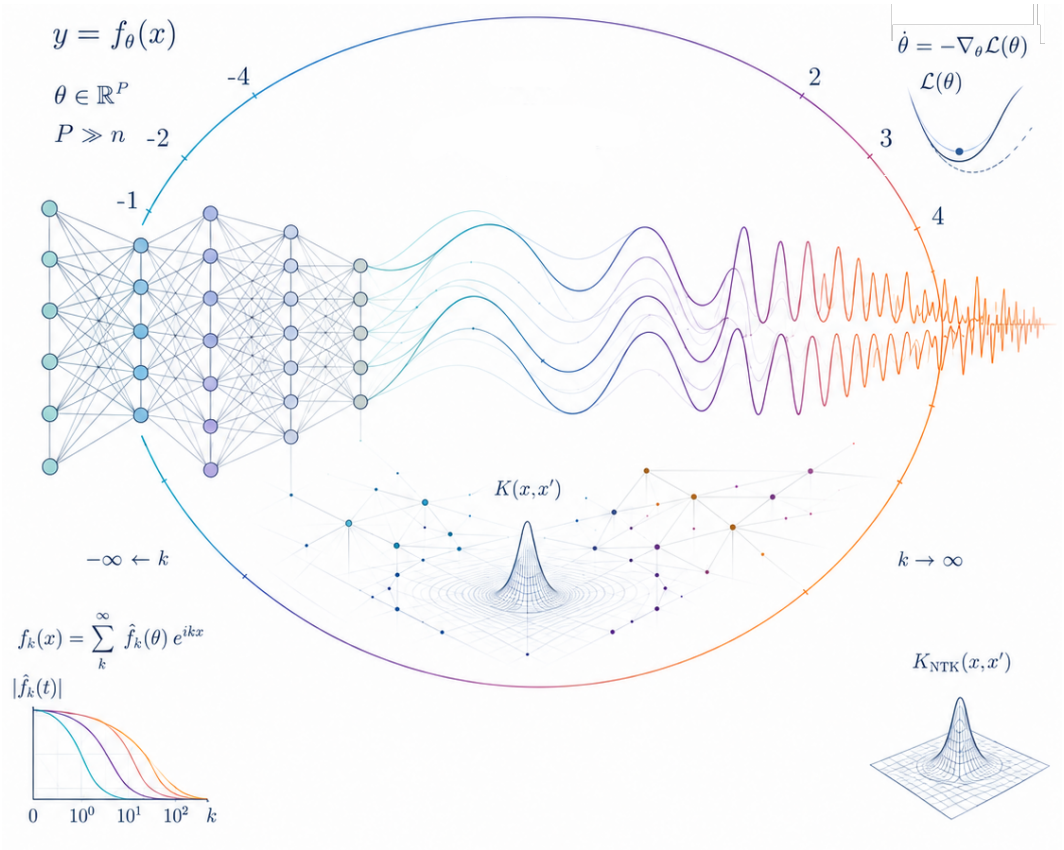
Training Dynamics of Overparameterized Neural Networks



Shreyas Kalvankar

Training Dynamics of Overparameterized Neural Networks

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Shreyas Kalvankar

Pattern Recognition & Bioinformatics Group
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
www.ewi.tudelft.nl

Training Dynamics of Overparameterized Neural Networks

Author: Shreyas Kalvankar
Student id: 6255191

Abstract

In this thesis, we study spectral bias, the tendency of gradient-based training to learn the low-frequency part of a target before its high-frequency part. We work in a setting simple enough to analyse explicitly: regression on the unit circle with a shallow ReLU network. In the infinite-width limit, the residual dynamics are governed by the Neural Tangent Kernel. Under the uniform measure on the circle this kernel depends only on the angle between two points, so the associated operator is a convolution and the Fourier modes are its eigenfunctions, each decaying at a rate set by its eigenvalue, and the lower the frequency, the larger the eigenvalue, so low frequencies are learned first.

Away from this idealised limit the picture degrades only gradually. On a fixed low-frequency subspace, both finite sampling and frozen finite width keep the operator close to the continuum Fourier prediction, with error of order $O(n^{-1/2})$ in the sample size n and $O(m^{-1/2})$ in the width m. The description breaks only once the kernel is allowed to evolve during training. At small width the evolving kernel reaches a lower loss by strengthening its lowest-frequency components, even as its alignment with the Fourier basis fails to improve. This reinforces the low-frequency bias rather than approximating the fixed-kernel dynamics. A formal theory of this evolving-kernel regime remains the main open problem.

Thesis Committee:

Advisor:	Prof. Dr. D. M. J. Tax, PRB Group, Faculty EEMCS, TU Delft
Supervisor:	Prof. Dr. A. Heinlein, Numerical Analysis, Faculty EEMCS, TU Delft
Committee Member:	Prof. Dr. A. Papapantoleon, Applied Probability, Faculty EEMCS, TU Delft

Acknowledgements

Like many analytically minded people, I have always been drawn to the theory and mathematics of systems. But a thesis is built on more than mathematics, and I owe a great deal to the people who carried me through this journey. For a long time I had a lingering curiosity about the foundations of deep learning, though it remained vague and unformed. I am truly grateful to David for guiding me since those formative days, long before the thesis began and before he formally became my advisor, helping turn an abstract interest into the concrete reality of this thesis. I am equally indebted to my supervisor, Alexander, for helping me find my footing on this topic. I owe both of them a great deal for their continued guidance and unwavering support throughout, and for anchoring me whenever I started hiding behind complex abstractions, gently pulling me back to earth and challenging me to find the true clarity beneath the technical language. I would also like to thank Francesca, whose work first showed me that this topic could be both mathematically beautiful and genuinely alive. I am deeply grateful to Prof. Dr. Jan van Neerven for teaching me functional analysis; many of the tools I used in this thesis began there, in a course that changed the way I saw mathematics. I would also like to thank Prof. Dr. Frank Redig for teaching me high-dimensional probability, a subject that ended up playing a much larger role in the thesis than I first expected. I am also grateful to Prof. Dr. Antonis Papapantoleon for serving on my committee.

To my grandmother and my parents, who have always believed in me with a steadiness I have often borrowed when my own confidence was less reliable. To Radha, who was there through the long stretches when this thesis was all I could talk about, and stayed anyway. And to my friends, both here and back home, who listened patiently while I tried to explain my thesis, sometimes with very little warning and often with the misplaced optimism that this time it would only take five minutes: your love, humour, patience, and reminders that there is life outside kernels, Fourier modes, and LaTeX error messages made this work feel less like a solitary climb and more like something I was lucky enough to carry among people who cared.

Shreyas Kalvankar
Delft, the Netherlands
June 4, 2026

Notation

The following list summarizes the notation used throughout this thesis. Symbols are grouped by theme and, within each group, ordered roughly by first appearance. Page or equation references are given where a symbol is formally defined. Throughout, the circle angle is written θ in the introductory linear-model discussion (Chapter 1) and φ from the methodology onward (Chapters 4–5); the two denote the same angular coordinate on S^1.

Sets, spaces, and measures

$\mathbb{R}$, $\mathbb{Z}$	The real numbers and the integers.
$\mathbb{R}^d$	d-dimensional Euclidean space.
S^1	The unit circle, identified with the angular domain $\Omega = [0, 2\pi)$ and embedded in $\mathbb{R}^2$ via $x(\varphi) = (\cos\varphi, \sin\varphi)$.
S^{d-1}	The unit sphere in $\mathbb{R}^d$ (the $(d-1)$-sphere).
$\mathcal{X}$, $\mathcal{Y}$	Input space ($\mathcal{X} \subseteq \mathbb{R}^d$) and target space (e.g. $\mathcal{Y} = \mathbb{R}^n$ in regression, $\mathcal{Y} = \{1, \dots, C\}$ in classification).
ρ	Joint data distribution over $\mathcal{X} \times \mathcal{Y}$; also used for the uniform input distribution.
μ	Uniform probability measure on S^1, $d\mu(\varphi) = d\varphi/(2\pi)$.
$L^2(\mu)$	Space of square-integrable functions on S^1 with respect to μ, with inner product $\langle \cdot, \cdot \rangle_{L^2(\mu)}$.
$\mathcal{H}$	A reproducing kernel Hilbert space (RKHS).
$\mathcal{H}_\theta$	Tangent parameter space ($\mathbb{R}^p$ in finite dimension; its limit in the infinite-width case).
Ω	The angular domain $[0, 2\pi)$; in the mean-field discussion, the neuron parameter space $\mathbb{R} \times \mathbb{R}^d$.
$\|\cdot\|_2$, $\|\cdot\|_F$, $\|\cdot\|_{\max}$	Euclidean (ℓ^2) norm, Frobenius norm, and entrywise max norm, respectively.
$\langle \cdot, \cdot \rangle$	Euclidean / Hilbert-space inner product.
$\mathbb{E}$, $\mathbb{P}$	Expectation and probability.

(Notation, continued)

$\mathcal{N}(0,\sigma^2 I_d)$	Gaussian distribution with mean 0 and variance σ^2; I_d denotes the $d \times d$ identity.

Network, data, and training

$f(x;\theta),\ f_\theta$	Neural network output at input x with parameter vector θ.
$f_m(x;\theta)$	One-hidden-layer ReLU network of width m in NTK parameterization, $f_m(x;\theta) = \frac{1}{\sqrt{m}} \sum_{i=1}^m a_i \sigma(w_i^\top x + b_i)$.
θ	Parameter vector, $\theta \in \mathbb{R}^p$; θ_0 its value at initialization.
p	Number of network parameters.
m	Network width (number of hidden units); also w as a width label in figures.
d	Input dimension; also the dimension $d = 2K+1$ of the truncated Fourier block $\mathcal{H}_K$.
$a_i,\ w_i,\ b_i$	Output weight, input weight vector, and bias of hidden unit i; initialized $a_i \sim \mathcal{N}(0,1)$, $w_i \sim \mathcal{N}(0, I_2)$, $b_i(0) = 0$.
$v_\alpha,\ w_\alpha$	Second-layer and first-layer parameters of neuron α in the generic two-layer network.
$\sigma,\ \varphi(\cdot)$	Activation function; $\sigma(t) = t_+ = \max\{t, 0\}$ is ReLU (φ also denotes a generic activation, and elsewhere the circle angle).
$W,\ v$	Hidden-layer weight matrix and output-weight vector of a two-layer network.
$\odot$	Hadamard (elementwise) product.
$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$	Training dataset of n i.i.d. samples.
n	Number of training samples / sampled points.
X	Data matrix ($X \in \mathbb{R}^{d \times n}$, columns x_i).
y	Vector / function of targets.
$\ell(\cdot,\cdot)$	Loss function.
$\mathcal{L},\ L$	Empirical (least-squares) risk; $\mathcal{L}_\rho$ denotes the population risk.
η	Gradient-descent step size / learning rate.
$t,\ k$ (subscript)	Training time / iteration index.
∇_θ	Gradient with respect to θ; $\dot{(\cdot)}$ denotes time derivative under gradient flow.
$\theta^\star,\ w^\star,\ f^\star$	Minimizers / fixed points of the relevant objective.
$A^+,\ A^*$	Moore–Penrose pseudoinverse and adjoint of an operator/matrix A.
$\mathrm{range}(\cdot),\ \ker(\cdot)$	Range (column space) and kernel (null space) of an operator.
$\lambda_{\max}(\cdot)$	Largest eigenvalue of a matrix.

(Notation, continued)

Neural Tangent Kernel and operators

$\phi(x),\ \Phi_t(x)$	Tangent feature map $\nabla_\theta f(x;\theta_0)$ (and its time-t version).
$\Theta_0(x,x')$	Neural Tangent Kernel at initialization, $\nabla_\theta f(x;\theta_0)^\top \nabla_\theta f(x';\theta_0)$.
$\Theta_t,\ \Theta_t^{(m)}$	NTK at training time t (the empirical, width-m kernel).
$\Theta_0^{(m)}$	Finite-width empirical NTK at initialization.
$\bar{\Theta},\ \Theta$	Limiting (infinite-width) deterministic NTK; also written as the Gram matrix $\bar{\Theta} = (\bar{\Theta}(x_i,x_j))_{i,j}$.
$\Theta(\delta)$	Limiting NTK on S^1 as a function of the angular distance $\delta = d(\varphi,\psi) \in [0,\pi]$.
$\Theta^{(\mathrm{nb})}$	Limiting NTK with the trainable hidden-bias contribution removed.
$\Theta^{(r)},\ \psi_r$	One-neuron kernel contribution and one-neuron tangent feature of neuron r.
J_0	Jacobian of network outputs w.r.t. parameters at initialization; $\bar{\Theta} = J_0 J_0^*$.
$T_{\bar{\Theta}},\ T,\ T_\infty$	Integral operator on $L^2(\mu)$ induced by the limiting kernel; $T_\infty = \lim_{m\to\infty} T_0^{(m)}$.
$T_t^{(m)},\ T_0^{(m)}$	Empirical tangent integral operator at time t and at initialization (frozen operator).
$T^{(r)}$	One-neuron integral operator; $T_0^{(m)} = \frac{1}{m}\sum_{r=1}^m T^{(r)}$.
T_p	Density-weighted integral operator for a non-uniform input density $p(\theta)$.
κ	The 2π-periodic kernel profile with $\bar{\Theta}(\theta,\theta') = \kappa(\theta-\theta')$; also a uniform kernel bound and (in the conditioning discussion) a matrix condition number.
$r_t,\ r$	Residual $f_t - y$ (function or vector).

Fourier basis, spectrum, and subspaces

$\theta,\ \varphi,\ \gamma$	Angular coordinate on S^1; γ is used as the angle variable in target functions.
$\phi_q(\theta) = e^{iq\theta}$	Complex Fourier mode of frequency $q \in \mathbb{Z}$.
$\phi_0,\ \phi_{k,c},\ \phi_{k,s}$	Real Fourier basis of $L^2(\mu)$: $\phi_0 = 1$, $\phi_{k,c} = \sqrt{2}\cos(k\varphi)$, $\phi_{k,s} = \sqrt{2}\sin(k\varphi)$.
ψ_j	Generic orthonormal eigenfunction of an integral operator.
$k,\ q$	Fourier frequency index; $k(p)$ is the frequency of basis function ϕ_p.
K	Frequency cutoff defining the retained low-frequency block.

(Notation, continued)

$\mathcal{H}_K$	Truncated Fourier subspace $\operatorname{span}\{\phi_0, \phi_{k,c}, \phi_{k,s} : 1 \le k \le K\}$, of dimension $d = 2K+1$.
$\mathcal{F}_k$	Frequency-k subspace: $\mathcal{F}_0 = \operatorname{span}\{\phi_0\}$, $\mathcal{F}_k = \operatorname{span}\{\phi_{k,c}, \phi_{k,s}\}$ for $k \ge 1$.
d_k	Dimension of $\mathcal{F}_k$: $d_0 = 1$, $d_k = 2$ for $k \ge 1$.
λ_q, λ_k	Fourier eigenvalue of T at frequency q (resp. k); $\lambda_p := \lambda_{k(p)}$.
$\lambda_k^{(\mathrm{nb})}$	Fourier eigenvalue of the no-bias kernel $\Theta^{(\mathrm{nb})}$.
Λ	Diagonal matrix of eigenvalues $\operatorname{diag}(\lambda_p)$; in the linear model, $A = Q\Lambda Q^\top$.
Λ_K	Continuum reference block $P_K T_\infty P_K = \operatorname{diag}(\lambda_0, \lambda_1, \lambda_1, \ldots, \lambda_K, \lambda_K)$.
$\alpha_p(t)$	Amplitude of the residual along Fourier mode ϕ_p; decays as $e^{-\lambda_p t}\alpha_p(0)$.
$a_j(t)$	Coefficient of the residual along eigenfunction ψ_j in the continuum operator analysis.

Finite-sample objects

$\varphi_1, \ldots, \varphi_n$	I.i.d. angles sampled uniformly on $[0, 2\pi)$.
Φ	Sampled Fourier feature matrix, $\Phi \in \mathbb{R}^{n \times d}$, $\Phi_{i,p} = \phi_p(\varphi_i)$; Φ_p its p-th column.
A	Normalized sampled kernel matrix, $A_{ij} = \frac{1}{n}\Theta(\varphi_i - \varphi_j)$.
G	Sampled Fourier Gram matrix, $G = \frac{1}{n}\Phi^\top\Phi$.
H	Compressed operator matrix, $H = \frac{1}{n}\Phi^\top A \Phi$; $\mathbb{E}[H] = \Lambda^{(n)}$.
$\lambda_p^{(n)}$	Finite-sample corrected Fourier eigenvalue, $(1 - \frac{1}{n})\lambda_p + \frac{\Theta(0)}{n}$.
$\Lambda^{(n)}$	Diagonal matrix $\operatorname{diag}(\lambda_p^{(n)})$ of corrected eigenvalues.
C, $C_n^{(K)}$	Coefficient-level restricted operator $G^{-1}H$ on the retained block; compared against $\Lambda^{(n)}$.
P_{th}, P_{emp}	Theory and empirical Fourier-block preconditioners.
B_{base}, B_{th}, B_{emp}	Baseline, theory-preconditioned, and empirically preconditioned sampled operators.
τ	Ridge-regularization parameter for ill-conditioned inverses.
u	Unit roundoff (machine precision); the floor κu sets the achievable relative accuracy.
$\widehat{r}(t)$	Sampled residual vector.
$\mathbf{1}$	All-ones vector in $\mathbb{R}^d$.

Finite-width objects and diagnostics

$B_m^{(K)}$	Frozen finite-width tangent block on $\mathcal{H}_K$, $P_K T_0^{(m)} P_K$; $\mathbb{E}[B_m^{(K)}] = \Lambda_K$.

(Notation, continued)

$C_t^{(K)}$	Evolving tangent block on $\mathcal{H}_K$, $P_K T_t^{(m)} P_K$; $C_0^{(K)} = B_m^{(K)}$.
$L_t^{(K)}$	High-to-low frequency coupling (leakage) operator, $P_K T_t^{(m)} (I - P_K)$.
$a_t,\ b_t$	Low- and high-frequency residual components, $a_t = P_K r_t$, $b_t = (I - P_K) r_t$.
$\xi_{r,pq}$	One-neuron matrix entry $\langle \phi_p, T^{(r)} \phi_q \rangle$.
P_K	Orthogonal projection $L^2(\mu) \to \mathcal{H}_K$.
$P_k,\ U_k$	Orthogonal projector onto the grid representation of $\mathcal{F}_k$, $P_k = U_k U_k^\top$, with U_k an orthonormal basis.
$\widehat{\mathcal{F}}_k(t),\ \widehat{P}_k(t),\ \widehat{U}_k(t)$	Matched empirical eigenspace of the tangent operator, its projector and orthonormal basis.
$v_j(t)$	Orthonormal eigenvectors of the grid-discretized tangent operator.
$N,\ \theta_j$	Number of uniform grid points and grid angles $\theta_j = 2\pi j/N$, $j = 0, \ldots, N-1$.
$\varepsilon_{k,F}^{(m)}(t)$	Projector error $\|\widehat{P}_k(t) - P_k\|_F$ for frequency k (frozen value at $t = 0$ written $\varepsilon_{k,F}^{(m)}$).
$\theta_{k,i}(t)$	i-th principal angle between $\widehat{\mathcal{F}}_k(t)$ and $\mathcal{F}_k$; $\cos\theta_{k,i} = s_i(U_k^\top \widehat{U}_k)$.
$s_i(\cdot)$	i-th largest singular value.
$s_k^{\cos}(t)$	Subspace cosine similarity, root-mean-square of the principal-angle cosines for frequency k.
$\mu_{\mathcal{F}_k}(t)$	Average spectral strength inside Fourier block $\mathcal{F}_k$, $\frac{1}{d_k} \operatorname{tr}(P_k T_t^{(m)} P_k)$.
$\mu_{\le K}(t)$	Cumulative low-frequency spectral strength over the retained block.
$\mu_m,\ \mu_t,\ \delta_{\theta_\alpha}$	Empirical neuron measure, its limit, and the Dirac mass at θ_α (mean-field discussion).

Targets and miscellaneous

$f_{\text{orig}},\ f_{\text{eq}},\ f_{\text{rev}}$	Original, equal-amplitude, and reversed-amplitude target functions used in the amplitude-variant experiments.
$A_i,\ k_i,\ \phi_i$	Amplitude, frequency, and phase of the i-th sinusoid in a synthetic Fourier-regression target.
c	Universal constant in the finite-width concentration bound (illustrative value $c = 0.2$).
δ	Failure probability in high-probability bounds; also the angular distance $d(\varphi, \psi)$ on S^1.
ε	Deviation level in concentration inequalities.
$R,\ R_\alpha$	Rotation matrix on $\mathbb{R}^2$ (rotation by angle α).

Contents

Chapter 1

Introduction

Deep neural networks have achieved remarkable empirical success across a wide range of application domains, most notably in computer vision and natural language processing. In vision, deep convolutional and residual architectures have led to dramatic improvements on large-scale benchmarks, establishing deep learning as the dominant paradigm for visual recognition tasks (Krizhevsky et al., 2012; He et al., 2016). Similarly, in natural language processing, attention-based Transformer architectures introduced by Vaswani et al. (2023) have enabled substantial advances in sequence modeling and representation learning, and now form the foundation of modern large-scale language models. Over the past decade, this empirical progress has been accompanied by a steady increase in model size, depth, and computational scale, often resulting in highly overparameterized models that generalize well despite their capacity to fit random labels (Zhang et al., 2017; Belkin et al., 2019). Much of this progress has been driven by empirical experimentation and architectural intuition rather than by a complete theoretical understanding of the mechanisms underlying training and generalization in deep neural networks.

