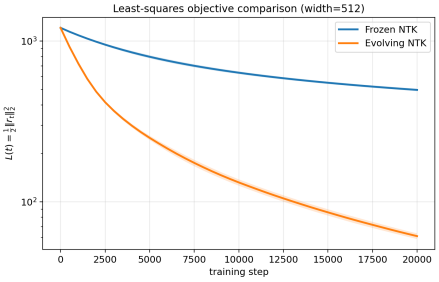
Figure 22: Cumulative geometry diagnostics over the nested subspaces $\mathcal{F}_{\leq k}$ at width $512$, for frozen (dashed) and evolving (solid) networks. Left: cumulative RMS-cosine alignment $S^{\mathrm{RMS}}_{\leq k}(t)$. Right: cumulative projection error $\|\widehat{P}_{\leq k}(t) - P_{\leq k}\|_F$. The evolving kernel becomes less Fourier-aligned over training, increasingly so as higher-frequency blocks are included.




(a) Width 128.

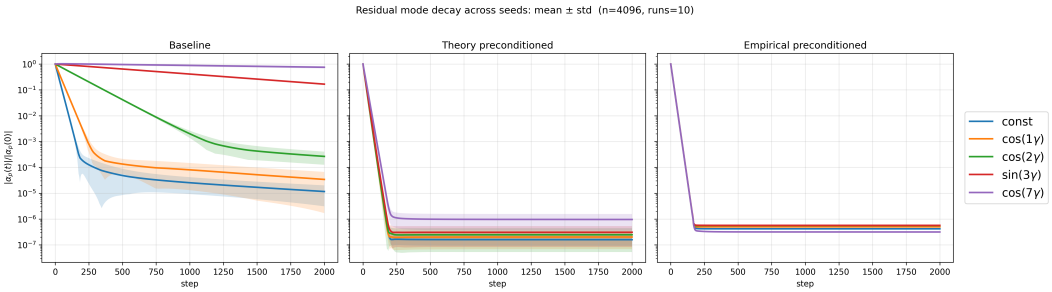



(b) Width 256.

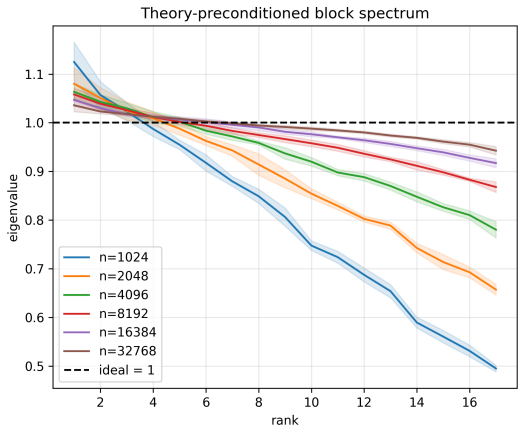



(c) Width 512.

Figure 23: Least-squares objective $L(t) = \frac{1}{2}\|r_t\|_2^2$ for the Gaussian mode-absent control of Figure 4b, frozen against evolving, one panel per width. Curves are means over 5 seeds, with $\pm 1$ standard-deviation bands. The sweep covers widths $128, 256$ and $512$.




(a) Mode-wise residual decay.




(b) Theory-rescaled block spectrum.

Figure 24: Consistency check at $n = 4096$, over 10 runs. (a) The baseline separates the retained modes by frequency; rescaling by the predicted diagonal (theory) or the observed block (empirical) flattens them. (b) As $n$ grows, the theory-rescaled block spectrum approaches 1, the same statement as $C_n^{(K)} \to \Lambda^{(n)}$.