Despite their practical success, the mechanisms governing the training dynamics and convergence behavior of neural networks remain only partially understood, even in highly simplified settings (Jacot et al., 2020; Mei et al., 2018; Chizat and Bach, 2018). From an optimization perspective, neural network training involves high-dimensional, non-convex objectives whose geometry depends intricately on architectural choices, initialization, and optimization dynamics (Bottou et al., 2018). Early analyses of neural network loss landscapes have highlighted the prevalence of saddle points and complex critical structures, underscoring the difficulty of directly characterizing training dynamics in parameter space (Choromanska et al., 2015). While more recent theoretical work has made progress in analyzing overparameterized models under restrictive assumptions such as two-layer networks or specific scaling regimes these results do not yet provide a unified explanation of neural network training in general settings (Du et al., 2019; Arora et al., 2019). More broadly, this gap between empirical performance and theoretical understanding has motivated the study of neural networks through simplified models and asymptotic limits, where training dynamics can be analyzed more precisely.

Structure of the thesis. This thesis aims to contribute to the theoretical study of neural-network training by focusing on one such simplified setting in which the learning dynamics can be analyzed explicitly. We begin by introducing the supervised-learning framework and empirical risk minimization, which provide the basic language for describing how models are trained from data. We then consider increasingly expressive model classes, moving from linear models to deep linear networks and nonlinear neural networks, to show where training dynamics can be characterized cleanly and where this understanding begins to break down. This progression motivates the use of simplified theoretical regimes: settings that are restrictive enough to permit precise analysis, but still rich enough to capture phenomena observed in neural-network training. After this foundation, we review existing theoretical approaches for studying wide neural networks. This discussion naturally leads to spectral bias, the empirical and theoretical observation that neural networks often learn smooth, low-frequency components of a target function before more oscillatory, high-frequency components. Spectral bias is the main object of study in this thesis. The related work then summarizes existing empirical and theoretical explanations of this phenomenon, leading to the research gap and the research questions addressed in the remainder of the thesis.

1.1 Problem setup

The statistical learning framework. The fundamental objective in supervised learning is to identify a functional relationship between an input space $\mathcal{X} \subseteq \mathbb{R}^d$ and a target space $\mathcal{Y}$, for instance $\mathcal{Y} = \mathbb{R}^n$ in regression or $\mathcal{Y} = \{1, \dots, C\}$ in classification where C denotes the number of classes. We assume the existence of a joint probability distribution ρ over $\mathcal{X} \times \mathcal{Y}$. Ideally, one seeks a measurable function $f : \mathcal{X} \to \mathcal{Y}$, that minimizes the expected risk (or population risk):

$$\mathcal{L}_\rho(f) = \mathbb{E}_\rho \ell(f(x), y) \tag{1.1}$$

where $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty]$ is a loss function quantifying the discrepancy between the prediction and the true target. (Vapnik, 1999; Shalev-Shwartz and Ben-David, 2014) In practice, two fundamental constraints prevent the direct minimization of the equation (1.1). First, searching over the space of all possible measurable functions is computationally and analytically intractable, so we usually restrict our search to a specific hypothesis class which is a family of functions f_θ parameterized by a vector $\theta \in \mathbb{R}^p$. Second, the underlying distribution ρ is unknown; we only have access to a finite training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ of n samples drawn i.i.d. from ρ, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$.

Empirical risk minimization. The second constraint leads us to the principle of empirical risk minimization (ERM). Given a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, a measurable function f_θ parameterized by θ, and a loss function $\ell(\cdot, \cdot)$, ERM seeks parameters that minimize the empirical risk,

$$\theta^\star \in \arg\min_\theta \frac{1}{n} \sum_{i=1}^n \ell(f_\theta(x_i), y_i). \tag{1.2}$$

This formulation isolates the key ingredients that determine the behavior of learning algorithms: the function class induced by the parameterization $\theta \mapsto f_\theta$, the geometry of the loss

landscape, and the optimization procedure used to minimize the empirical risk. In modern neural network training, the empirical risk is typically minimized using gradient-based methods (Bottou et al., 2018), which induce a dynamical system in parameter space whose properties depend intricately on both the model architecture and the chosen loss function.

Linear models and gradient descent. To build intuition, we first consider the case of models that are linear in inputs and in parameters. We define a function $f_w(x) : \mathbb{R}^d \to \mathbb{R}$, parameterized by $w \in \mathbb{R}^d$ with the prescription $f_w(x) = x^\top w$. We specialize the loss function to the squared error, $\ell(f(x), y) = \frac{1}{2}(f_w(x) - y)^2$. Let $X \in \mathbb{R}^{d \times n}$ denote the data matrix whose columns are $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$, and let $y \in \mathbb{R}^n$ denote the vector of targets. With slight abuse of notation, the ERM problem simplifies to a least-squares objective:

$$\arg\min_w \ \mathcal{L}(w) = \frac{1}{2}\sum_{i=1}^{n}(f_w(x_i) - y_i)^2 = \frac{1}{2}\left\|X^\top w - y\right\|_2^2.$$

This objective is convex and differentiable, with gradient

$$\nabla_w \mathcal{L}(w) = X(X^\top w - y) = XX^\top w - Xy. \tag{1.3}$$

At a stationary point $w^\star$, the gradient vanishes, yielding the equations

$$XX^\top w^\star = Xy, \tag{1.4}$$

which characterize the set of global minimizers of $\mathcal{L}$ owing to convexity. The structure of the minimizers depends on the rank of the empirical covariance operator $A := XX^\top$. If A is positive definite (equivalently, if X has full row rank), then the minimizer is unique and given by $w^\star = A^{-1}Xy$. In contrast, in the overparameterized regime where $\operatorname{rank}(X) < d$, the matrix A is singular and the least-squares objective admits infinitely many global minimizers (Boyd and Vandenberghe, 2004; Golub and Van Loan, 2013). Among these minimizers, the Moore–Penrose pseudoinverse selects the one with smallest Euclidean norm. Since A is symmetric, its pseudoinverse A^+ acts as the inverse on $\operatorname{range}(A)$ and annihilates $\ker(A)$ (Penrose, 1955; Golub and Van Loan, 2013), yielding the minimum-norm least-squares solution

$$w^\star_{\min} = A^+ Xy.$$

Here minimum-norm means that $w^\star_{\min}$ minimizes $\|w\|_2$ among all global minimizers of $\mathcal{L}$ (Gunasekar et al., 2018). Gradient descent with step size $\eta > 0$ takes the explicit form

$$w_{t+1} = w_t - \eta \nabla_w \mathcal{L}(w_t) = \left(I - \eta XX^\top\right) w_t + \eta Xy. \tag{1.5}$$

Assuming $w_0 = 0$, we can write a closed form solution

$$w_t = \eta \sum_{s=0}^{t-1} \left(I - \eta XX^\top\right)^s Xy. \tag{1.6}$$

Let $\lambda_{\max}(A)$ denote the largest eigenvalue of A. If the step size satisfies $\eta \in (0, 2/\lambda_{\max}(A))$, then for any least-squares minimizer $w^\star$ the error $e_t := w_t - w^\star$ evolves as

$$e_{t+1} = (I - \eta A)e_t. \tag{1.7}$$

Since A is symmetric, it admits an eigendecomposition

$$A = Q\Lambda Q^\top,$$

where $Q \in \mathbb{R}^{d\times d}$ is orthogonal ($Q^\top Q = I$) and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ contains the (real) eigenvalues of A. Moreover, A is positive semidefinite because $v^\top A v = \|X^\top v\|_2^2 \geq 0$ for all v, hence $\lambda_i \geq 0$ and in particular $\lambda_i \in [0, \lambda_{\max}(A)]$.

Starting from the error recursion

$$e_{t+1} = (I - \eta A)e_t,$$

we express the error in the eigenbasis of A by defining $\tilde{e}_t := Q^\top e_t$. Left-multiplying by $Q^\top$ and using $e_t = Q\tilde{e}_t$ gives

$$\tilde{e}_{t+1} = Q^\top (I - \eta A) Q \tilde{e}_t = \big(I - \eta Q^\top A Q\big)\tilde{e}_t = \big(I - \eta\Lambda\big)\tilde{e}_t.$$

Since Λ is diagonal, this yields the component-wise dynamics

$$\tilde{e}_{t+1,i} = (1 - \eta\lambda_i)\tilde{e}_{t,i}, \qquad i = 1, \ldots, d. \tag{1.8}$$

If $\eta \in (0, 2/\lambda_{\max}(A))$, then for any $\lambda_i > 0$ we have $0 < \eta\lambda_i \leq \eta\lambda_{\max}(A) < 2$, hence $-1 < 1 - \eta\lambda_i < 1$, i.e. $|1 - \eta\lambda_i| < 1$. Under the stated condition on η, we have $|1 - \eta\lambda_i| < 1$ for all $\lambda_i > 0$, implying geometric decay of all components of e_t in range(A). In particular, if A is positive definite, then $w_t \to w^\star$ as $t \to \infty$. More generally, if $w_0 \in \mathrm{range}(A)$ (for example, $w_0 = 0$), then w_t converges to the minimum-norm least-squares solution $w^\star = A^+ X y$.

The eigen-decomposition above shows that gradient descent acts as a linear dynamical system whose behavior is governed by the spectrum of the empirical covariance operator $XX^\top$. Directions corresponding to larger eigenvalues converge more rapidly, while components in the null space of $XX^\top$ remain unchanged. In the overparameterized regime, where multiple *interpolating solutions* exist, i.e., solutions satisfying $X^\top w = y$, gradient descent from standard initializations converges to a particular interpolating solution, namely the minimum-norm solution (Boyd and Vandenberghe, 2004; Golub and Van Loan, 2013). This illustrates how the optimization algorithm induces an implicit bias even in this simplest linear setting.

This simple example already contains the basic spectral mechanism that will reappear throughout the thesis: gradient descent does not reduce all components of the error at the same speed; instead, the decay rate of each component is determined by the eigenvalue of the operator governing the dynamics. In the linear model above, these eigendirections are determined by the empirical covariance matrix $XX^\top$, so they do not yet have to correspond to frequencies.

Deep linear network. Even in the absence of nonlinear activation functions, introducing an additional layer changes the parameterization of the model, even though the resulting input-output map remains linear. This already leads to substantially more intricate training dynamics (Saxe et al., 2014). Consider a two-layer linear network with scalar output,

$$f_{v,W}(x) := v^\top W x, \tag{1.9}$$

where $W \in \mathbb{R}^{m\times d}$, $v \in \mathbb{R}^m$, and m denotes the width of the hidden layer. The least-squares empirical risk is

$$\mathcal{L}(v,W) := \frac{1}{2}\left\|X^\top W^\top v - y\right\|_2^2. \tag{1.10}$$

Although the resulting function is linear in the input, the loss is no longer convex in the parameters (v,W) due to their multiplicative coupling.

Let

$$r := X^\top W^\top v - y \in \mathbb{R}^n \tag{1.11}$$

denote the residual vector. The gradients of $\mathcal{L}$ are given by

$$\nabla_v \mathcal{L} = WXr, \qquad \nabla_W \mathcal{L} = v(Xr)^\top. \tag{1.12}$$

Similar to the linear case, the parameters of this model are updated iteratively using gradient descent with step size $\eta > 0$

$$v_{k+1} = v_k - \eta \nabla_v \mathcal{L}(v_k, W_k), \qquad W_{k+1} = W_k - \eta \nabla_W \mathcal{L}(v_k, W_k).$$

The analysis of learning dynamics is often simplified by considering the infinitesimal learning rate limit ($\eta \to 0$). In this regime, the discrete updates converge to a system of coupled ordinary differential equations (ODEs) known as gradient flow.

$$\dot{v}_t = -\nabla_v \mathcal{L}(v_t, W_t) = W_t X\left(y - X^\top W_t^\top v_t\right), \quad \dot{W}_t = -\nabla_W \mathcal{L}(v_t, W_t) = v_t\left(y - X^\top W_t^\top v_t\right)^\top X^\top. \tag{1.13}$$

This coupled system consists of a vector-valued and a matrix-valued ordinary differential equation and is nonlinear in the parameters, despite the underlying predictor being linear in the input. As a result, the training dynamics no longer admit a simple spectral characterization as in the one-layer case.

Nevertheless, recent work has shown that deep linear networks admit a more structured description in suitable asymptotic regimes. In particular, Chizat et al. (2024) study gradient flow for deep linear networks in an infinite-width limit, where the dynamics are described at the level of a measure-valued evolution. While no closed-form solution of the finite-dimensional ODE system is obtained, this framework establishes global well-posedness, convergence to global minimizers, and the emergence of implicit regularization effects induced by depth (Chizat et al., 2024).

Nonlinear networks. The analysis of training dynamics becomes more involved once nonlinear activation functions are introduced. Consider again a two-layer network of the form

$$f_{v,W}(x) = v^\top \phi(Wx),$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable activation function applied elementwise. For the least-squares objective, the continuous-time evolution of the parameters is governed by the gradient flow system:

$$\dot{v}_t = \phi(W_t X)(y - \phi(W_t X)^\top v_t), \qquad \dot{W}_t = \left(\phi(W_t X) \odot \left(v_t (y - \phi(W_t X)^\top)^\top\right)\right) X^\top, \tag{1.14}$$

where $\odot$ denotes the Hadamard (elementwise) product. The gradient flow equations involve both the activations $\phi(Wx)$ and their derivatives $\phi'(Wx)$. In particular, the evolution of the parameters mixes standard matrix products with elementwise nonlinear operations. As a consequence, the training dynamics of nonlinear networks do not admit a simple closed-form or spectral description in general.

Nevertheless, the linear example suggests the kind of structure one would like to recover: an operator-level description in which different components of the error decay at different rates. For nonlinear neural networks, identifying such an operator and understanding its spectrum is substantially more difficult, because the relevant dynamics depend on the evolving parameters of the network. This is one reason why studying frequency-dependent learning, or spectral bias, is non-trivial in general neural networks. The remainder of the thesis therefore focuses on simplified regimes in which an operator governing the dynamics can be made explicit, allowing the connection between its spectrum and the observed learning behavior to be analyzed more precisely.

1.2 Thesis motivation

At a high level, this thesis is motivated by the question of how these networks' training dynamics can be studied in a mathematically tractable setting without discarding the phenomena one ultimately seeks to understand. The objective is not to provide a complete description of neural network training in full generality, but rather to identify a setting in which the dynamics can be analyzed explicitly.

This motivates the study of analytical regimes in which the dynamics become tractable. Although such regimes necessarily involve simplifications like taking large-width limits, they can still provide useful insight into mechanisms that are otherwise difficult to access in the full non-linear parameter dynamics. Two prominent examples of such approaches are the Neural Tangent Kernel (NTK) and mean-field formulations, which offer complementary perspectives on the behavior of wide neural networks during training.

In this thesis, that regime is a single-hidden-layer ReLU network with inputs on the unit circle (1-sphere S^1). In the infinite-width limit, the NTK defines a deterministic integral operator. For the uniform distribution on the circle, this operator is diagonalized by the Fourier basis (the orthonormal basis of trigonometric functions: $\sin(k\theta), \cos(k\theta)$). This gives a clean reference picture for spectral bias: each Fourier mode (specific harmonic i.e. frequency k) has its own learning rate. The rest of the thesis studies how this picture changes at finite sample size and finite width.

Chapter 2

Theoretical Background

2.1 Neural Tangent Kernel

As discussed in Chapter 1, for nonlinear neural networks the gradient flow dynamics are parameter-dependent and do not admit a simple closed-form description. One approach to recovering a tractable analytical framework is to consider regimes in which the network behaves approximately linearly around its random initialization. The *Neural Tangent Kernel* (NTK), introduced by Jacot et al. (2020), formalizes this idea by studying training dynamics through a first-order linearization in parameter space, leading to a kernel-based description of learning in wide neural networks.

Linearization around initialization Let $f(x;\theta)$ denote a neural network with parameters $\theta \in \mathbb{R}^p$, initialized at θ_0. A first-order Taylor expansion around θ_0 yields

$$f(x;\theta) \;\approx\; f(x;\theta_0) + \nabla_\theta f(x;\theta_0)^\top (\theta - \theta_0), \tag{2.1}$$

where $f(x;\theta_0)$ is the network output at initialization and

$$\phi(x) := \nabla_\theta f(x;\theta_0)$$

defines the tangent feature map. Locally around initialization, the network thus behaves as a linear model in parameter space,

$$f(x;\theta) \approx f(x;\theta_0) + \phi(x)^\top (\theta - \theta_0). \tag{2.2}$$

Definition 2.1 (Neural Tangent Kernel (Jacot et al., 2020))**.** Given an initialization θ_0, the neural tangent kernel is defined as

$$\Theta_0(x,x') := \nabla_\theta f(x;\theta_0)^\top \nabla_\theta f(x';\theta_0).$$

The NTK measures how infinitesimal parameter updates couple the network outputs at different inputs and plays the role of an empirical covariance operator in the tangent feature space.

Training dynamics induced by the NTK We consider training with the squared loss using gradient descent with step size $\eta > 0$,

$$L(\theta) = \frac{1}{2}\sum_{i=1}^{n}\big(f(x_i;\theta) - y_i\big)^2, \qquad \theta_{t+1} = \theta_t - \eta\nabla_\theta L(\theta_t), \qquad t = 0,1,2,\dots$$

Define $f_t(x) := f(x;\theta_t)$. The gradient of the loss is

$$\nabla_\theta L(\theta_t) = \sum_{j=1}^{n}\big(f_t(x_j) - y_j\big)\,\nabla_\theta f(x_j;\theta_t).$$

Applying a first-order Taylor expansion of $f(\cdot;\theta)$ around θ_t gives

$$f_{t+1}(x_i) \approx f_t(x_i) + \nabla_\theta f(x_i;\theta_t)^\top(\theta_{t+1} - \theta_t).$$

Substituting the gradient descent update $\theta_{t+1} - \theta_t = -\eta\nabla_\theta L(\theta_t)$ yields

$$f_{t+1}(x_i) = f_t(x_i) - \eta\sum_{j=1}^{n}\nabla_\theta f(x_i;\theta_t)^\top\nabla_\theta f(x_j;\theta_t)\big(f_t(x_j) - y_j\big),$$

which can be written compactly as

$$f_{t+1}(x_i) = f_t(x_i) - \eta\sum_{j=1}^{n}\Theta_t(x_i,x_j)\big(f_t(x_j) - y_j\big), \tag{2.3}$$

where the iteration-dependent NTK is

$$\Theta_t(x_i,x_j) := \nabla_\theta f(x_i;\theta_t)^\top\nabla_\theta f(x_j;\theta_t).$$

If we stack predictions into $f_t := (f_t(x_1),\dots,f_t(x_n))^T \in \mathbb{R}^n$, in vector form, the dynamics read

$$f_{t+1} = f_t - \eta\,\Theta_t(f_t - y). \tag{2.4}$$

In general, Θ_t evolves during training, reflecting changes in the network's tangent features.

One-hidden-layer NTK and infinite-width limit To make the NTK explicit, consider a two-layer network of width m,

$$f(x) = \frac{1}{\sqrt{m}}\sum_{\alpha=1}^{m} v_\alpha\,\varphi(w_\alpha^\top x + b_\alpha), \qquad v_\alpha, b_\alpha \in \mathbb{R}, \quad w_\alpha, x \in \mathbb{R}^d. \tag{2.5}$$

Each neuron is parameterized by $\theta_\alpha = (v_\alpha, w_\alpha, b_\alpha)$ and randomly initialized as

$$v_\alpha \sim \mathcal{N}(0,\sigma_v^2), \qquad w_\alpha \sim \mathcal{N}\big(0,\sigma_w^2 I_d\big), \qquad b_\alpha \sim \mathcal{N}(0,\sigma_b^2),$$

independently across $\alpha = 1,\dots,m$. $\varphi : \mathbb{R} \to \mathbb{R}$ is a non-linear activation function.

Finite-width NTK at initialization. By direct computation,

$$\nabla_{v_\alpha} f(x) = \frac{1}{\sqrt{m}} \varphi(w_\alpha^\top x + b_\alpha), \tag{2.6}$$

$$\nabla_{w_\alpha} f(x) = \frac{1}{\sqrt{m}} v_\alpha \varphi'(w_\alpha^\top x + b_\alpha) x, \tag{2.7}$$

$$\nabla_{b_\alpha} f(x) = \frac{1}{\sqrt{m}} v_\alpha \varphi'(w_\alpha^\top x + b_\alpha). \tag{2.8}$$

Substituting into the definition of the NTK yields the empirical kernel

$$\Theta_0^{(m)}(x,x') = \frac{1}{m}\sum_{\alpha=1}^{m} \varphi(w_\alpha^\top x + b_\alpha)\,\varphi(w_\alpha^\top x' + b_\alpha) + \frac{1}{m}\sum_{\alpha=1}^{m} v_\alpha^2\,\varphi'(w_\alpha^\top x + b_\alpha)\,\varphi'(w_\alpha^\top x' + b_\alpha)\,\big(x^\top x' + 1\big). \tag{2.9}$$

Infinite-width limit. Each sum in (2.9) is an empirical average of i.i.d. terms. By the strong law of large numbers,

$$\Theta_0^{(m)}(x,x') \xrightarrow{\text{a.s.}} \bar{\Theta}(x,x') \qquad \text{as } m \to \infty,$$

where the limiting NTK is deterministic and given by

$$\bar{\Theta}(x,x') = \mathbb{E}_{w,b}\big[\varphi(w^\top x + b)\,\varphi(w^\top x' + b)\big] + \sigma_v^2\,\big(x^\top x' + 1\big)\,\mathbb{E}_{w,b}\big[\varphi'(w^\top x + b)\,\varphi'(w^\top x' + b)\big]. \tag{2.10}$$

This argument extends to deep fully connected networks, where both the forward covariance kernel and the NTK satisfy recursive layerwise equations in the infinite-width limit (Jacot et al., 2020; Lee et al., 2020).

2.1.1 Constant-kernel regime and spectral dynamics

In the infinite-width limit, the NTK remains constant throughout training, $\Theta_t \equiv \bar{\Theta}$. Equation (2.4) then reduces to the linear iteration

$$f_{t+1} = f_t - \eta\,\bar{\Theta}(f_t - y). \tag{2.11}$$

Letting $r_t := f_t - y$, we obtain

$$r_{t+1} = (I - \eta\bar{\Theta})\,r_t.$$

Assuming $\eta < 2/\lambda_{\max}(\bar{\Theta})$, the residual admits the closed form

$$r_t = (I - \eta\bar{\Theta})^t r_0, \qquad f_t = y + (I - \eta\bar{\Theta})^t (f_0 - y). \tag{2.12}$$

As in the linear setting of Chapter 1, convergence occurs independently along the eigenvectors of $\bar{\Theta}$, with geometric rates determined by the corresponding eigenvalues.

NTK dynamics, solution structure, and implicit bias In the constant-kernel (infinite-width) regime, the NTK prediction dynamics reduce to a linear iteration in prediction space. Equation (2.11), admits fixed points $f^\star$ satisfying

$$\bar{\Theta}(f^\star - y) = 0. \tag{2.13}$$

Equation (2.13) characterizes the set of global minimizers of the squared loss in the NTK regime. As in linear least squares, the structure of this solution set depends on the rank of the operator $\bar{\Theta}$.

If $\bar{\Theta}$ is strictly positive definite on the training set, the solution is unique and satisfies $f^\star = y$. When $\bar{\Theta}$ is singular, the loss admits infinitely many interpolating solutions in prediction space, differing by elements of $\ker(\bar{\Theta})$. In this case, gradient descent converges to a particular fixed point determined by the initialization.

Tangent features and operator factorization. To formulate the NTK in a way that remains meaningful in the infinite-width limit, it is convenient to view parameter perturbations as elements of a Hilbert space $\mathcal{H}_\theta$ equipped with an inner product $\langle \cdot, \cdot \rangle$. In the finite-dimensional case, $\mathcal{H}_\theta = \mathbb{R}^p$ with the Euclidean inner product; in the infinite-width limit, $\mathcal{H}_\theta$ denotes the corresponding limit space of parameter perturbations, often referred to as the tangent parameter space (Jacot et al., 2020; Lee et al., 2020).

In this setting, the limiting NTK Gram matrix on the training set can be expressed as

$$\bar{\Theta} = J_0 J_0^*, \tag{2.14}$$

where $J_0 : \mathcal{H}_\theta \to \mathbb{R}^n$ denotes the Jacobian of the network outputs with respect to parameters at initialization, viewed as a linear operator,

$$(J_0 h)_i = \langle \nabla_\theta f(x_i; \theta_0), h \rangle,$$

and J_0^* is its adjoint. This representation makes explicit that $\bar{\Theta}$ is symmetric and positive semidefinite, and that it acts as an empirical covariance operator for the tangent features induced by the network at initialization (Jacot et al., 2020).

Implicit bias and representer viewpoint. When $\bar{\Theta}$ is singular, the Moore–Penrose pseudoinverse $\bar{\Theta}^+$ acts as the inverse on $\mathrm{range}(\bar{\Theta})$ and annihilates $\ker(\bar{\Theta})$. Consequently, constant-kernel NTK training eliminates only the residual component lying in $\mathrm{range}(\bar{\Theta})$ while preserving the component in $\ker(\bar{\Theta})$. This behavior reflects an implicit regularization effect analogous to linear least squares and can be interpreted as convergence to a minimum-norm solution in the reproducing kernel space associated with $\bar{\Theta}$, as formalized by representer-theorem results (Bietti and Mairal, 2019; Bartolucci et al., 2021).

Limitations of stable kernels and feature learning. Although the NTK framework provides a clean and tractable description of training dynamics, the stability of the kernel also introduces inherent limitations. When the kernel does not change during training, learning is effectively confined to a fixed feature space. Indeed, any positive semidefinite kernel

admits a representation $K(x,x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$, so that optimization in the NTK regime corresponds to fitting a linear model in the associated reproducing kernel Hilbert space $\mathcal{H}$.

In contrast, for finite-width networks the NTK typically evolves during training, implying that the tangent features $\Phi_t(x) = \nabla_\theta f(x;\theta_t)$ also change over time. This evolution allows the network to adapt its representation to the data, a phenomenon commonly referred to as *feature learning*. Such effects are absent in a strictly stable-kernel regime, where only the coefficients of fixed features are adjusted.

The importance of feature learning is already visible in simplified settings. Even purely linear networks exhibit nontrivial training dynamics due to the coupling of parameters across layers, leading to implicit biases that cannot be captured by a fixed kernel model (Saxe et al., 2014; Tu et al., 2024). From this viewpoint, the NTK regime can be seen as a useful but restrictive approximation, which motivates studying regimes where kernel evolution and feature learning play an explicit role.

2.2 Spectral dynamics on the circle

The constant-kernel NTK regime gives a linear description of training in function space. We now specialize this viewpoint to the setting used throughout the thesis: inputs on the unit circle S^1. This gives a clean reference model in which the operator governing training is diagonalized by Fourier modes.

Let S^1 be parameterized by $\theta \in [0, 2\pi)$, with embedding

$$x(\theta) = (\cos\theta, \sin\theta) \in \mathbb{R}^2,$$

and let ρ denote the uniform probability measure on S^1, $d\rho(\theta) = d\theta/(2\pi)$. Given a limiting NTK $\bar{\Theta}(\theta,\theta')$, define the associated integral operator

$$(T_{\bar{\Theta}} g)(\theta) := \int_0^{2\pi} \bar{\Theta}(\theta,\theta') g(\theta') \frac{d\theta'}{2\pi}. \tag{2.15}$$

In the continuum constant-kernel regime, the residual $r_t(\theta) = f_t(\theta) - f^\star(\theta)$ evolves according to

$$\partial_t r_t = -T_{\bar{\Theta}} r_t. \tag{2.16}$$

If $\{\psi_j\}_{j\geq 0}$ is an orthonormal eigenbasis of $T_{\bar{\Theta}}$, with

$$T_{\bar{\Theta}} \psi_j = \lambda_j \psi_j,$$

then the residual decomposes as

$$r_t = \sum_{j\geq 0} a_j(t)\psi_j, \qquad a_j(t) = \langle r_t, \psi_j \rangle_{L^2(\rho)}.$$

Substituting this expansion into (2.16) gives

$$a_j(t) = e^{-\lambda_j t} a_j(0). \tag{2.17}$$

Thus each eigendirection is learned independently, and the decay rate is determined by the corresponding eigenvalue. Larger eigenvalues correspond to faster decay.

This is the operator form of the NTK training dynamics introduced by Jacot et al. (2020). In the context of spectral bias, Cao et al. (2020) use this eigenfunction viewpoint to show that components aligned with larger NTK eigenvalues are learned faster.

For the circle, the Fourier basis appears when the limiting kernel is rotation-invariant. This is the same symmetry principle used in analyses of NTK spectra on the circle and sphere, where uniform input distributions lead to Fourier modes on S^1 and spherical harmonics on S^{d-1} (Ronen et al., 2019; Bietti and Mairal, 2019; Cao et al., 2020). Under isotropic initialization, the distribution of w is invariant under rotations. Using the infinite-width expression (2.10), this implies

$$\bar{\Theta}(Rx, Rx') = \bar{\Theta}(x, x')$$

for every rotation R. Since $x(\theta + \alpha) = R_\alpha x(\theta)$, the kernel depends only on the angular difference:

$$\bar{\Theta}(\theta, \theta') = \kappa(\theta - \theta') \tag{2.18}$$

for some 2π-periodic function κ.

With the uniform measure, (2.15) becomes a circular convolution:

$$(T_{\bar{\Theta}} g)(\theta) = \int_0^{2\pi} \kappa(\theta - \theta') g(\theta') \frac{d\theta'}{2\pi} = (\kappa * g)(\theta). \tag{2.19}$$

Therefore the complex Fourier modes

$$\phi_q(\theta) = e^{iq\theta}, \qquad q \in \mathbb{Z},$$

are eigenfunctions of $T_{\bar{\Theta}}$. Indeed,

$$(T_{\bar{\Theta}} \phi_q)(\theta) = \int_0^{2\pi} \kappa(u) e^{iq(\theta - u)} \frac{du}{2\pi} \tag{2.20}$$

$$= e^{iq\theta} \int_0^{2\pi} \kappa(u) e^{-iqu} \frac{du}{2\pi} = \lambda_q \phi_q(\theta), \tag{2.21}$$

where

$$\lambda_q = \int_0^{2\pi} \kappa(u) e^{-iqu} \frac{du}{2\pi}. \tag{2.22}$$

When the kernel is real and even, one may equivalently use the real Fourier basis

$$1, \qquad \sqrt{2}\cos(k\theta), \qquad \sqrt{2}\sin(k\theta), \qquad k \geq 1.$$

In this basis, each Fourier frequency subspace evolves independently under the continuum infinite-width operator.

This gives the basic spectral picture behind the NTK explanation of frequency-dependent learning. In the idealized continuum infinite-width setting, the Fourier modes are exact eigenfunctions of the training operator, and their learning rates are determined by the corresponding eigenvalues λ_q. The next section briefly contrasts this fixed-operator view with another large-width limit, before the thesis returns to the NTK framework.

2.3 Another large-width viewpoint: mean-field

The previous section explains why the NTK viewpoint is useful for this thesis. In the constant-kernel regime, training is described by a fixed operator on functions. On S^1, this operator is diagonal in the Fourier basis, so each Fourier mode has a well-defined learning rate. This is the structure needed for a mode-by-mode analysis of spectral bias.

There is another important large-width viewpoint, usually called the *mean-field* viewpoint. It starts from a different scaling of a two-layer network,

$$f_\theta(x) = \frac{1}{m}\sum_{\alpha=1}^{m} v_\alpha \varphi(w_\alpha^\top x).$$

Here each neuron has parameters

$$\theta_\alpha = (v_\alpha, w_\alpha) \in \Omega := \mathbb{R}\times\mathbb{R}^d.$$

Instead of following the m neurons one by one, the mean-field viewpoint tracks how the neurons are distributed in parameter space. This is done through the empirical measure

$$\mu_m := \frac{1}{m}\sum_{\alpha=1}^{m} \delta_{\theta_\alpha},$$

where δ_{θ_α} places one unit of mass at the parameter value θ_α. With this notation, the network output can be written as

$$f_\theta(x) = \int_\Omega v\varphi(w^\top x)\, d\mu_m(v,w).$$

Thus the network is represented by the distribution of its hidden units. During training, this distribution moves through parameter space.

This gives a different limiting description from the NTK regime. In suitable large-width limits, the empirical measure μ_m converges to a limiting time-dependent measure μ_t, and training is described by an evolution equation for μ_t. This perspective was developed for two-layer networks by Mei et al. (2018), and related interacting-particle descriptions were studied by Rotskoff and Vanden-Eijnden (2022). From an optimization viewpoint, Chizat and Bach (2018) studied the corresponding many-particle limit using tools from optimal transport. Extensions and rigorous frameworks for deeper networks have also been studied (Nguyen, 2019; Araújo et al., 2019; Sirignano and Spiliopoulos, 2019; Nguyen and Pham, 2023).

The reason we do not use this viewpoint as the main framework is not that it is unrelated to neural-network training. It answers a different kind of question. Mean-field limits are useful for describing how the hidden units move and how the representation changes during training. In contrast, the spectral-bias question studied here requires an operator acting on functions: we want to ask which Fourier modes are eigenfunctions, what their eigenvalues are, and how the corresponding eigenspaces behave. The NTK viewpoint gives such an operator directly. The mean-field viewpoint describes the evolution of a parameter distribution, so the same fixed Fourier-mode picture is not available in the same direct form.

For this reason, the rest of the thesis uses the NTK formulation as the main analytical framework. The mean-field derivation and its comparison with the NTK scaling are given in Appendix A.

Chapter 3

Related Work

3.1 Convergence theory for overparameterized neural networks

A central question in the theory of neural-network training is why gradient-based optimization can reach small or even zero training loss despite the non-convexity of the parameter-space objective. A major line of work addresses this question in overparameterized regimes by proving that gradient descent or stochastic gradient descent can converge to global minimizers when the network is sufficiently wide (Du et al., 2019; Allen-Zhu et al., 2019; Arora et al., 2019). A closely related viewpoint shows that, in large-width regimes, networks can remain close to their linearization around initialization, so that the tangent kernel stays stable during training (Jacot et al., 2020; Lee et al., 2020). In this regime, the training dynamics can be analyzed as kernel dynamics rather than as a general non-convex optimization problem.

Infinite-width NTK dynamics. The Neural Tangent Kernel framework of Jacot et al. (2020) gives one of the cleanest such descriptions. In the infinite-width limit, the NTK converges to a deterministic limiting kernel and remains constant during training (Jacot et al., 2020; Lee et al., 2020). For the squared loss, this reduces the evolution of the network function to a linear differential equation in function space. Convergence is then controlled by the spectrum of the limiting kernel, and in particular by its positive-definiteness on the training data (Jacot et al., 2020). This provides a direct link between training dynamics, kernel eigenvalues, and convergence rates.

Finite-width global convergence in the overparameterized regime. A related line of work proves that sufficiently wide but finite networks can be trained to global minima by gradient descent or stochastic gradient descent. For example, Du et al. (2019) show that gradient descent achieves zero training loss for overparameterized deep networks by controlling the stability of the Gram matrix induced by the network. Similarly, Allen-Zhu et al. (2019) prove convergence of gradient-based optimization for sufficiently overparameterized deep networks, with width polynomial in the number of samples and depth under suitable

assumptions on the data. These results show that overparameterization can make the local geometry around random initialization regular enough for global convergence, even though the original objective is non-convex.

Optimization and generalization in the NTK regime. Other works refine this kernel-based view by studying both optimization and generalization. Arora et al. (2019) analyze overparameterized two-layer ReLU networks through a kernel perspective, showing that the network behaves similarly to kernel regression in the overparameterized regime. Such results help explain why training can converge rapidly in wide networks, but they also emphasize that the conclusions rely on regimes where the tangent features do not move too far from initialization (Arora et al., 2019; Lee et al., 2020).

Limits of the fixed-kernel picture. The NTK regime gives a tractable description of training, but it is also a restrictive approximation. In the infinite-width limit with fixed depth, the tangent kernel becomes deterministic and stays constant during training, so learning is described by a fixed feature map (Jacot et al., 2020; Lee et al., 2020). At finite width, this picture can fail in two ways: the tangent kernel is random at initialization, and it can also change during training.

Hanin and Nica (2019) show that finite depth and finite width can produce non-negligible fluctuations in the NTK at initialization. In particular, they find that the variance of the NTK depends on the ratio between depth and width, so the infinite-width deterministic-kernel picture is not automatically accurate when depth and width grow together. They also show that, in such regimes, the NTK can have non-trivial evolution during training. In other words, the tangent kernel itself can move during training, not merely fluctuate around its initial value. This suggests that finite networks need not behave exactly like their frozen infinite-width kernel counterparts.

Bordelon and Pehlevan (2023) study the dynamics of finite-width kernel and prediction fluctuations during training. Their analysis separates the lazy regime, where the kernel is random but essentially static, from feature learning regimes, where kernel fluctuations and prediction fluctuations evolve together. This is relevant here because the evolving tangent kernel can change not only the size of the learning rates, but also the directions in function space along which learning takes place.

Finally, Vyas et al. (2022) give empirical evidence that NTK models can fail to reproduce the generalization behavior of finite neural networks on realistic tasks. They also find that the empirical NTK can continue to evolve during much of training, rather than stabilizing after a short initial period. This reinforces the point that the fixed-kernel NTK should be treated as a useful reference model, not as a complete description of finite network training.

These works motivate the comparison made in this thesis. We use the continuum infinite-width NTK as the reference operator, but we also study the frozen finite-width tangent kernel at initialization and the evolving finite-width tangent kernel during training.

3.2 Spectral bias

A striking and widely reported phenomenon in neural network training is that gradient-based optimization does not fit all components of a target function at the same rate. Instead, networks tend to learn simple, slowly varying structure first and refine fine-scale or highly oscillatory structure later. This preference is commonly referred to as *spectral bias* or the *frequency principle*. In its classical formulation, spectral bias asserts that, when a target function is decomposed into Fourier modes, lower-frequency components are learned earlier and at a faster rate than higher-frequency components (Rahaman et al., 2019; Xu et al., 2019; Zhi-Qin et al., 2020).

In this section, we briefly review empirical evidence for spectral bias and summarize theoretical viewpoints that relate it to the spectrum of the operator governing training dynamics in the kernel (NTK) regime. We also highlight extensions that study how the phenomenon depends on the input distribution and how it manifests outside the training set.

3.2.1 Empirical evidence and the frequency principle

A standard way to empirically probe spectral bias is to choose a target with a controlled frequency decomposition and to track the frequency content of the network's prediction f_t during training.

Synthetic Fourier regression. A canonical experiment is 1D regression on a target constructed as a sum of sinusoids,

$$y(z) = \sum_{i=1}^{r} A_i \sin(2\pi k_i z + \phi_i), \qquad z \in [0,1],$$

and monitoring the discrete Fourier spectrum of the learned predictor as a function of training time. Rahaman et al. (2019) report that, for deep ReLU networks trained by (full-batch) gradient descent, Fourier coefficients at smaller frequencies k_i grow substantially earlier than those at larger k_i, even when the amplitudes are matched or when high-frequency components have larger amplitude (Rahaman et al., 2019). This "frequency-dependent learning speed" is one of the most direct empirical signatures of spectral bias.

Related work under the name *frequency principle* reports a similar ordering in the frequency domain: lower frequencies are fitted earlier during training, while higher frequencies are fitted later (Xu et al., 2019; Zhi-Qin et al., 2020). These observations support the view that spectral bias is a statement about the training dynamics, not just about which functions can be represented by the network.

Robustness and perturbation tests. A complementary protocol is to train to near-zero training error and then apply random perturbations in parameter space, comparing how the Fourier spectrum of the realized function changes. Rahaman et al. (2019) observe that lower-frequency components of the learned function are substantially more robust to such perturbations than higher-frequency components, suggesting that the network parameterization itself represents low frequencies in a more stable manner (Rahaman et al., 2019).

This matters because the perturbations are applied around the minimizer reached by training. If the low-frequency part changes little under such local perturbations, then small movements in parameter space do not strongly affect those components of the learned function, while higher-frequency components are more sensitive to the same perturbations. This provides further evidence that gradient-based training learns low-frequency components in a more stable way than high-frequency components, not only that it learns them earlier.

3.2.2 Theoretical explanations in the NTK regime

The NTK viewpoint connects convergence theory to spectral bias; see Section 2.1.1. In the fixed-kernel regime, the residual evolves under a kernel operator. When this operator is diagonalized into eigenfunctions, each residual component decays at a rate determined by the corresponding eigenvalue. Cao et al. (2020) make this connection explicit for the NTK regime by showing that components aligned with larger eigenvalues are learned faster. Spectral bias can therefore be viewed as a refined convergence statement, where some directions in function space decay faster than others.

This operator viewpoint becomes a frequency statement in settings with enough symmetry. Ronen et al. (2019), Bietti and Mairal (2019), and Cao et al. (2020) analyze settings in which one-hidden-layer ReLU networks and related NTK models are considered on the circle S^1 or sphere S^{d-1}, with isotropic initialization and uniform input distribution. In these settings, the limiting NTK is rotation-invariant, and the associated operator is diagonal in a harmonic basis.[1] On S^1, the eigenfunctions are Fourier modes. On the sphere, the corresponding eigenfunctions are spherical harmonics, which play the same role as Fourier modes for rotationally invariant operators. In these settings, lower-frequency components typically correspond to larger eigenvalues, giving a precise kernel-regime explanation of the low-frequency-first behavior.

The same operator viewpoint also explains why spectral bias is not only a statement about Fourier frequency. In kernel regimes, the preferred components are the leading eigenfunctions of the operator induced by the architecture, activation, and input distribution (Jacot et al., 2020; Cao et al., 2020; Bowman and Montufar, 2022). When the architecture, activation, or input distribution changes, the induced operator changes as well. The relevant eigenfunctions then need not be Fourier modes, and they may be difficult to characterize explicitly. The next subsection discusses this issue for non-uniform input distributions.

3.2.3 Uniform vs. non-uniform sampling

The question studied by Basri et al. (2020) is what happens when the input distribution is no longer uniform. If inputs are distributed according to a density $p(\theta)$ on S^1, then the relevant operator as described in equation (2.19) becomes

$$(T_p g)(\theta) := \int_0^{2\pi} \kappa(\theta - \theta')\, g(\theta')\, p(\theta')\, \frac{d\theta'}{2\pi}.$$

[1] A harmonic basis is an orthogonal decomposition of L^2 into eigenspaces of the Laplace–Beltrami operator. See: https://en.wikipedia.org/wiki/Spherical_harmonics

Even when κ depends only on $\theta - \theta'$, the factor $p(\theta')$ means that T_p is no longer a pure convolution operator. Its eigenfunctions therefore need not be Fourier modes. As a result, the components learned first are determined by the leading eigenfunctions of the density-weighted operator, rather than only by the lowest Fourier frequencies.

Basri et al. (2020) show that this can make learning spatially dependent. Non-uniform densities can produce eigenfunctions with localized oscillatory structure, and denser regions can exhibit faster learning of finer-scale structure.

For practical data distributions, this means that the spectral bias is shaped by both the model and the distribution of inputs. This is related to the manifold hypothesis, which says that high-dimensional data often lie near a lower-dimensional structured set (Fefferman et al., 2016). Because the integral operator seen earlier is defined with respect to the data distribution, its eigenfunctions are affected by the geometry of the set where the data concentrates. In this setting, "low frequency" need not mean low frequency in the ambient coordinates; it can instead mean slow variation along the data manifold. For instance, on image data, the ambient input space is the space of all possible pixel arrays, while natural images occupy only a small and highly structured part of that space. This is consistent with empirical work by Fridovich-Keil et al. (2022), who measure frequency-dependent behavior in modern image classifiers and find that the learned function's frequency content depends on the structure of the image data. Similar issues arise for audio and time-series data, where the possible input space contains all sequences, but realistic signals form a much smaller structured subset. For text, the raw input is a sequence of tokens, and these tokens are often mapped to high-dimensional embedding vectors; natural language only uses a highly structured subset of all possible token sequences or embedding patterns (Mikolov et al., 2013). This makes the corresponding NTK eigenfunctions harder to describe explicitly than in the uniform circle or sphere setting. The same operator principle still applies, since training favours the directions emphasized by the NTK operator under the data distribution (Bowman and Montufar, 2022; Basri et al., 2020).

3.2.4 Spectral bias beyond the training set

Most early demonstrations of spectral bias describe what happens on the training samples. For example, one can track Fourier coefficients computed from sampled predictions, or project the training residual onto eigenvectors of the empirical NTK matrix. These are useful diagnostics, but they only describe a finite vector of values on the training set.

Bowman and Montufar (2022) ask whether the same bias holds for the learned function itself. Here, "outside the training set" means that we evaluate the residual not only at the sampled points $x_1, \ldots, x_n$, but as a function over the full input domain. In their notation, this is measured in $L^2(\mathcal{X}, \rho)$, where $\mathcal{X}$ is the set of possible inputs and ρ is the input distribution.

Their result compares a finite-width network trained on finitely many samples with the idealized infinite-width, infinite-data kernel dynamics. In the idealized dynamics, the residual decomposes along eigenfunctions of the NTK integral operator, and each component decays at a rate determined by its eigenvalue. They show that the finite network stays close to this idealized trajectory, up to a stopping time. This implies that the network inherits the

spectral bias of the NTK integral operator as a function-space effect, not only as a pattern visible on the training samples (Bowman and Montufar, 2022).

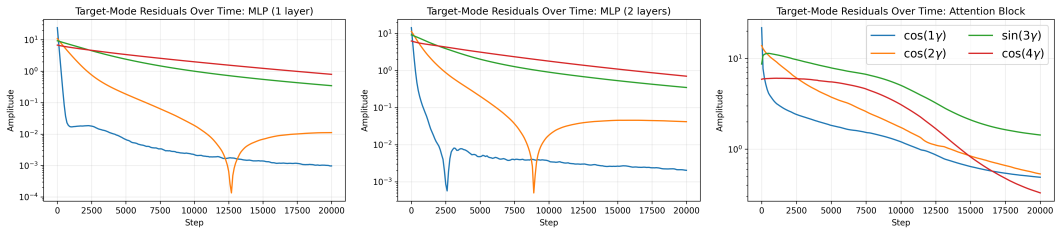
3.2.5 Discussion and connection to feature learning

The operator-spectrum viewpoint gives a clean account of spectral bias in kernelized regimes, where learning rates are dictated by the spectrum of a fixed or nearly fixed operator. At finite width, however, the tangent kernel is random at initialization and may also change during training. The relevant operator is then no longer exactly the continuum NTK operator, and its eigenvalues and eigenspaces need not match the Fourier reference.

Finite-width corrections to the NTK have been studied by Hanin and Nica (2019), while Bordelon and Pehlevan (2023) study finite-width kernel and prediction fluctuations during training. Vyas et al. (2022) further emphasize that the NTK alone may not fully explain generalization in practical deep networks. These works point to a limitation of the ideal fixed-kernel picture: it explains spectral bias cleanly when the kernel is fixed, but it does not by itself describe how the spectral structure changes when the kernel evolves.

3.3 Preliminary experiments across architectures

Before turning to the analysis, we ran preliminary experiments of our own to check whether the frequency-dependent ordering seen in the literature also appears in the architectures and finite-width, finite-data setting we care about here.



(a) Single-hidden-layer MLP. (b) Two-layer MLP. (c) Attention block.

Figure 3.1: Preliminary experiments across architectures trained at finite width on finite data. Each panel shows the residual amplitude of four target modes, $\cos\gamma$, $\cos 2\gamma$, $\sin 3\gamma$, and $\cos 4\gamma$, against training step on a logarithmic scale.

Figure 3.1 shows these experiments. In every case the lowest-frequency mode $\cos\gamma$ collapses first, within the first few hundred steps, while the higher-frequency modes decay more slowly, so the broad low-frequency-first ordering holds across all three architectures. The ordering is not strict: in the MLPs $\cos 2\gamma$ overtakes the higher modes and its coefficient passes through zero, and in the attention block the modes are more tangled and $\cos 4\gamma$ eventually decays past the others. Frequency-dependent learning is therefore not specific to the one-hidden-layer setting, but the precise spectral picture depends on the operator induced by the architecture, the finite sample, and the finite-width kernel. This is what motivates the

controlled analysis that follows, where the infinite-width reference operator is explicit and the effects of finite sampling and finite width can be separated.

3.4 Research gap

The preceding discussion shows that spectral bias has a clean explanation in idealized fixed-kernel settings. In the infinite-width NTK regime, training is governed by a fixed operator whose spectrum determines the decay rates of the corresponding components of the residual (Jacot et al., 2020; Cao et al., 2020). In symmetric settings, such as the circle or sphere with the uniform measure, this operator can be diagonalized by Fourier or harmonic modes, giving a precise version of the "low frequencies first" phenomenon (Ronen et al., 2019; Basri et al., 2020; Bietti and Mairal, 2019).

Bowman and Montufar (2022) prove bounds comparing the trajectory of a finite-width network trained on finitely many samples with the idealized NTK dynamics obtained at infinite width and infinite data. The comparison holds in L^2 over the input distribution and up to a finite time horizon, so it describes the learned function away from the training samples as well. Their result shows that the network can inherit the bias of the NTK integral operator toward its leading eigenfunctions. However, their analysis is general as it does not focus on a setting where the eigenfunctions can be written explicitly as Fourier modes, nor does it separately track how finite sampling and finite width perturb those Fourier eigenspaces.

This motivates studying a setting where the ideal operator is explicit enough that its perturbations can be inspected directly. The circle S^1 with the uniform measure provides such a setting. The limiting NTK is a convolution operator, the Fourier modes are exact eigendirections, and the corresponding eigenvalues can be computed explicitly. This lets us separate the effects that are otherwise mixed together: the effect of replacing the continuum measure by a finite sample, the effect of replacing the infinite-width kernel by a finite-width random kernel, and the effect of allowing the tangent kernel to evolve during training.

What is still missing is therefore a direct account of how the Fourier description on S^1 changes under these perturbations. One would expect the low-frequency Fourier structure to persist approximately under finite sampling and finite width. In other words, the perturbed eigenvalues should remain close to their continuum values, and the eigenspaces of the perturbed operator should remain close to the corresponding Fourier frequency subspaces. At the same time, these approximations should become more fragile when eigenvalues are small, when gaps between eigenvalues are narrow (e.g. in higher frequencies), or when the kernel evolves during training.

3.5 Research questions

This thesis studies spectral bias on S^1 by comparing the residual dynamics induced by the continuum infinite-width NTK operator with those induced by its finite-sample and finite-width counterparts.

Since the full operator acts on an infinite-dimensional function space, we make the comparisons on a finite Fourier subspace. For $K > 0$, let

$$\mathcal{H}_K := \text{span}\{1, \sqrt{2}\cos(k\theta), \sqrt{2}\sin(k\theta) : 1 \leq k \leq K\}.$$

This subspace contains the constant mode and the first K non-zero Fourier frequencies. Working on $\mathcal{H}_K$ gives a controlled way to test how far the ideal Fourier-mode explanation of spectral bias survives at finite sample size and finite width.

Main research question. To what extent does the spectral decomposition of the continuum infinite-width NTK operator on S^1 persist on a fixed low-frequency Fourier subspace $\mathcal{H}_K$ under finite-sample and finite-width perturbations?

Subquestions.

(RQ1) **Finite-sample perturbation of the continuum operator.** In the infinite-width regime, how does replacing the uniform measure on S^1 by n i.i.d. sampled points affect the action of the continuum NTK operator on the fixed truncated Fourier subspace $\mathcal{H}_K$? More precisely, how close is the sampled operator to the diagonal Fourier prediction on this subspace, and what are the consequences for the associated residual dynamics?

(RQ2) **Finite-width perturbation in the continuum setting.** With the continuum measure on S^1 kept fixed, how does finite network width affect the tangent-kernel dynamics on the fixed truncated Fourier subspace $\mathcal{H}_K$, both at initialization and during training? More precisely, how close is the finite-width tangent kernel at initialization to the infinite-width tangent kernel, and how does the evolution of this tangent kernel during training change the decay of Fourier modes, their alignment, and their effective learning rates?

3.6 Contributions

- We derive the infinite-width reference model for the specific network studied in this thesis. Closely related two-layer analyses often simplify the NTK by training only the first-layer weights and biases while keeping the output weights fixed (Ronen et al., 2019; Basri et al., 2020). Here, we keep the output weights trainable and include their contribution to the NTK. This gives an explicit kernel on S^1, explicit Fourier learning rates, and the baseline used in the rest of the thesis.

- We study how this Fourier-frequency picture changes when the continuum input distribution is replaced by finitely many sampled points. For a fixed low-frequency subspace, we prove non-asymptotic bounds showing that the sampled Fourier structure remains close to the idealized continuum prediction when the sample size is large enough.

- We support the finite-sample analysis with numerical experiments. These experiments compare the predicted frequency-wise learning rates with the rates observed on sampled data, and show that the agreement improves as the number of samples increases.
- We study the effect of finite network width when the tangent kernel is frozen at initialization. On the same low-frequency subspace, we prove that the finite-width operator concentrates around the infinite-width prediction as width increases, and we verify this numerically using eigenspace-alignment diagnostics.
- We empirically study the evolving finite-width tangent kernel during training. We compare frozen and evolving dynamics across widths, and track mode-wise residual decay, Fourier-subspace alignment, effective decay-rates, and kernel drift. The experiments suggest that kernel evolution helps at small width mainly by increasing low-frequency spectral strength, even though it can move some eigenspaces away from the Fourier reference. This advantage becomes smaller as width increases.

Together, these results show how the clean Fourier description from the infinite-width model changes when we move to finite data and finite network width. For finite sampling and frozen finite width, we give explicit non-asymptotic bounds. For the evolving finite-width kernel during training, we provide an empirical analysis showing how kernel evolution changes residual decay, Fourier alignment, and spectral strength. Developing comparable theory for this evolving-kernel regime remains the main open direction.

Chapter 4

Methodology

This chapter defines the operators, projections, and diagnostics used throughout the thesis, and states what each object measures and how it enters the finite-sample and finite-width experiments. Detailed derivations are given in the appendices.

4.1 Mathematical setting

4.1.1 Domain and measure

We identify the unit circle S^1 with the angular domain $\Omega = [0, 2\pi)$, equipped with the uniform probability measure

$$d\mu(\varphi) = \frac{d\varphi}{2\pi}. \tag{4.1}$$

The embedding into $\mathbb{R}^2$ is given by

$$x(\varphi) = (\cos\varphi, \sin\varphi) \in S^1 \subset \mathbb{R}^2. \tag{4.2}$$

This parametrization makes the rotational symmetry of the problem explicit and provides the natural Fourier basis for $L^2(\mu)$.

4.1.2 Network model

We consider a one-hidden-layer ReLU network in NTK parameterization. For inputs $x \in S^1 \subset \mathbb{R}^2$,

$$f_m(x;\theta) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \sigma(w_i^\top x + b_i), \tag{4.3}$$

where $\theta = \{(a_i, w_i, b_i)\}_{i=1}^m$, $\sigma(t) = t_+ = \max\{t, 0\}$, and the parameters are initialized according to

$$a_i \sim \mathcal{N}(0,1), \qquad w_i \sim \mathcal{N}(0, I_2), \qquad b_i(0) = 0, \tag{4.4}$$

independently across $i = 1, \ldots, m$. The bias parameters are initialized at zero but remain trainable throughout.

4.1.3 Least-squares loss and residual dynamics

Let $y \in L^2(\mu)$ be the target function, and write

$$f_t(\varphi) := f_m(x(\varphi); \theta(t))$$

for the network output at training time t. We consider the least-squares loss

$$\mathcal{L}(\theta) = \frac{1}{2}\|f_m(\cdot;\theta) - y\|^2_{L^2(\mu)} = \frac{1}{2}\int_0^{2\pi} \big(f_m(\varphi;\theta) - y(\varphi)\big)^2 \, d\mu(\varphi). \tag{4.5}$$

We define the residual

$$r_t(\varphi) = f_t(\varphi) - y(\varphi). \tag{4.6}$$

At finite width, the empirical neural tangent kernel at time t is

$$\Theta_t^{(m)}(\varphi,\psi) = \big\langle \nabla_\theta f_m(\varphi;\theta(t)), \nabla_\theta f_m(\psi;\theta(t)) \big\rangle, \tag{4.7}$$

and the corresponding integral operator $T_t^{(m)} : L^2(\mu) \to L^2(\mu)$ is

$$(T_t^{(m)} f)(\varphi) = \int_0^{2\pi} \Theta_t^{(m)}(\varphi,\psi) f(\psi) \, d\mu(\psi). \tag{4.8}$$

Under gradient flow, the residual evolves according to

$$\partial_t r_t = -T_t^{(m)} r_t. \tag{4.9}$$

Consequently, the loss evolves as

$$\frac{d}{dt}\frac{1}{2}\|r_t\|^2_{L^2(\mu)} = -\langle r_t, T_t^{(m)} r_t \rangle_{L^2(\mu)}. \tag{4.10}$$

Loss decay therefore depends not only on the eigenspaces of $T_t^{(m)}$, but also on how strongly the operator acts on the directions where the residual has mass.

This equation contains both the frozen and evolving regimes. In the frozen regime, $T_t^{(m)}$ is replaced by its initial value

$$T_0^{(m)} := T_t^{(m)}\big|_{t=0}, \tag{4.11}$$

whereas in the evolving regime the full time dependence of $T_t^{(m)}$ is retained.

4.2 Continuum reference model

4.2.1 Infinite-width tangent operator

The first object studied in the thesis is the infinite-width limiting tangent operator

$$T_\infty := \lim_{m\to\infty} T_0^{(m)}, \tag{4.12}$$

whose kernel is denoted by Θ. Under the initialization (4.4), the limiting kernel exists and is deterministic. The explicit formula for Θ and the corresponding Fourier eigenvalues are stated in Chapter 5 and derived in Appendix B.

4.2.2 Translation invariance and Fourier diagonalization

Because the input distribution is uniform and the limiting kernel is rotation-invariant, $\Theta(\varphi,\psi)$ depends only on the angular difference. The limiting operator therefore takes the convolution form

$$(T_\infty f)(\varphi) = \int_0^{2\pi} \Theta(\varphi-\psi) f(\psi)\, d\mu(\psi), \tag{4.13}$$

and is diagonalized by the real Fourier basis of $L^2(\mu)$. Its eigenvalues determine the decay rates of the Fourier components of the residual in the ideal fixed-kernel continuum model.

This continuum Fourier decomposition serves as the reference point for the finite-sample and finite-width analyses below.

4.2.3 Real Fourier basis and truncated subspaces

We define the real Fourier basis of $L^2(\mu)$ by

$$\phi_0(\varphi) = 1, \tag{4.14}$$

and, for each $k \geq 1$,

$$\phi_{k,c}(\varphi) = \sqrt{2}\cos(k\varphi), \qquad \phi_{k,s}(\varphi) = \sqrt{2}\sin(k\varphi). \tag{4.15}$$

For a fixed cutoff $K \geq 1$, we consider the truncated subspace

$$\mathcal{H}_K = \operatorname{span}\{\phi_0, \phi_{k,c}, \phi_{k,s} : 1 \leq k \leq K\}, \tag{4.16}$$

with dimension

$$d = 2K+1. \tag{4.17}$$

Let

$$P_K : L^2(\mu) \to \mathcal{H}_K \tag{4.18}$$

denote the orthogonal projection onto $\mathcal{H}_K$.

The truncation to $\mathcal{H}_K$ is used throughout the thesis to reduce the continuum problem to a finite-dimensional low-frequency block that can be compared across the continuum, sampled, frozen finite-width, and evolving finite-width settings.

4.3 Finite-sample methodology

4.3.1 Finite-sample question

We first study the effect of replacing the measure μ by finitely many sampled points. Let

$$\varphi_1, \ldots, \varphi_n \overset{\text{iid}}{\sim} \text{Unif}[0, 2\pi). \tag{4.19}$$

The main question is whether the diagonal Fourier decomposition of the continuum operator remains approximately valid after this replacement.

4.3.2 Sampled objects

To study the sampled operator on $\mathcal{H}_K$, we introduce the feature matrix $\Phi \in \mathbb{R}^{n\times d}$,

$$\Phi_{i,p} = \phi_p(\varphi_i), \qquad 1 \le i \le n, \quad 1 \le p \le d, \tag{4.20}$$

and the normalized sampled kernel matrix A with

$$A_{ij} = \frac{1}{n}\Theta(\varphi_i - \varphi_j). \tag{4.21}$$

From these we form the Gram matrix

$$G = \frac{1}{n}\Phi^\top \Phi, \tag{4.22}$$

and the compressed operator matrix

$$H = \frac{1}{n}\Phi^\top A \Phi. \tag{4.23}$$

We also introduce the diagonal matrix

$$\Lambda^{(n)} = \operatorname{diag}(\lambda_p^{(n)}), \tag{4.24}$$

whose entries are the finite-sample corrected Fourier eigenvalues defined in Chapter 5.

4.3.3 Quantities to be controlled

The finite-sample analysis controls how far three sampled quantities deviate from the values they would take if the sampled Fourier structure were exact. Each is written as the difference between the sampled quantity and its ideal value:

$$\underbrace{G - I_d}_{\text{basis vs. orthonormal}}, \qquad \underbrace{H - \Lambda^{(n)}}_{\text{operator vs. diagonal}}, \qquad \underbrace{A\Phi - \Phi\Lambda^{(n)}}_{\text{modes vs. eigenvectors}}. \tag{4.25}$$

The first is small when the sampled Fourier basis is close to orthonormal, so its Gram matrix G is close to the identity. The second is small when the sampled operator restricted to $\mathcal{H}_K$, represented by H, is close to the diagonal finite-sample prediction $\Lambda^{(n)}$. The third is small when the sampled Fourier modes are close to being eigenvectors of the sampled operator, i.e. applying the operator A to the modes Φ reproduces them scaled by the predicted eigenvalues. Together these are the quantities needed to answer Q1.

4.4 Fourier-block preconditioning methodology

The finite-sample analysis gives a diagonal reference for the sampled operator on a fixed Fourier block. We use it to define a preconditioning experiment that tests whether the predicted Fourier eigenvalues explain the observed rate separation between retained Fourier modes.

Fix a retained Fourier block

$$\mathcal{H}_K = \operatorname{span}\{\phi_0, \phi_{1,c}, \phi_{1,s}, \dots, \phi_{K,c}, \phi_{K,s}\},$$

with dimension $d = 2K+1$. Let

$$\Phi \in \mathbb{R}^{n\times d}$$

be the sampled Fourier feature matrix, and let $A \in \mathbb{R}^{n\times n}$ be the normalized sampled kernel operator. Recall that

$$G = \frac{1}{n}\Phi^\top \Phi, \qquad H = \frac{1}{n}\Phi^\top A \Phi. \tag{4.26}$$

The matrix G is the sampled Fourier Gram matrix, while H records the action of the sampled operator A on the retained Fourier basis. When G is invertible, define

$$C_n^{(K)} := G^{-1}H. \tag{4.27}$$

This matrix describes what the sampled operator A does to the Fourier coefficients on the retained block. Indeed, if $u = \Phi z$, then z is the coefficient vector of u in the sampled Fourier basis. After applying A, the least-squares coefficients of Au in the same sampled Fourier basis are

$$G^{-1}\frac{1}{n}\Phi^\top (Au) = G^{-1}\frac{1}{n}\Phi^\top A\Phi z = C_n^{(K)} z. \tag{4.28}$$

Thus $C_n^{(K)}$ is the object that should be compared with the diagonal finite-sample prediction

$$\Lambda^{(n)} = \operatorname{diag}(\lambda_p^{(n)}). \tag{4.29}$$

The finite-sample theory predicts that, on the retained block,

$$C_n^{(K)} \approx \Lambda^{(n)}. \tag{4.30}$$

This means that, on the retained Fourier block, the sampled operator is expected to act mainly by rescaling each Fourier coefficient by its predicted eigenvalue, with only small mixing between different retained Fourier modes.

We now define two preconditioners from this comparison. First note that, for any sampled vector $v \in \mathbb{R}^n$,

$$G^{-1}\frac{1}{n}\Phi^\top v$$

gives the least-squares Fourier coefficients of the projection of v onto the sampled span of the retained modes. Thus, to correct the action of A on the retained block, we first apply A, then project back onto the retained Fourier span, and finally rescale the resulting Fourier coefficients.

The theory preconditioner uses the diagonal prediction $\Lambda^{(n)}$:

$$P_{\text{th}} := \frac{1}{n}\Phi\big(\Lambda^{(n)}\big)^{-1} G^{-1}\Phi^\top. \tag{4.31}$$

This operator projects a sampled vector onto the retained Fourier span and rescales each retained Fourier coefficient by the inverse predicted eigenvalue. It uses the analytic prediction $\Lambda^{(n)}$, and therefore does not use the observed matrix $C_n^{(K)}$.

The empirical preconditioner uses the observed retained block:

$$P_{\text{emp}} := \frac{1}{n}\Phi\big(C_n^{(K)}\big)^{-1}G^{-1}\Phi^\top. \tag{4.32}$$

This operator uses the actual sampled matrix $C_n^{(K)}$. It therefore corrects both the diagonal scaling and the mixing between retained Fourier modes that is present for the particular sampled points.

We compare three sampled dynamics:

$$B_{\text{base}} := A, \qquad B_{\text{th}} := P_{\text{th}}A, \qquad B_{\text{emp}} := P_{\text{emp}}A. \tag{4.33}$$

The baseline B_{base} is the original sampled kernel operator. The operators B_{th} and B_{emp} apply the sampled kernel operator and then correct its action through the retained Fourier block.

For sampled predictions $f_t \in \mathbb{R}^n$, targets $y \in \mathbb{R}^n$, and residuals

$$r_t = f_t - y,$$

we compare the gradient-flow dynamics

$$\frac{dr_t}{dt} = -Br_t, \tag{4.34}$$

with

$$B \in \{B_{\text{base}}, B_{\text{th}}, B_{\text{emp}}\}.$$

The effect is easiest to see on a vector $u = \Phi z$ in the retained Fourier span. Applying A and projecting back to the retained block gives the coefficient vector $C_n^{(K)}z$. Therefore

$$B_{\text{th}}\Phi z = \Phi\big(\Lambda^{(n)}\big)^{-1}C_n^{(K)}z, \tag{4.35}$$

while

$$B_{\text{emp}}\Phi z = \Phi\big(C_n^{(K)}\big)^{-1}C_n^{(K)}z = \Phi z. \tag{4.36}$$

Thus on the retained block the empirical preconditioner reduces the action to the identity: every retained Fourier mode is driven at the same rate, so the frequency-dependent rate separation is removed, provided $C_n^{(K)}$ is invertible. The theory preconditioner achieves this only insofar as the observed block $C_n^{(K)}$ is close to the diagonal prediction $\Lambda^{(n)}$; from (4.35), it leaves the residual factor $(\Lambda^{(n)})^{-1}C_n^{(K)}$, which equals the identity exactly when $C_n^{(K)} = \Lambda^{(n)}$.

In the numerical experiments, ill-conditioned inverses are replaced by small ridge-regularized inverses when needed, for example $(\Lambda^{(n)} + \tau I)^{-1}$ or $(C_n^{(K)} + \tau I)^{-1}$ with $\tau > 0$. If B_{th} and B_{emp} behave similarly, the predicted Fourier eigenvalues explain most of the rate differences on the retained block; if B_{emp} is noticeably better, the sampled points introduce mixing between retained Fourier modes that the diagonal prediction does not capture.

4.5 Finite-width frozen-kernel methodology

4.5.1 Finite-width question

We next remove sampling randomness and return to the continuum input space S^1, but keep the network width finite. The question is then whether the low-frequency block of the frozen finite-width tangent operator remains close to the continuum Fourier block predicted by the infinite-width theory.

4.5.2 One-neuron tangent features and operators

For each neuron r, define its tangent feature at initialization by

$$\psi_r(\varphi) = \nabla_{(a_r,w_r,b_r)}\big(a_r\sigma(w_r^\top x(\varphi)+b_r)\big). \tag{4.37}$$

The corresponding one-neuron kernel contribution is

$$\Theta^{(r)}(\varphi,\psi) = \psi_r(\varphi)^\top\psi_r(\psi), \tag{4.38}$$

and the associated one-neuron operator is

$$(T^{(r)}f)(\varphi) = \int_0^{2\pi}\Theta^{(r)}(\varphi,\psi)f(\psi)\,d\mu(\psi). \tag{4.39}$$

At initialization, the frozen finite-width tangent operator is the average

$$T_0^{(m)} = \frac{1}{m}\sum_{r=1}^{m}T^{(r)}. \tag{4.40}$$

4.5.3 Projected frozen operator

Let $P_K : L^2(\mu)\to\mathcal{H}_K$ denote the orthogonal projection onto the truncated Fourier subspace $\mathcal{H}_K$. The low-frequency block of the frozen finite-width operator is

$$B_m^{(K)} := P_K T_0^{(m)} P_K. \tag{4.41}$$

The corresponding continuum reference block is

$$\Lambda_K := P_K T_\infty P_K = \operatorname{diag}(\lambda_0,\lambda_1,\lambda_1,\ldots,\lambda_K,\lambda_K). \tag{4.42}$$

The finite-width frozen analysis asks whether $B_m^{(K)}$ remains close to Λ_K with high probability as m increases.

4.5.4 Matrix representation of the frozen block

Using the ordered basis

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

the projected operator $B_m^{(K)}$ is represented by the matrix with entries

$$(B_m^{(K)})_{pq} = \langle \phi_p, T_0^{(m)} \phi_q \rangle_{L^2(\mu)}. \tag{4.43}$$

Equivalently,

$$(B_m^{(K)})_{pq} = \frac{1}{m} \sum_{r=1}^{m} \xi_{r,pq}, \qquad \xi_{r,pq} := \langle \phi_p, T^{(r)} \phi_q \rangle, \tag{4.44}$$

so that $B_m^{(K)}$ is an empirical average of m i.i.d. one-neuron matrices. Because the entries are averages of independent terms, they concentrate around their mean as m grows, which is what the finite-width frozen concentration result in the next chapter exploits.

4.5.5 Fourier-plane alignment diagnostics

The continuum operator is diagonal in the Fourier basis. For each frequency $k \geq 1$, the cosine and sine modes have the same continuum eigenvalue λ_k. Because $\phi_{k,c}$ and $\phi_{k,s}$ share the eigenvalue λ_k, the continuum operator does not single out either one; only the plane they span is determined. The object to compare against is therefore this two-dimensional frequency subspace,

$$\mathcal{F}_k := \operatorname{span}\{\phi_{k,c}, \phi_{k,s}\}, \qquad k \geq 1,$$

For the constant mode, we set

$$\mathcal{F}_0 := \operatorname{span}\{\phi_0\}.$$

Thus, the dimensionality $d_0 = 1$ and $d_k = 2$ for $k \geq 1$.

We use subspace-alignment diagnostics to compare these reference Fourier subspaces with matched eigenspaces of the finite-width tangent operator. In the frozen case this operator is $T_0^{(m)}$. In the evolving case it is the time-dependent operator $T_t^{(m)}$.

The operator is defined on the continuum S^1, but it cannot be eigen-decomposed in closed form at finite width, so for the numerical diagnostics we discretise it onto a uniform grid of N points,

$$\theta_j = \frac{2\pi j}{N}, \qquad j = 0, \ldots, N-1,$$

and carry out the eigendecomposition and subspace comparisons on this grid.

All vectors and matrices in this diagnostic are therefore finite-dimensional representations of continuum functions on this grid.

Let $U_k \in \mathbb{R}^{N \times d_k}$ be an orthonormal basis for the grid representation of $\mathcal{F}_k$, and define the corresponding orthogonal projector

$$P_k := U_k U_k^\top. \tag{4.45}$$

At each training time t, we eigendecompose the grid representation of the tangent operator, an $N \times N$ symmetric matrix, and write its orthonormal eigenvectors as $v_1(t), \ldots, v_N(t)$. For each eigenvector $v_j(t)$, the quantity

$$\|U_k^\top v_j(t)\|_2^2 \in [0,1]$$

is the squared length of its projection onto $\mathcal{F}_k$: it is 1 when $v_j(t)$ lies entirely in $\mathcal{F}_k$ and 0 when it is orthogonal to it. We build the matched empirical subspace $\widehat{\mathcal{F}}_k(t)$ by taking the d_k eigenvectors with the largest such values, the ones lying closest to $\mathcal{F}_k$. If $\widehat{U}_k(t) \in \mathbb{R}^{N\times d_k}$ is an orthonormal basis of $\widehat{\mathcal{F}}_k(t)$, define

$$\widehat{P}_k(t) := \widehat{U}_k(t)\widehat{U}_k(t)^\top. \tag{4.46}$$

The projection error is

$$\varepsilon_{k,F}^{(m)}(t) := \|\widehat{P}_k(t) - P_k\|_F. \tag{4.47}$$

The projectors P_k and $\widehat{P}_k(t)$ depend only on the subspaces, not on the bases chosen for them, so this distance is invariant to sign changes of eigenvectors and to rotations of the cosine-sine basis inside a two-dimensional Fourier plane, neither of which should count as a mismatch. It therefore measures the geometric distance between the matched eigenspace and the reference Fourier subspace, and (4.50) below expresses it through the principal angles. For the frozen operator, we write

$$\varepsilon_{k,F}^{(m)} := \varepsilon_{k,F}^{(m)}(0).$$

We also report principal angles between $\widehat{\mathcal{F}}_k(t)$ and $\mathcal{F}_k$. Principal angles are a standard way to compare finite-dimensional subspaces (Björck and Golub, 1973). If U_k and $\widehat{U}_k(t)$ are orthonormal bases for the two subspaces, then the cosines of the principal angles are the singular values of $U_k^\top \widehat{U}_k(t)$. We define

$$0 \le \theta_{k,1}(t) \le \cdots \le \theta_{k,d_k}(t) \le \frac{\pi}{2}$$

by

$$\cos\theta_{k,i}(t) = s_i(U_k^\top \widehat{U}_k(t)), \qquad i = 1,\dots,d_k, \tag{4.48}$$

where $s_i(\cdot)$ is the i-th largest singular value, so the cosines of the principal angles are the singular values of $U_k^\top \widehat{U}_k(t)$ ordered from largest to smallest.

For $k = 0$, there is one principal angle because both spaces are one-dimensional. For $k \ge 1$, there are two principal angles because both spaces are two-dimensional. The first angle measures the best aligned direction between the two planes. The second angle measures the remaining independent direction. Both angles must be small for the full Fourier plane to be well aligned with the matched empirical eigenspace.

For the evolving-kernel experiments, we also summarize the principal angles by the subspace cosine similarity

$$s_k^{\cos}(t) := \left(\frac{1}{d_k}\sum_{i=1}^{d_k}\cos^2\theta_{k,i}(t)\right)^{1/2}. \tag{4.49}$$

This quantity is the root-mean-square of the cosines of the principal angles. We use this aggregation to summarize one-dimensional and two-dimensional Fourier subspaces with a single score per frequency. The score lies between 0 and 1, with larger values indicating

better alignment. It equals 1 exactly when the matched empirical subspace is the same as the reference Fourier subspace.

The projection distance and the principal angles are related by the standard projector identity (Björck and Golub, 1973)

$$\|\widehat{P}_k(t) - P_k\|_F^2 = 2\sum_{i=1}^{d_k} \sin^2 \theta_{k,i}(t). \tag{4.50}$$

Thus the projection error gives one scalar measure of subspace mismatch, while the principal angles and subspace cosine similarity give more detailed views of the same alignment.

4.6 Finite-width evolving-kernel methodology

4.6.1 Projected evolving operator

To study the actual training dynamics, we retain the time dependence of the empirical tangent operator $T_t^{(m)}$. Its block on the same retained subspace $\mathcal{H}_K$, the frequencies up to K used throughout, is

$$C_t^{(K)} := P_K T_t^{(m)} P_K. \tag{4.51}$$

In the ordered Fourier basis of $\mathcal{H}_K$, this block has entries

$$(C_t^{(K)})_{pq} = \langle \phi_p, T_t^{(m)} \phi_q \rangle_{L^2(\mu)}. \tag{4.52}$$

At initialization,

$$C_0^{(K)} = B_m^{(K)}.$$

If $b_t = (I - P_K) r_t$ is the residual component outside the retained Fourier block, then $T_t^{(m)} b_t$ need not remain outside that block. Its projection back into $\mathcal{H}_K$ is

$$L_t^{(K)} b_t = P_K T_t^{(m)} (I - P_K) r_t. \tag{4.53}$$

We therefore define

$$L_t^{(K)} := P_K T_t^{(m)} (I - P_K). \tag{4.54}$$

This operator represents high-to-low frequency coupling. It measures how residual components outside the retained block can influence the evolution of the retained Fourier coefficients after the tangent operator is applied.

4.6.2 Projected residual dynamics

Let

$$a_t := P_K r_t, \qquad b_t := (I - P_K) r_t. \tag{4.55}$$

Projecting the exact residual equation (4.9) gives

$$\partial_t a_t = -C_t^{(K)} a_t - L_t^{(K)} b_t. \tag{4.56}$$

Thus the low-frequency dynamics are driven by two terms: the operator restricted to $\mathcal{H}_K$, and the contribution from residual components outside $\mathcal{H}_K$.

4.6.3 Frozen versus evolving comparison

The frozen and evolving regimes differ only through the time dependence of the operator. In the frozen regime,

$$C_t^{(K)} \equiv C_0^{(K)} = B_m^{(K)},$$

whereas in the evolving regime the operator changes during training. This motivates the decomposition

$$C_t^{(K)} = \Lambda_K + \big(C_0^{(K)} - \Lambda_K\big) + \big(C_t^{(K)} - C_0^{(K)}\big), \tag{4.57}$$

which separates three effects:

1. the ideal infinite-width Fourier block Λ_K,
2. the finite-width initialization error $C_0^{(K)} - \Lambda_K$,
3. the time-dependent kernel drift $C_t^{(K)} - C_0^{(K)}$.

The frozen-kernel theory studies the second term. The evolving-kernel experiments study the third term empirically: how much the tangent operator changes during training, and how these changes affect the Fourier-block structure.

4.6.4 Frequency-wise spectral strength

The subspace-alignment diagnostics measure whether the eigenspaces of the finite-width tangent operator remain close to the Fourier subspaces. They do not, however, measure how strongly the operator acts on those subspaces. For the residual dynamics, this distinction matters because the loss decay is controlled by $\langle r_t, T_t^{(m)} r_t \rangle_{L^2(\mu)}$, as shown in (4.10). Thus the loss decay depends not only on the orientation of the eigenspaces, but also on the magnitude of the operator in directions where the residual has mass. For each Fourier frequency subspace $\mathcal{F}_k$, with orthogonal projector P_k and dimension d_k, we define the average spectral strength

$$\mu_{\mathcal{F}_k}(t) := \frac{1}{d_k} \operatorname{tr}\Big(P_k T_t^{(m)} P_k\Big). \tag{4.58}$$

Equivalently, if U_k is an orthonormal basis for $\mathcal{F}_k$, then

$$\mu_{\mathcal{F}_k}(t) = \frac{1}{d_k} \operatorname{tr}\Big(U_k^\top T_t^{(m)} U_k\Big). \tag{4.59}$$

This quantity is basis-invariant within $\mathcal{F}_k$. If $\mathcal{F}_k$ is an invariant eigenspace of the operator, then $\mu_{\mathcal{F}_k}(t)$ equals the corresponding eigenvalue. In general, it should instead be interpreted as the average kernel strength inside that Fourier block.

For $k \le K$, the same quantity can be read directly from the low-frequency block $C_t^{(K)}$. It is the average trace of the diagonal block corresponding to $\mathcal{F}_k$:

$$\mu_{\mathcal{F}_k}(t) = \frac{1}{d_k} \operatorname{tr}\Big((C_t^{(K)})_{\mathcal{F}_k,\mathcal{F}_k}\Big). \tag{4.60}$$

We also use the cumulative low-frequency strength

$$\mu_{\leq K}(t) := \frac{1}{2K+1} \operatorname{tr}\Big(P_K T_t^{(m)} P_K\Big) = \frac{1}{2K+1} \sum_{k=0}^{K} d_k \mu_{\mathcal{F}_k}(t). \tag{4.61}$$

This summarizes the average tangent-kernel strength over the retained low-frequency block.

4.7 Summary of the methodology

The methodology consists of three comparisons.

First, the continuum infinite-width operator T_∞ gives the ideal Fourier-diagonal reference model on S^1. In this model, the Fourier modes are exact eigenfunctions and the eigenvalues determine the residual decay rates.

Second, the finite-sample analysis studies what changes when the continuum measure is replaced by finitely many sampled points. The main objects are the sampled Fourier Gram matrix G, the matrix H describing the action of the sampled operator on the retained Fourier basis, and the difference $A\Phi - \Phi\Lambda^{(n)}$ between the sampled operator action and the diagonal Fourier prediction. These objects are used to answer Q1. The rate separation between modes comes from the spread of eigenvalues, so rescaling each mode by its predicted eigenvalue should cancel it; the preconditioning experiment uses this to test how much of the separation the eigenvalues explain.

Third, the finite-width analysis studies what changes when the infinite-width operator is replaced by the tangent operator of a finite-width network. The frozen analysis focuses on $B_m^{(K)} = P_K T_0^{(m)} P_K$, the low-frequency block of the tangent operator at initialization. The evolving analysis tracks the time-dependent block $C_t^{(K)} = P_K T_t^{(m)} P_K$ and the high-to-low coupling $L_t^{(K)} = P_K T_t^{(m)} (I - P_K)$.

For the evolving experiments, we use two complementary kinds of diagnostics. Projection error, principal angles, and subspace cosine similarity measure whether the eigenspaces of the tangent operator remain close to the Fourier subspaces. Frequency-wise spectral strength measures how strongly the tangent operator acts on each Fourier block. Together, these diagnostics separate changes in Fourier geometry from changes in the magnitude of the kernel on the retained frequency blocks.

Chapter 5

Results

5.1 Continuum kernel on S^1

We first state the explicit form of the limiting neural tangent kernel on the circle and the corresponding Fourier eigenvalues of the integral operator T introduced in Chapter 4. The calculation uses the standard infinite-width NTK description of Jacot et al. (2020), together with the arc-cosine formulas for ReLU Gaussian averages (Cho and Saul, 2009). Related spectral calculations for ReLU NTKs on the sphere appear in Ronen et al. (2019); Bietti and Mairal (2019); Cao et al. (2020). The derivation in the assumptions used here is given in Appendix B.

Proposition 5.1 (Limiting NTK on S^1)**.** *Consider the finite-width network f_m and initialization from* (4.3)–(4.4)*. For $x(\varphi) = (\cos\varphi, \sin\varphi)$, define the finite-width NTK at initialization by*

$$\Theta_m(\varphi,\psi) = \langle \nabla_\theta f_m(x(\varphi);\theta_0), \nabla_\theta f_m(x(\psi);\theta_0)\rangle\,.$$

The limiting NTK is the deterministic kernel

$$\Theta(\varphi,\psi) = \lim_{m\to\infty} \Theta_m(\varphi,\psi),$$

where the limit is taken pointwise over the random initialization.

If $\delta = d(\varphi,\psi) \in [0,\pi]$ denotes the angular distance between $x(\varphi)$ and $x(\psi)$, then $\Theta(\varphi,\psi)$ depends only on δ and is given by

$$\Theta(\delta) = \frac{1}{2\pi}\Big(\sin\delta + 2(\pi-\delta)\cos\delta + (\pi-\delta)\Big). \tag{5.1}$$

Extending $\Theta(\delta)$ to an even 2π-periodic function, the associated integral operator is

$$(Tf)(\varphi) = \int_0^{2\pi} \Theta(\varphi-\psi) f(\psi)\, d\mu(\psi). \tag{5.2}$$

Since T is a convolution operator, it is diagonalized by the real Fourier basis introduced in (4.14)–(4.15).

Theorem 5.2 (Fourier eigenvalues of the limiting NTK)**.** *Let λ_k denote the eigenvalue of T corresponding to the Fourier frequency k introduced in* (4.14)–(4.15)*. Then*

$$\lambda_0 = \frac{1}{4} + \frac{3}{\pi^2}, \qquad \lambda_1 = \frac{1}{4} + \frac{1}{\pi^2}, \tag{5.3}$$

and for $k \geq 2$,

$$\lambda_k = \begin{cases} \dfrac{k^2+3}{\pi^2(k^2-1)^2}, & k \geq 2 \text{ even}, \\ \dfrac{1}{\pi^2 k^2}, & k \geq 3 \text{ odd}. \end{cases} \tag{5.4}$$

In particular, the decay rates of the Fourier components of the residual are determined by these eigenvalues.

The detailed derivation is provided in Appendix B.4.

The eigenvalues decrease with frequency. Both even and odd modes satisfy $\lambda_k = O(k^{-2})$ for large k, so higher-frequency residual components decay more slowly under fixed-kernel gradient flow. This gives the continuum spectral-bias baseline used in the rest of the chapter.

Corollary 5.3 (Spectrum without the trainable bias)**.** *Let $\Theta^{(\mathrm{nb})}$ denote the limiting NTK obtained by removing the trainable hidden-bias contribution, while keeping the output-weight and input-weight contributions. Then, on S^1,*

$$\Theta^{(\mathrm{nb})}(\delta) = \frac{1}{2\pi}\Big(\sin\delta + 2(\pi-\delta)\cos\delta\Big). \tag{5.5}$$

Let $\lambda_k^{(\mathrm{nb})}$ denote the corresponding Fourier eigenvalue. Then

$$\lambda_0^{(\mathrm{nb})} = \frac{3}{\pi^2}, \qquad \lambda_1^{(\mathrm{nb})} = \frac{1}{4}, \tag{5.6}$$

and, for $k \geq 2$,

$$\lambda_k^{(\mathrm{nb})} = \begin{cases} \dfrac{k^2+3}{\pi^2(k^2-1)^2}, & k \geq 2 \text{ even}, \\ 0, & k \geq 3 \text{ odd}. \end{cases} \tag{5.7}$$

Consequently, comparing with Theorem 5.2, the trainable bias leaves the even frequencies $k \geq 2$ unchanged, increases the constant and $k = 1$ eigenvalues, and supplies the nonzero eigenvalues of all odd modes $k \geq 3$. Thus the odd modes $k \geq 3$ are absent from the no-bias fixed-kernel dynamics but present in the model studied here.

The detailed derivation is provided in Appendix B.5.

The trainable hidden bias affects the nullspace of the continuum operator. Corollary 5.3 shows that, without the bias contribution, the odd modes $k \geq 3$ have zero eigenvalue. With the bias contribution, these modes acquire nonzero but small eigenvalues. Thus, the bias determines whether these modes can be learned in the fixed-kernel continuum dynamics. This is consistent with earlier harmonic analyses of two-layer ReLU networks on the sphere and circle, where bias-free two-layer models do not express odd harmonics beyond the first frequency (Ronen et al., 2019; Basri et al., 2020).

5.1.1 Continuum fixed-kernel dynamics

The eigenvalues in Theorem 5.2 determine the ideal fixed-kernel dynamics. Write the residual from Section 4.1.3 in the real Fourier basis as

$$r_t = \sum_p \alpha_p(t)\phi_p.$$

Here, α_p is the amplitude associated with mode p, λ_p denotes the eigenvalue associated with the Fourier frequency of ϕ_p and $\Lambda = \mathrm{diag}(\lambda_p)$. Since $T\phi_p = \lambda_p\phi_p$, gradient flow gives

$$\alpha_p(t) = \exp(-\lambda_p t)\alpha_p(0). \tag{5.8}$$

Thus the normalized amplitude $|\alpha_p(t)|/|\alpha_p(0)|$ is set by the eigenvalue of the corresponding Fourier mode.

Figure 5.1 shows this diagonal continuum dynamics for the kernel with and without the trainable-bias contribution. Low-frequency components decay faster than high-frequency components.

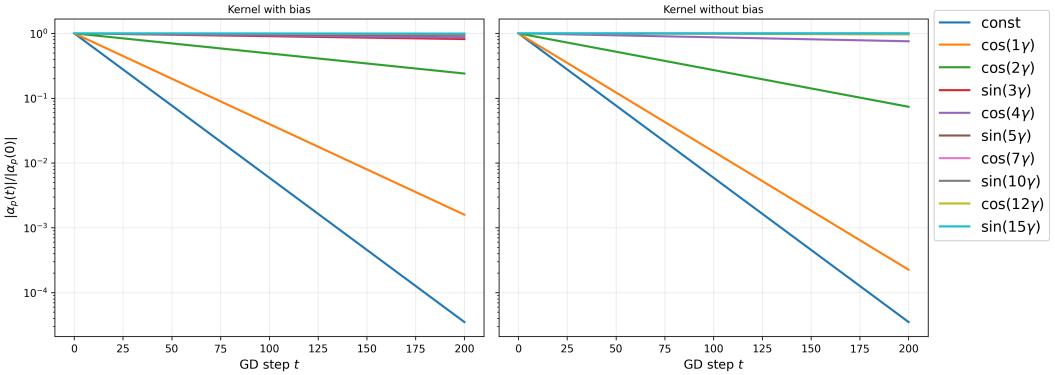


Figure 5.1: Continuum fixed-kernel dynamics for the first 200 gradient descent steps, comparing the kernel with and without the trainable-bias contribution. The curves show normalized mode amplitudes of the residual $|\alpha_p(t)|/|\alpha_p(0)|$ under the diagonal Fourier dynamics.

Figure 5.2 isolates odd Fourier modes. With the bias contribution, these modes have nonzero eigenvalues. Without the bias contribution, the plotted odd modes have zero eigenvalues and remain constant under the fixed-kernel dynamics.

Together, these calculations and plots give the continuum reference case. Here the operator is the deterministic infinite-width integral operator T, defined using the uniform measure on S^1. Thus the Fourier frequency subspaces are exact invariant subspaces of the dynamics, and each residual component decays at the rate given by the corresponding eigenvalue. This is the standard NTK spectral mechanism in an explicit setting. Components aligned with larger kernel eigenvalues decay faster (Jacot et al., 2020; Lee et al., 2020; Cao et al., 2020). Because the domain and measure are rotationally symmetric, this eigenvalue ordering is also a frequency ordering (Bietti and Mairal, 2019; Ronen et al., 2019; Basri et al., 2020).

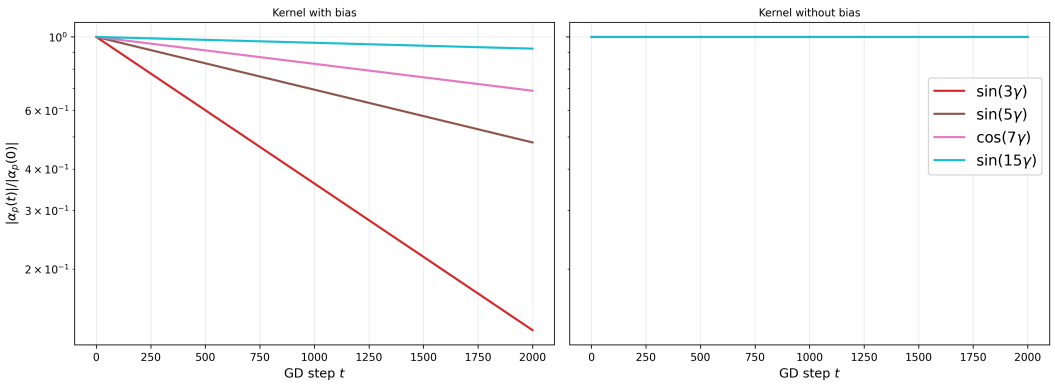


Figure 5.2: Odd-mode continuum fixed-kernel dynamics with and without the trainable-bias contribution. With the bias contribution, the plotted odd modes have nonzero eigenvalues. Without it, the plotted odd modes have zero eigenvalues and therefore remain constant under the fixed-kernel dynamics.

With this continuum baseline in place, we now ask how much of this Fourier structure survives after replacing the uniform measure by finitely many sampled points.

5.2 Finite-sample control on a truncated Fourier subspace

We now turn to the sampled setting introduced in Section 4.3. Throughout, $K \geq 1$ is fixed, $d = 2K+1$, and the truncated Fourier subspace $\mathcal{H}_K$ is given by (4.16). Recall from Section 4.2.3 that, for a basis function in $\mathcal{H}_K$, its frequency means its Fourier index. Thus ϕ_0 has frequency 0, corresponding to the constant mode, while both $\phi_{k,c}$ and $\phi_{k,s}$ have frequency k. The matrices Φ, A, G, and H are defined in (4.20)–(4.23).

For each basis index p, let $k(p)$ denote the frequency of the corresponding basis function ϕ_p, and write

$$\lambda_p := \lambda_{k(p)}.$$

Proposition 5.4 (Expectation of the restricted sampled operator)**.** *Define*

$$\lambda_p^{(n)} = \left(1 - \frac{1}{n}\right)\lambda_p + \frac{\Theta(0)}{n}, \tag{5.9}$$

and let

$$\Lambda^{(n)} = \operatorname{diag}(\lambda_p^{(n)}). \tag{5.10}$$

Then

$$\mathbb{E}[H] = \Lambda^{(n)}. \tag{5.11}$$

The detailed proof is provided in Appendix C.2.

The finite-sample eigenvalue $\lambda_p^{(n)}$ differs from the continuum value λ_p by a term of order $\frac{\Theta(0)}{n}$. This correction shrinks with n, so the sampled eigenvalues approach the continuum

values as the sample size grows. For the kernel values and block dimensions used in the experiments, this correction is small once n is a few hundred.

The next theorem gives explicit concentration bounds for the three sampled quantities that appear in the methodology chapter.

Theorem 5.5 (Finite-sample concentration on the truncated Fourier subspace)**.** *Assume that*

$$\kappa := \sup_{\theta\in[0,2\pi)} |\Theta(\theta)| < \infty. \tag{5.12}$$

Then, for every $\varepsilon > 0$*,*

$$\mathbb{P}(\|G - I_d\|_{\max} \geq \varepsilon) \leq d(d+1)\exp\left(-\frac{n\varepsilon^2}{4+\frac{4}{3}\varepsilon}\right), \tag{5.13}$$

$$\mathbb{P}\left(\|H - \Lambda^{(n)}\|_{\max} \geq \varepsilon\right) \leq 2d^2\exp\left(-\frac{n\varepsilon^2}{32\kappa^2}\right), \tag{5.14}$$

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

$$\mathbb{P}\left(\|A\Phi - \Phi\Lambda^{(n)}\|_{\max} \geq \varepsilon\right) \leq 2nd\exp\left(-\frac{n\varepsilon^2}{2\kappa^2+2\kappa\varepsilon}\right). \tag{5.15}$$

The detailed proofs of the three bounds are provided in Appendix C.

Combining the preceding concentration bounds gives a single high-probability event on which all three finite-sample quantities are controlled at once. Namely, with probability at least $1-\delta$,

$$\|G - I_d\|_{\max}, \qquad \|H - \Lambda^{(n)}\|_{\max}, \qquad \|A\Phi - \Phi\Lambda^{(n)}\|_{\max}$$

are all of order

$$O\left(\sqrt{\frac{\log(nd/\delta)}{n}}\right),$$

up to logarithmic-over-n terms and constants depending on the kernel bound κ. Thus the same sampled points simultaneously make the sampled Fourier basis nearly orthonormal, the compressed operator nearly diagonal, and the sampled operator action close to its corrected diagonal form. The explicit version with constants is given in Appendix C.6.

Theorem 5.5 shows that the sampled operator remains close to the continuum Fourier decomposition on the truncated subspace $\mathcal{H}_K$. To express this directly at the level of the induced matrix on Fourier coefficients, define

$$C := G^{-1}H, \tag{5.16}$$

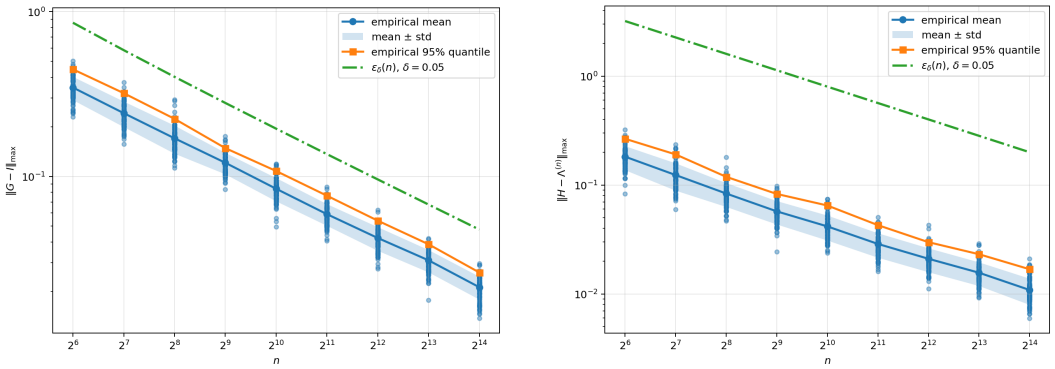
whenever G is invertible.

The factor G^{-1} appears because the sampled Fourier basis is not exactly orthonormal under the empirical inner product. Thus, when G is close to I_d and H is close to $\Lambda^{(n)}$, the coefficient-level restricted operator $C = G^{-1}H$ is close to the diagonal finite-sample prediction $\Lambda^{(n)}$.

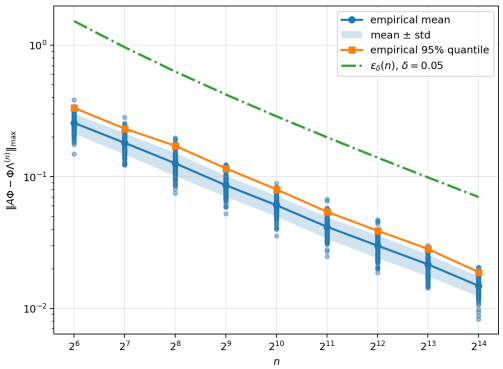
5.2.1 Empirical validation of finite-sample Fourier structure

The finite-sample bounds above give a target-independent test of whether the sampled operator still preserves the continuum Fourier structure on the truncated subspace $\mathcal{H}_K$. The three errors have different roles. The Gram error $\|G - I_d\|_{\max}$ measures how close the sampled Fourier basis is to being orthonormal on the sampled points. The compressed-operator error $\|H - \Lambda^{(n)}\|_{\max}$ measures how close the sampled operator is to the diagonal finite-sample prediction on the retained Fourier coordinates. The action error $\|A\Phi - \Phi\Lambda^{(n)}\|_{\max}$ measures whether applying the sampled operator to sampled Fourier modes produces the predicted mode-wise action.

Figure 5.3 shows the empirical behavior of these three errors appearing in Theorem 5.5. All three errors decrease as n increases. The plotted high-probability envelopes are conservative, but they have the same decreasing trend.



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



(c) $\|A\Phi - \Phi\Lambda^{(n)}\|_{\max}$

Figure 5.3: Finite-sample diagnostics for the sampled Fourier approximation. The three plotted errors decrease with n. The theoretical envelopes are conservative, but they show the same decreasing trend.

These bounds naturally lead to the question of what finite sampling does to the residual dynamics. In the continuum setting, (5.8) shows that Fourier components decay indepen-

dently. On the retained subspace, this gives the coefficient system

$$\dot{\alpha}(t) = -\Lambda\alpha(t).$$

After sampling, if the residual is mainly represented in the retained Fourier span, then

$$\widehat{r}(t) \approx \Phi\alpha(t),$$

and the sampled kernel dynamics

$$\dot{\widehat{r}}(t) = -A\widehat{r}(t)$$

gives

$$\Phi\dot{\alpha}(t) \approx -A\Phi\alpha(t).$$

Using the finite-sample decomposition

$$A\Phi = \Phi\Lambda^{(n)} + \big(A\Phi - \Phi\Lambda^{(n)}\big),$$

we obtain

$$\Phi\dot{\alpha}(t) \approx \underbrace{-\Phi\Lambda^{(n)}\alpha(t)}_{\text{finite-sample eigenvalue prediction}} \underbrace{-\big(A\Phi - \Phi\Lambda^{(n)}\big)\alpha(t)}_{\text{finite-sample action error}}.$$

If the action-error term were zero, each retained Fourier component would follow

$$\alpha_p(t) = \exp(-\lambda_p^{(n)}t)\alpha_p(0) \tag{5.17}$$

The action-error term is small by the finite-sample bounds, but it is not zero. To see when it can become visible, suppose that the retained coefficients are normalized so that $\alpha_p(0) = 1$ for all retained basis indices p in $\mathcal{H}_K$, and suppose initially that they follow the finite-sample eigenvalue prediction from (5.17). Then

$$\alpha(t) \approx \exp(-\Lambda^{(n)}t)\mathbf{1},$$

and hence

$$\widehat{r}(t) \approx \Phi\exp(-\Lambda^{(n)}t)\mathbf{1} = \sum_p \exp(-\lambda_p^{(n)}t)\Phi_p,$$

where Φ_p is the sampled vector of the p-th retained Fourier basis function and $\mathbf{1} \in \mathbb{R}^d$ is the vector with all entries equal to one.

Substituting this expression into the action-error term gives

$$\big(A\Phi - \Phi\Lambda^{(n)}\big)\alpha(t) \approx \big(A\Phi - \Phi\Lambda^{(n)}\big)\exp(-\Lambda^{(n)}t)\mathbf{1}.$$

Therefore

$$\Big\|\big(A\Phi - \Phi\Lambda^{(n)}\big)\alpha(t)\Big\|_{\max} \le \|A\Phi - \Phi\Lambda^{(n)}\|_{\max}\sum_p \exp(-\lambda_p^{(n)}t).$$

This is a crude upper bound on the extra motion caused by finite sampling, and gives a scale at which deviations from the predicted eigenvalue decay become visible.

For a single retained component p, the eigenvalue prediction gives

$$|\alpha_p(t)| \approx \exp(-\lambda_p^{(n)} t), \qquad |\dot{\alpha}_p(t)| \approx \lambda_p^{(n)} \exp(-\lambda_p^{(n)} t).$$

Thus, the predicted movement of component p is $\lambda_p^{(n)} \exp(-\lambda_p^{(n)} t)$. This eigenvalue prediction should therefore be reliable while

$$\lambda_p^{(n)} \underbrace{\exp(-\lambda_p^{(n)} t)}_{\alpha_p(t)} \gg \|A\Phi - \Phi\Lambda^{(n)}\|_{\max} \sum_q \underbrace{\exp(-\lambda_q^{(n)} t)}_{\alpha_q(t)}.$$

The left-hand side is the motion predicted for component p by its own eigenvalue. This term becomes small as $\alpha_p(t)$ decays. The right-hand side is a bound on the extra motion coming from the finite-sample action error. This term is small when the action error is small, but it depends on all retained components, including components that decay more slowly.

The comparison above suggests a two-stage picture. At early times, the retained components have not yet decayed much, so the eigenvalue-driven motion is larger than the finite-sample action error. The sampled curves should therefore follow the finite-sample eigenvalue prediction. At later times, after some components have decayed substantially, the same action error can become visible relative to the remaining motion. Figures 5.4 and 5.5 show these two regimes for $n = 4096$ uniformly sampled points.

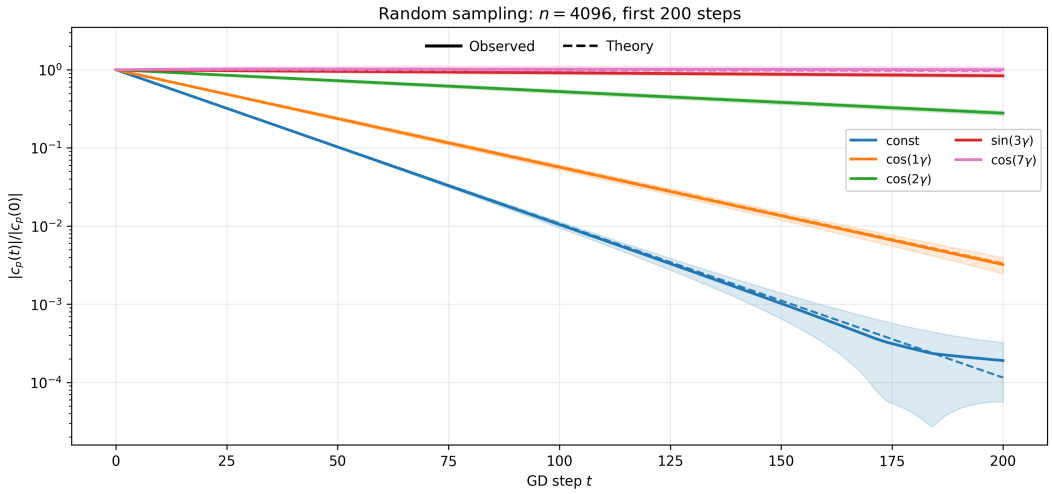


Figure 5.4: Random sampling with $n = 4096$, first 200 gradient descent steps. Solid curves show the observed normalized residual amplitudes $|\alpha_p(t)|/|\alpha_p(0)|$, while dashed curves show the diagonal prediction from the finite-sample corrected eigenvalues $\lambda_p^{(n)}$.

However, the level at which an individual curve starts to flatten cannot be computed directly from the max-norm bounds in Theorem 5.5 because these bounds control the largest possible error over the whole retained Fourier block. They are therefore useful for explaining why deviations can appear once the predicted motion becomes small, but they do not determine the deviation level of each Fourier component separately. That level depends on

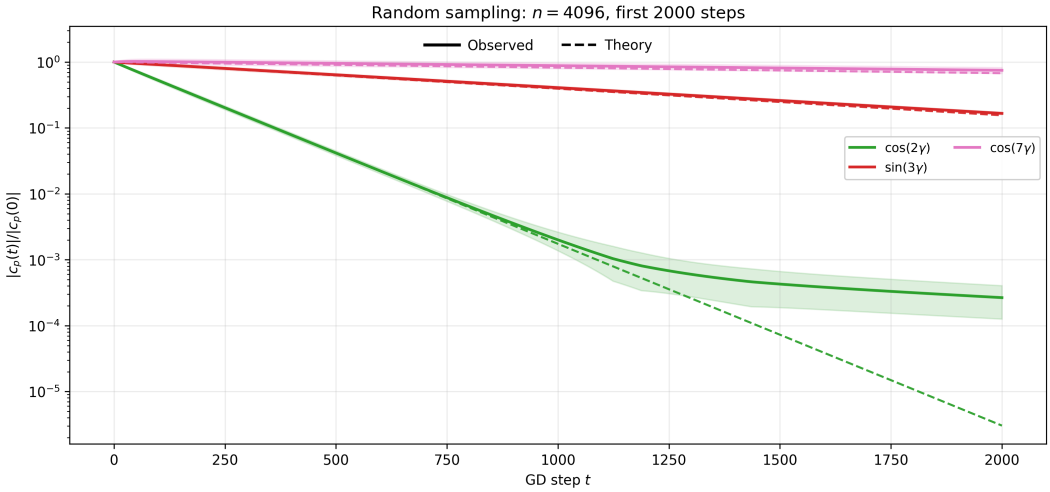


Figure 5.5: Random sampling with $n = 4096$, higher-frequency retained modes over 2000 gradient descent steps. The diagonal prediction captures the initial decay rates, while small deviations from this prediction become visible at later times.

the actual sampled operator for the particular draw of training points, and on which other Fourier components are still present when the component under consideration has already decayed.

5.2.2 Interpretation of the finite-sample results

The finite-sample results show that the continuum Fourier picture is not only a population-level statement. For a fixed low-frequency block $\mathcal{H}_K$, random sampling perturbs the Fourier structure in a controlled way. The sampled basis is not exactly orthonormal, and the sampled Fourier modes are not exact eigenvectors of the sampled operator, but the errors measured by $G - I_d$, $H - \Lambda^{(n)}$, and $A\Phi - \Phi\Lambda^{(n)}$ decrease with n.

This supports the usual NTK spectral-bias mechanism in a finite-sample setting. In the continuum model, the Fourier modes diagonalize the NTK operator and the eigenvalues determine the decay rates. The results above show that, on a fixed retained block, the sampled operator remains close to this diagonal Fourier description. This complements the population-level harmonic analyses of NTK spectra on the sphere and circle (Ronen et al., 2019; Bietti and Mairal, 2019; Cao et al., 2020; Basri et al., 2020) by making explicit how the same structure survives after replacing the uniform measure by finitely many sampled points.

The residual experiments then show the dynamical consequence. The corrected finite-sample eigenvalues $\lambda_p^{(n)}$ predict the early decay of the retained Fourier components, while the finite-sample action error explains why late-time deviations can appear once some components have become small. Thus finite sampling does not remove the spectral-bias mechanism; it turns the exact continuum diagonal dynamics into an approximate low-frequency description.

This suggests a direct test. If the decay of each retained component is mainly governed by its eigenvalue, then an eigenvalue-based rescaling should remove most of the frequency-dependent decay. The next experiment tests this prediction directly.

5.3 Preconditioned sampled dynamics

We next apply the Fourier-block preconditioning construction from Section 4.4. The goal is to compare the baseline sampled dynamics with two block-corrected dynamics: one using the diagonal finite-sample prediction and one using the observed sampled Fourier block.

5.3.1 Experimental setup

The three sampled dynamics are

$$B_{\text{base}} = A, \qquad B_{\text{th}} = P_{\text{th}}A, \qquad B_{\text{emp}} = P_{\text{emp}}A.$$

Here P_{th} is defined in (4.31) and uses the diagonal prediction $\Lambda^{(n)}$. The empirical preconditioner P_{emp}, defined in (4.32), uses the observed coefficient-space block $C_n^{(K)} = G^{-1}H$.

In this experiment, we use the target

$$f^*(\gamma) = 1 + 0.9\cos(\gamma) + 0.5\cos(2\gamma) - 0.35\sin(3\gamma) + 0.25\cos(7\gamma). \tag{5.18}$$

The retained preconditioning block contains all active target frequencies shown in the residual plots.

5.3.2 Mode-wise residual decay

Figure 5.6 shows the residual amplitudes for the baseline, theory-preconditioned, and empirically preconditioned dynamics. The baseline dynamics show separated decay rates across retained modes. Both preconditioned dynamics reduce this separation, with the empirical preconditioner giving the flattest curves.

Figure 5.7 compares the final prediction on a dense probe grid. The baseline fit leaves a visible residual, while both preconditioned flows more closely match the target.

5.3.3 Theory versus empirical preconditioning

Figure 5.8 shows the spectrum obtained after applying the theory preconditioner to the sampled Fourier block. At the coefficient level, the empirical block is $C = G^{-1}H$. The empirical preconditioner would invert this block directly, giving $C_{\text{emp}}^{-1}C = I_d$, so all eigenvalues would be equal to 1. The theory preconditioner instead rescales by the predicted diagonal eigenvalues $\Lambda^{(n)}$. Thus it is exact only when $C = \Lambda^{(n)}$. The figure shows that, as n increases, the theory-preconditioned spectrum moves closer to the ideal value 1. This is the empirical counterpart of the finite-sample bounds above: the sampled block C approaches the diagonal prediction $\Lambda^{(n)}$, so the diagonal theory preconditioner becomes a better approximation to the empirical block inverse.

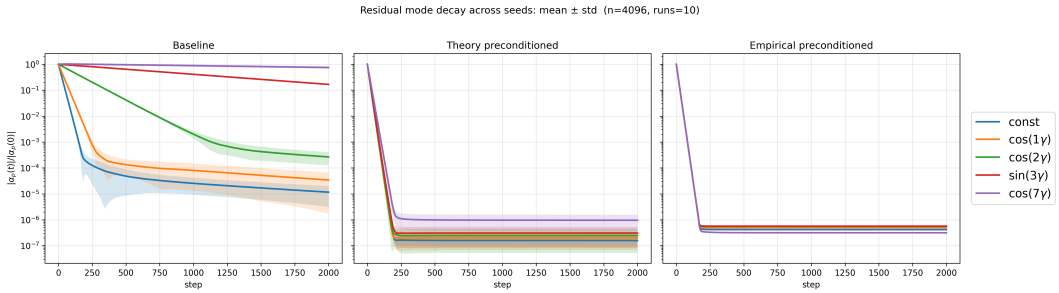


Figure 5.6: Preconditioned sampled dynamics at $n = 4096$. The baseline dynamics shows separated decay rates across retained Fourier modes. Both preconditioned dynamics make the retained residual amplitudes decay on more similar time scales. The empirical preconditioner gives the most uniform decay, while the theory preconditioner uses only the predicted diagonal scaling.

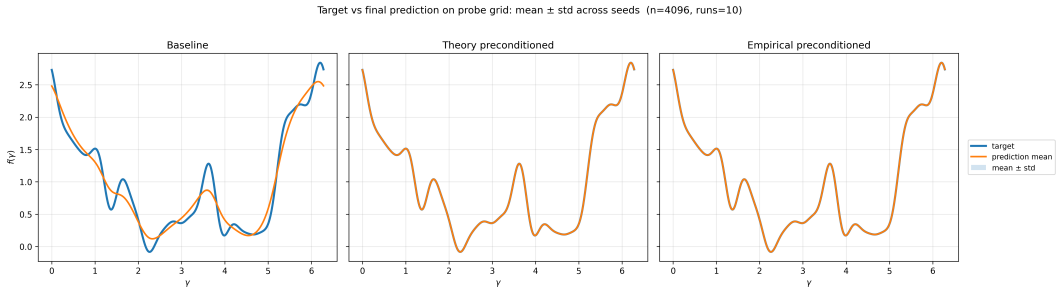


Figure 5.7: Final prediction on a dense probe grid for the same experiment as Figure 5.6. The baseline fit leaves a visible residual, while both preconditioned dynamics closely match the target. The theory and empirical preconditioners are nearly indistinguishable at the level of the final fitted function.

Figure 5.9 compares the theory and empirical constructions at two levels. The preconditioner P is an operator on sampled function values. Given a vector in $\mathbb{R}^n$, it projects that vector onto the retained Fourier block, rescales the retained Fourier coefficients, and maps the result back to $\mathbb{R}^n$. Thus, the gap between P_{th} and P_{emp} measures how different the two correction steps are before they are combined with the sampled kernel operator.

The product PA is the operator that is actually used in the residual dynamics. It first applies the sampled kernel operator A to the residual, and then applies the correction P to the retained Fourier components of the result. Thus, the gap between B_{th} and B_{emp} measures how different the resulting training dynamics are. Both gaps decrease with sample size.

The preconditioning experiment gives a second test of the finite-sample interpretation. The theory preconditioner uses only the predicted diagonal block $\Lambda^{(n)}$, while the empirical preconditioner uses the observed block $C_n^{(K)} = G^{-1}H$. If most of the rate separation is caused by the Fourier eigenvalues, then rescaling by $\Lambda^{(n)}$ should already make the retained modes decay on more similar time scales. If the remaining finite-sample mixing inside the retained block is important, then the empirical preconditioner should give a visibly different

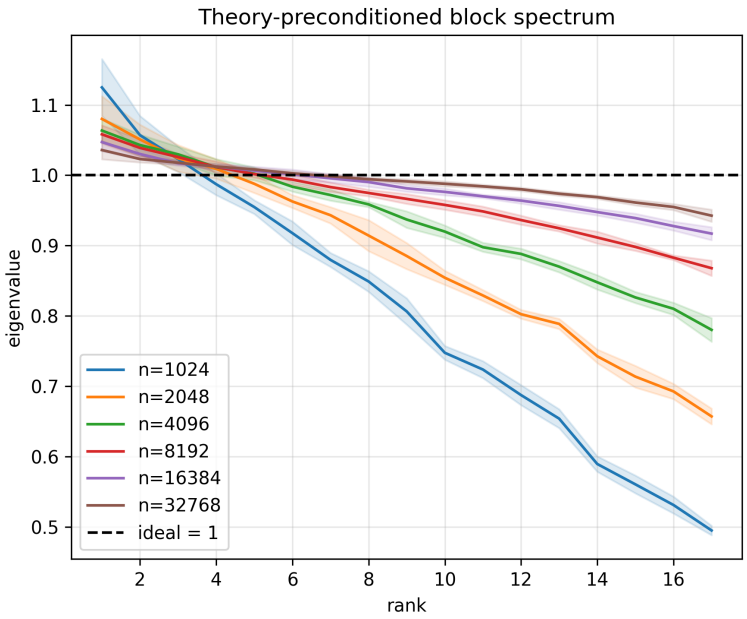
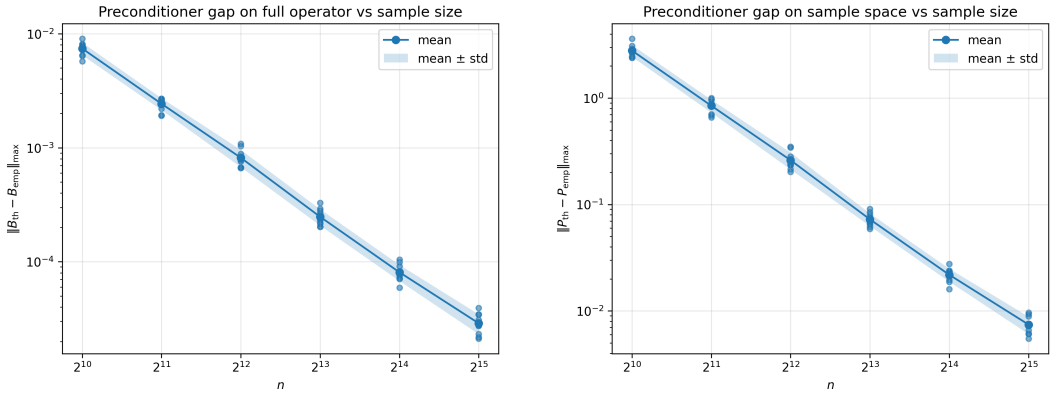


Figure 5.8: Spectrum of the theory-preconditioned sampled Fourier block as a function of sample size. If the theory preconditioner exactly inverted the sampled block, all eigenvalues would be equal to 1. As n increases, the spectrum becomes flatter and moves closer to this ideal value.



(a) $\|B_{\text{th}} - B_{\text{emp}}\|_{\max}$ (b) $\|P_{\text{th}} - P_{\text{emp}}\|_{\max}$

Figure 5.9: Gap between the theory and empirical preconditioners as sample size increases. The empirical preconditioner uses the observed block $C = G^{-1}H$, while the theory preconditioner uses the diagonal approximation $\Lambda^{(n)}$. The decreasing gap is consistent with the finite-sample result $C \approx \Lambda^{(n)}$.

dynamics.

Figures 5.6, 5.8, and 5.9 show that the theory preconditioner already removes most of the rate separation on the retained block. The empirical preconditioner is closer to the ideal flattened dynamics because it also corrects the sample-dependent mixing in $C_n^{(K)}$. The decreasing gap between the two preconditioners as n increases is consistent with the finite-sample result $C_n^{(K)} \approx \Lambda^{(n)}$.

5.3.4 Conditioning and the numerical floor

The preconditioned curves flatten at a level set by the numerical conditioning of the problem. The condition number κ of a matrix measures how much it amplifies relative errors when used to solve a linear system: a large κ means small input errors can produce large errors in the solution, while κ close to 1 means errors are barely amplified. The relevant matrix here is the retained Fourier block. Before preconditioning this is $C = G^{-1}H$, the block whose deviation from $\Lambda^{(n)}$ we have been tracking; after the theory preconditioner it becomes $C_{\text{th}} = (\Lambda^{(n)})^{-1}C$, which the preconditioner aims to bring close to the identity.

Table 5.1 reports both. The uncorrected block C has condition number around 330 across sample sizes. This stays roughly constant because it reflects the conditioning of the continuum block that C converges to, set largely by the spread of the retained eigenvalues $\lambda_0, \ldots, \lambda_K$, the largest of which is a few hundred times the smallest. This spread does not depend on n, so only the small finite-sample correction makes $\kappa(C)$ drift slightly as n grows. After the theory preconditioner, $\kappa(C_{\text{th}})$ drops toward 1 as n grows, which is the conditioning counterpart of $C_n^{(K)} \to \Lambda^{(n)}$: as the sampled block approaches the diagonal prediction, applying the inverse of that diagonal prediction leaves an operator close to the identity. We omit the empirical preconditioner, since it inverts the block exactly, $C_{\text{emp}}^{-1}C = I_d$, giving $\kappa = 1$ by construction.

Table 5.1: Conditioning of the retained Fourier block and the resulting finite-precision floors, over 10 runs. $\kappa(C)$ is the condition number of the uncorrected sampled block, $\kappa(C_{\text{th}})$ the condition number after the theory preconditioner is applied. The last two columns give the relative-accuracy floor κu, the condition number times the unit roundoff u, at single and double precision.

n	$\kappa(C)$	$\kappa(C_{\text{th}})$	κu (fp32)	κu (fp64)
1024	340.8	2.273	2.7×10^{-7}	5.0×10^{-16}
2048	333.4	1.644	1.9×10^{-7}	3.6×10^{-16}
4096	329.5	1.364	1.6×10^{-7}	3.0×10^{-16}
8192	328.6	1.220	1.4×10^{-7}	2.7×10^{-16}
16384	327.1	1.142	1.4×10^{-7}	2.5×10^{-16}
32768	325.7	1.099	1.3×10^{-7}	2.4×10^{-16}

Why conditioning sets the floor is visible in Figure 5.10, which repeats the experiment in single and double precision. In finite-precision arithmetic, every operation introduces a relative error of size u, the unit roundoff ($u \approx 10^{-7}$ in single precision, 10^{-16} in double).

Solving with a matrix of condition number κ amplifies these errors, so the relative accuracy of the result cannot be driven below about κu (Trefethen and Bau, 2022). Because the theory preconditioner makes $\kappa(C_{\text{th}})$ close to 1, this floor is essentially u itself, and the curves plateau there. Moving from fp32 to fp64 lowers u by nine orders of magnitude and lowers the plateaus correspondingly, which confirms the floor is numerical.

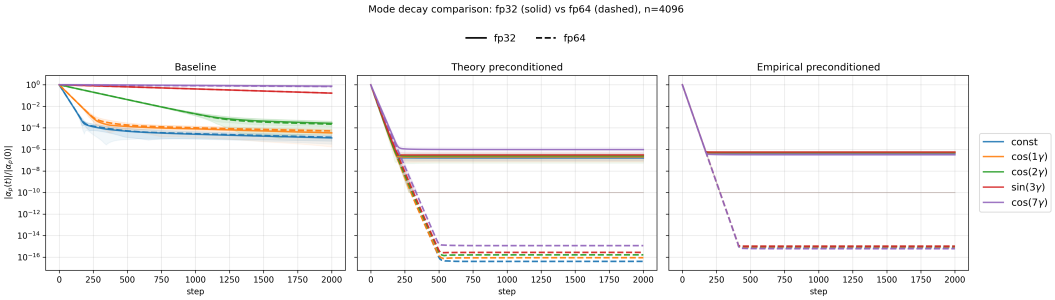


Figure 5.10: Mode decay at $n = 4096$ in single (solid) and double (dashed) precision, for the baseline, theory-preconditioned, and empirically preconditioned dynamics. The preconditioned plateaus drop by several orders of magnitude from fp32 to fp64, identifying them as a finite-precision floor.

5.3.5 Targets with out-of-block frequencies

The experiments so far kept all active target frequencies inside the retained block. Figure 5.11 repeats the comparison with a target that also contains frequencies outside the block, $\sin(10\gamma)$ and $\cos(12\gamma)$. The preconditioner is still built only on the retained block, so it accelerates the in-block modes as before but leaves the out-of-block modes near their initial amplitude.

The in-block modes also plateau higher than in Figure 5.6. This floor is not numerical: the single- and double-precision curves reach nearly the same plateaus, so roundoff is ruled out. We expect the cause to be coupling from the out-of-block components. The sampled Fourier modes are only approximate eigenvectors of the sampled operator, so the uncorrected modes at $k = 10$ and $k = 12$ can feed back into the retained coordinates, and the in-block residuals would then settle at a level set by this coupling. We have not isolated this mechanism directly, but it is consistent with the modes being only approximate eigenvectors and with the precision comparison ruling out roundoff.

5.3.6 Discussion

The preconditioning results connect to existing work on modifying NTK training dynamics by changing the effective kernel spectrum. In particular, Geifman et al. (2024) use preconditioned gradient descent to alter the trajectory of wide-network training by modifying the spectrum of the associated kernel. Recent work also studies preconditioned methods as a way to reduce spectral bias and accelerate exploration of the NTK space (Jiang et al., 2026;

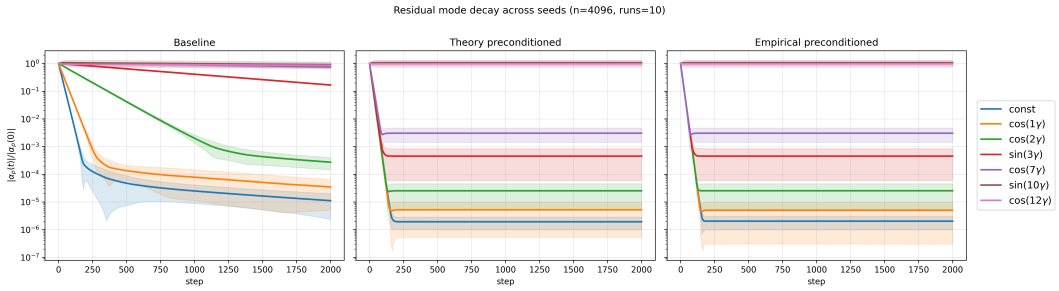


Figure 5.11: Preconditioned sampled dynamics with target frequencies outside the retained block, $\sin(10\gamma)$ and $\cos(12\gamma)$, at $n = 4096$ over 10 runs. The preconditioner accelerates the in-block modes but leaves the out-of-block modes near their initial amplitude. The in-block plateaus are higher than when all target modes lie inside the block, reflecting coupling from the uncorrected out-of-block components.

Yang, 2024). Our setting is more limited, since the preconditioner acts only on a fixed retained Fourier block. The advantage is that the effect is explicit: rescaling by the predicted eigenvalues directly compensates for the different decay rates of the retained modes.

A practical use of this observation is as a block-level correction when the target is known, or assumed, to be concentrated on a prescribed low-frequency Fourier range. In that case, the theory preconditioner does not require estimating the full sampled block inverse. It only uses the predicted finite-sample eigenvalues $\lambda_p^{(n)}$, and can make the retained components converge on more similar time scales. The empirical preconditioner gives the best correction when the sampled block is available, while the theory preconditioner provides a cheaper approximation whose accuracy improves as the finite-sample Fourier block becomes closer to diagonal.

Overall, the finite-sample results show that random sampling turns the exact continuum Fourier dynamics into an approximate low-frequency description. The sampled operator is no longer exactly diagonal in the Fourier basis, but its deviation from the corrected diagonal prediction decreases with n and explains the observed residual dynamics.

So far, however, the kernel itself has still been the deterministic infinite-width NTK. In an actual finite-width network, the tangent kernel at initialization is built from randomly initialized weights, so even before training it is a random finite-width approximation of the continuum kernel. The next section therefore asks whether the same Fourier structure survives this second perturbation: finite width.

5.4 Finite-width control of the frozen tangent operator

We now turn from sampling error to width-induced error. Throughout this section, the input space remains the continuum S^1, and the object of interest is the frozen finite-width operator $B_m^{(K)}$ on the truncated Fourier subspace $\mathcal{H}_K$, introduced in (4.41). The corresponding continuum reference block Λ_K was defined in (4.42).

The next theorem shows that, for each fixed K, the frozen finite-width block is a controlled perturbation of the continuum Fourier block.

Theorem 5.6 (Finite-width concentration of the frozen low-mode operator)**.** *Fix $K \geq 1$, and let $d = 2K+1$. Then*

$$\mathbb{E}[B_m^{(K)}] = \Lambda_K. \tag{5.19}$$

Moreover, there exists a universal constant $c > 0$ such that, for every $\varepsilon > 0$,

$$\mathbb{P}\Big(\|B_m^{(K)} - \Lambda_K\|_{\max} \geq \varepsilon \Big) \leq 2d^2 \exp\left(-cm \min\left(\frac{\varepsilon^2}{32^2}, \frac{\varepsilon}{32} \right) \right). \tag{5.20}$$

Equivalently, for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|B_m^{(K)} - \Lambda_K\|_{\max} \leq 32 \max\left\{ \sqrt{\frac{1}{cm} \log\left(\frac{2d^2}{\delta} \right)}, \frac{1}{cm} \log\left(\frac{2d^2}{\delta} \right) \right\}. \tag{5.21}$$

In particular, for fixed K,

$$\|B_m^{(K)} - \Lambda_K\|_{\max} = O\left(\sqrt{\frac{\log K}{m}} \right). \tag{5.22}$$

The detailed proof is provided in Appendix D.

Theorem 5.6 shows that, on a fixed truncated Fourier subspace, the frozen finite-width tangent block is a random perturbation of the continuum Fourier block. The perturbation decreases with width at the usual $m^{-1/2}$ concentration scale, up to logarithmic factors in the block dimension. However, the theorem contains an unspecified universal constant $c > 0$. This constant comes from Bernstein's inequality for averages of centered sub-exponential random variables (Vershynin, 2018a). In explicit proofs of such inequalities, constants of order $1/4$ often appear after converting a ψ_1-norm bound into a moment generating function bound. We next check this concentration behavior numerically where we use $c = 0.2$ as a conservative representative value, not an optimized constant for this problem. Changing this constant shifts the bound but does not change the predicted rate as m increases.

For each width m, we sample 100 independent initializations and compute

$$\|B_m^{(K)} - \Lambda_K\|_{\max}.$$

Figure 5.12 shows the median and 95% quantile of this error, together with the rate envelope suggested by Theorem 5.6. The empirical errors decrease steadily with width. The theoretical envelope is much larger than the observed errors, but it has the correct qualitative trend: the finite-width block concentrates around the continuum Fourier block as m increases.

This is the finite-width analogue of the finite-sample control from the previous section. In the finite-sample case, the randomness comes from replacing the uniform measure by sampled points. Here, the randomness comes from the finite number of hidden units at initialization. In both cases, on a fixed low-frequency Fourier block, the operator remains

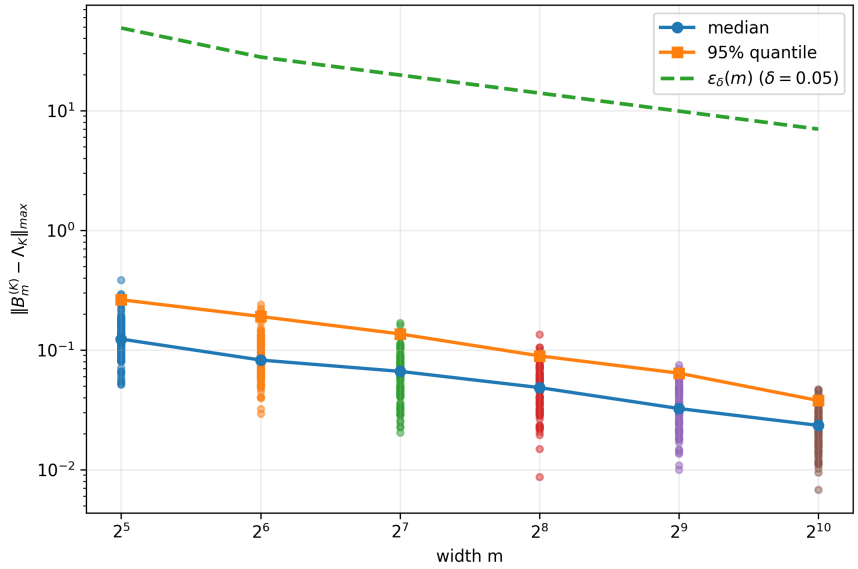


Figure 5.12: Finite-width concentration of the frozen tangent block. For each width m, the plot shows the median and 95% quantile of $\|B_m^{(K)} - \Lambda_K\|_{\max}$ across 100 independent initializations. The dashed curve shows the rate envelope from Theorem 5.6 with $\delta = 0.05$ and the illustrative constant $c = 0.2$. The empirical errors decrease with width, while the explicit envelope is conservative.

close to the deterministic continuum Fourier block, and the deviation decreases as the relevant size parameter (number of samples or width of network) increases.

Thus finite width alone does not destroy the Fourier structure of the frozen tangent operator. For a particular random initialization, the operator is not exactly diagonal in the Fourier basis, but Theorem 5.6 and Figure 5.12 show that this mismatch shrinks with width. This supports the use of the infinite-width NTK as a baseline, while also showing why finite-width fluctuations can matter quantitatively (Jacot et al., 2020; Lee et al., 2020; Bordelon and Pehlevan, 2023; Vyas et al., 2022).

The next question is whether this entrywise concentration also corresponds to geometric agreement of the frequency subspaces themselves.

5.4.1 Empirical alignment of Fourier frequency subspaces

The block concentration result controls the matrix $B_m^{(K)}$ on the retained Fourier coordinates. We next examine the corresponding eigenspaces directly. Using the quantities introduced in Section 4.5.5, we compare each Fourier frequency subspace $\mathcal{F}_k$ with a matched empirical eigenspace $\widehat{\mathcal{F}}_k$ of the frozen finite-width operator.

We first summarize the comparison using the projector error

$$\varepsilon_{k,F}^{(m)} = \|\widehat{P}_k - P_k\|_F.$$

This gives one scalar mismatch measure for each frequency subspace. By (4.50), it is equivalent to combining the principal angles between $\mathcal{F}_k$ and $\widehat{\mathcal{F}}_k$.

Figure 5.13 shows the mean and standard deviation of $\varepsilon_{k,F}$ across random initializations. For all displayed frequencies, the projector error decreases as the width increases. Higher frequencies generally have larger projector error at a fixed width. One exception in this experiment is $k = 2$, which has the smallest projector error among the displayed frequencies. A plausible explanation is the structural role of $k = 0$ and $k = 1$ in this chosen architecture. The constant mode receives a direct eigenvalue contribution from the trainable hidden-bias term (refer Theorem 5.2). The first frequency, $k = 1$, is tied to the coordinate embedding,

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

and it is also the only odd frequency present in the no-bias kernel (Corollary 5.3), with

$$\lambda_1^{(\mathrm{nb})} = \frac{1}{4}, \qquad \lambda_k^{(\mathrm{nb})} = 0 \quad \text{for odd } k \geq 3.$$

These features may make the finite-width perturbation act differently on $k = 0$ and $k = 1$ than on the higher frequency blocks. We do not investigate this mechanism further here, so the $k = 2$ behavior is an empirical observation and not a theoretical claim.

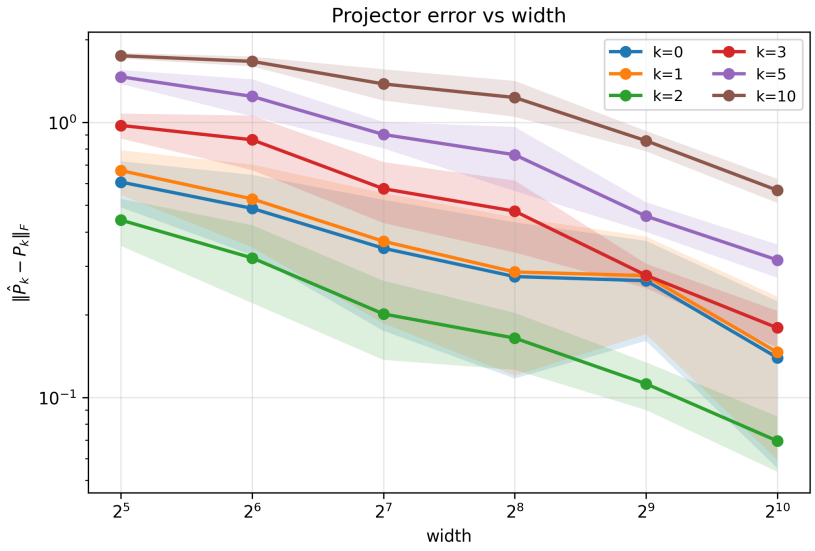


Figure 5.13: Projector error $\|\widehat{P}_k - P_k\|_F$ between the Fourier frequency subspace $\mathcal{F}_k$ and the matched empirical eigenspace $\widehat{\mathcal{F}}_k$ of the frozen finite-width operator at initialization. The markers show the mean over random initializations and the shaded regions show one standard deviation. The error decreases with width for all displayed frequencies.

The projector error gives one scalar measure of subspace mismatch. As a complementary summary, we also report the subspace cosine similarity defined in (4.49). This quantity averages the squared cosines of the principal angles between $\mathcal{F}_k$ and $\widehat{\mathcal{F}}_k$, and lies between

0 and 1. Larger values indicate better alignment, with value 1 corresponding to perfect agreement of the two subspaces.

Figure 5.14 shows this alignment score across widths. The alignment improves with m for all displayed frequencies, matching the trend in the projector-error plot. Lower frequencies are already well aligned at small width, while higher frequencies require larger widths to approach the continuum Fourier subspaces. As before, $k = 2$ is unusually well aligned compared with neighbouring displayed frequencies.

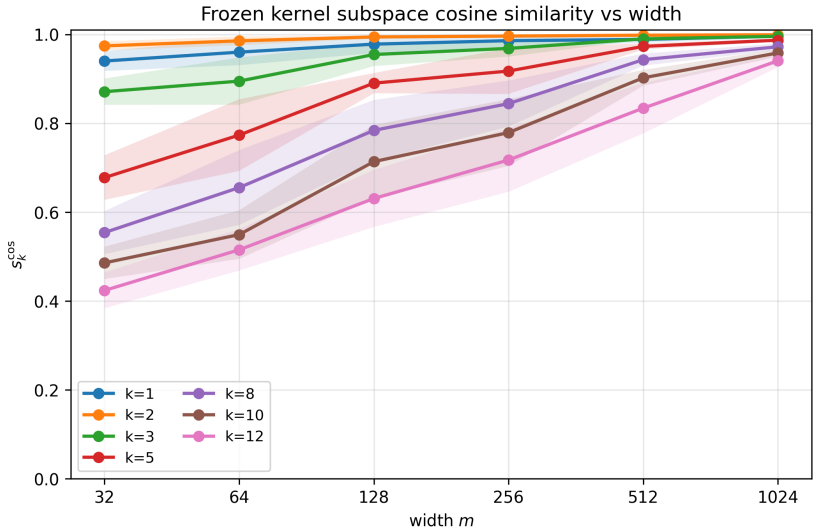


Figure 5.14: Subspace cosine similarity between the Fourier frequency subspace $\mathcal{F}_k$ and the matched empirical eigenspace $\widehat{\mathcal{F}}_k$ of the frozen finite-width operator at initialization. The markers show the mean over random initializations and the shaded regions show one standard deviation. Larger values indicate better subspace alignment, with value 1 corresponding to perfect alignment. The alignment improves with width for all displayed frequencies, with higher frequencies requiring larger widths to approach the continuum Fourier subspaces.

Taken together, Figures 5.12, 5.13, and 5.14 show the same finite-width trend from complementary viewpoints. The matrix block $B_m^{(K)}$ approaches the continuum block Λ_K as m increases, and the matched empirical eigenspaces become closer to the corresponding Fourier frequency subspaces. The improvement is strongest at small and moderate widths for the higher frequencies, which are the most poorly aligned at small width.

The concentration theorem controls the absolute error $|B_m^{(K)} - \Lambda_K|_{\max}$. To interpret its effect on a particular Fourier component, this error should be compared with the size of the corresponding continuum eigenvalue. In the continuum fixed-kernel model, mode k decays at rate λ_k, and Theorem 5.2 gives

$$\lambda_k = O(k^{-2}) \qquad (k \to \infty),$$

for both even and odd frequencies. Thus, the same absolute perturbation is less important for low-frequency modes with larger eigenvalues, and can be more visible for high-frequency

modes with smaller eigenvalues. This explains why the finite-width perturbation is expected to matter more for higher frequencies.

Figures 5.14 and 5.13, support this interpretation geometrically. For $k \geq 1$, each Fourier frequency corresponds to a two-dimensional plane spanned by the cosine and sine modes. At finite width, the matched empirical eigenspace need not coincide exactly with this plane. The projector-error and cosine-similarity plots show that this mismatch decreases with width. They also show that the amount of alignment can vary substantially across frequencies at the same width. One plausible explanation is that eigenspace stability depends not only on the size of the finite-width perturbation, but also on the local spectral separation of the corresponding continuum eigenspace. Since higher-frequency eigenvalues are smaller and closer together, their eigenspaces may be more sensitive to the same absolute perturbation.

Overall, the frozen finite-width results show that finite width does not destroy the Fourier structure at initialization. It introduces a random perturbation of the continuum Fourier operator, and this perturbation decreases with width. The result is, therefore, a frozen-kernel baseline: before training changes the tangent operator, the continuum Fourier eigenspaces remain a useful reference, but finite-width fluctuations can already affect higher-frequency components more strongly. This is consistent with empirical observations that higher-frequency components are learned later and are more sensitive to the data distribution and to perturbations (Rahaman et al., 2019; Ronen et al., 2019; Basri et al., 2020).

The part not settled here is the evolving finite-width case. During training, the tangent operator can change, so the frozen Fourier block structure need not remain fixed. The finite-width result therefore gives us a frozen-kernel baseline but not a complete theory of finite-width training (Bordelon and Pehlevan, 2023; Vyas et al., 2022; Golikov et al., 2022).

5.5 Evolving finite-width tangent kernel

The previous section studied the frozen tangent operator at initialization. We now compare this frozen baseline with the tangent operator that evolves during training. Unlike the finite-width concentration result above, this part is empirical. The goal is to understand how kernel evolution changes the Fourier picture at finite width.

5.5.1 Loss decay across widths

We first compare the least-squares objective under frozen and evolving finite-width tangent-kernel dynamics across widths. The cross-width comparison in Figure 5.15 gives the overall trend, while Figure 5.16 shows the width-by-width frozen versus evolving comparison in more detail.

In the frozen regime, the objective decreases more quickly as the width increases. In the evolving regime, the curves are more tightly clustered across widths. At widths 128 and 256, the evolving kernel reaches a lower final objective than the frozen baseline. At width 512, the difference is smaller. At width 1024, the frozen kernel reaches the lower final objective.

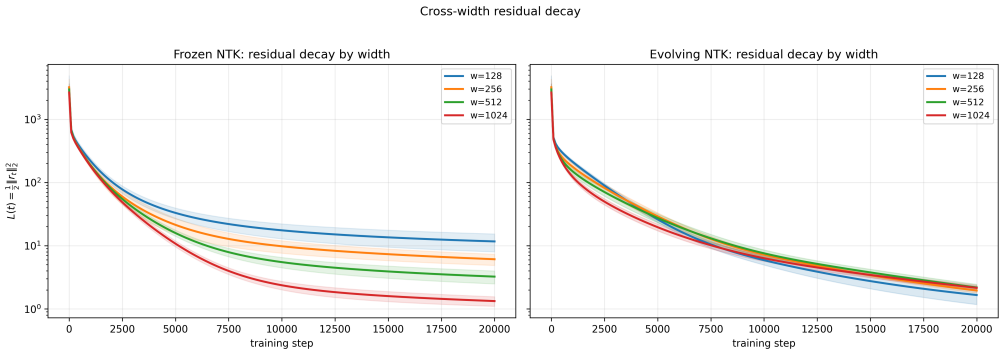
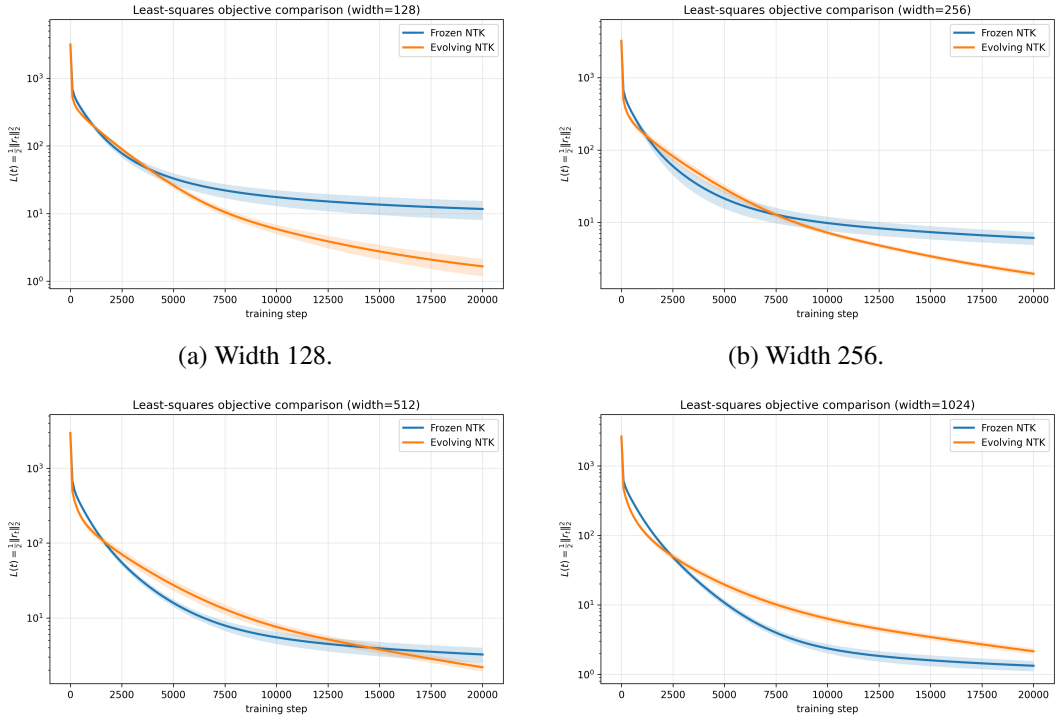


Figure 5.15: Cross-width comparison of the least-squares objective $L(t) = \frac{1}{2}\|r_t\|_2^2$ under frozen and evolving tangent-kernel dynamics. In the frozen regime, larger width gives consistently better residual decay. In the evolving regime, the curves are closer across widths.



(a) Width 128.

(b) Width 256.

(c) Width 512.

(d) Width 1024.

Figure 5.16: Least-squares objective comparison between frozen and evolving finite-width tangent-kernel dynamics across widths. The evolving kernel gives a clear improvement at widths 128 and 256, still improves the final objective at width 512, and loses its advantage by width 1024, where the frozen kernel reaches the lower final loss.

This raises the main question for the rest of the section: what changes in the evolving tangent kernel allow it to help at small width, but not at large width?

5.5.2 Mode-wise decay, Fourier geometry, and spectral strength

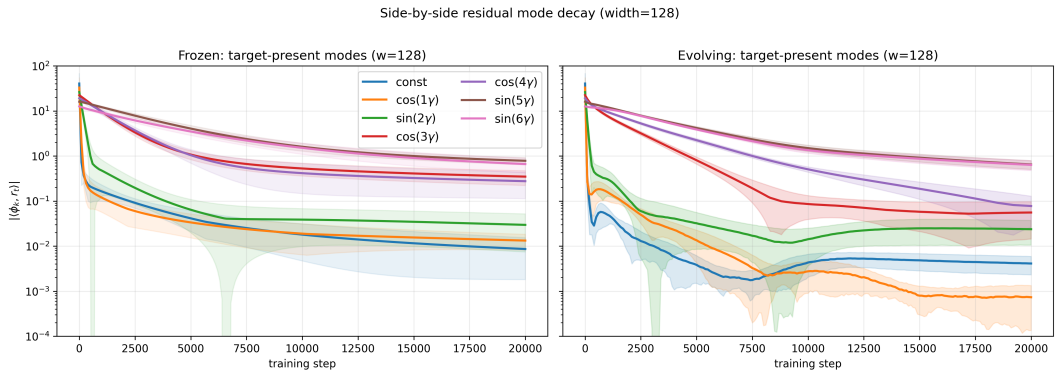
To understand the loss comparison above, we examine three quantities together. The mode-wise residual decay plots show how individual Fourier components are learned over time. The subspace cosine similarity plots measure how closely the eigenspaces of the evolving tangent-kernel remain aligned with the Fourier subspaces. The spectral-strength plots measure the average eigenvalue inside each Fourier block. This quantity can be read as the average strength with which the current kernel acts on that Fourier subspace: larger spectral strength means that residual components in that block are pushed down more strongly by the kernel dynamics.

If the evolving kernel helps because it preserves Fourier geometry better, then one would expect improved alignment in the subspaces over time. If instead it helps by redistributing operator strength, then one should expect improved decay of the low-frequency residual modes together with increased low-frequency spectral strength, even if the Fourier alignment worsens.

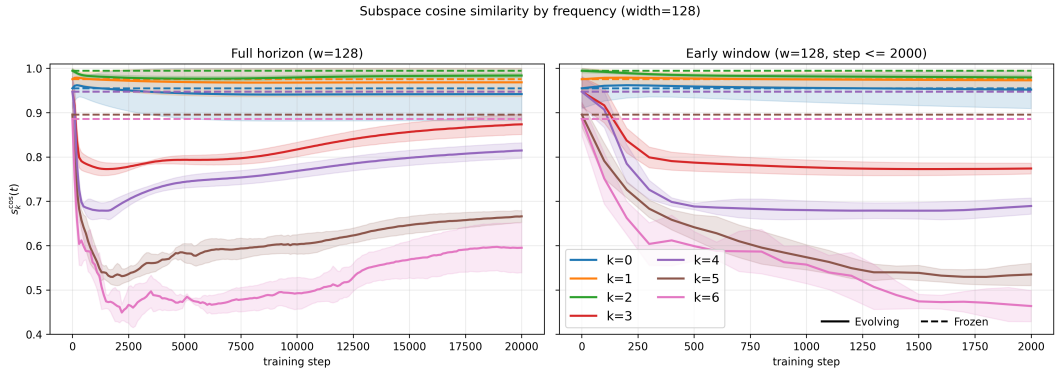
Figures 5.17–5.20 show three recurring trends. First, the mode-wise residual plots show that the early training dynamics still largely follows the frequency ordering: lower-frequency components decay before higher-frequency components. Kernel evolution does not remove this ordering, but changes the relative decay rates within the retained target modes. At widths 128 and 256, the evolving dynamics give smaller final residuals for several low and intermediate modes, including the constant mode, $\cos(\varphi)$, $\sin(2\varphi)$, $\cos(3\varphi)$, and $\cos(4\varphi)$. At width 512, this advantage is less uniform. At width 1024, the frozen dynamics are better on several components with the largest remaining residual amplitudes, including $\cos(3\varphi)$, $\cos(4\varphi)$, $\sin(5\varphi)$, and $\sin(6\varphi)$. Since the least-squares objective is quadratic in the residual, these larger remaining components have the greatest effect on the final loss.

Second, the subspace cosine similarity plots show that kernel evolution does not improve Fourier alignment. The low-frequency subspaces, especially $k = 0$, $k = 1$, and to a lesser extent $k = 2$, remain close to their Fourier subspaces during training. In contrast, several higher-frequency subspaces lose alignment early in training and only partially recover later. The frozen alignment improves with width, as expected from the frozen finite-width results. The evolving alignment, however, reaches a similar range across widths for some higher frequencies. Thus, at larger width, the evolving kernel can move farther away from the better-aligned initialization, even though the low-frequency subspaces remain relatively stable.

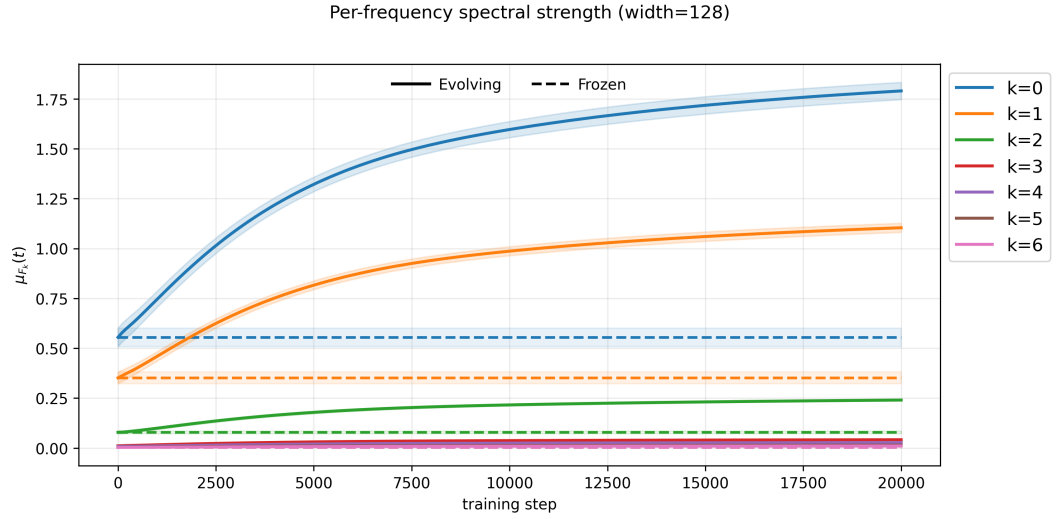
Third, the spectral-strength plots show that the largest change in operator strength occurs precisely in the low-frequency blocks that remain well aligned. At width 128, the average strength in the constant block and the first frequency increases by a large factor during training, with a smaller increase for $k = 2$. The same effect is present at width 256, weaker at width 512, and much smaller at width 1024. The higher-frequency blocks remain weak in comparison and do not show the same increase in strength.



(a) Mode-wise residual decay.

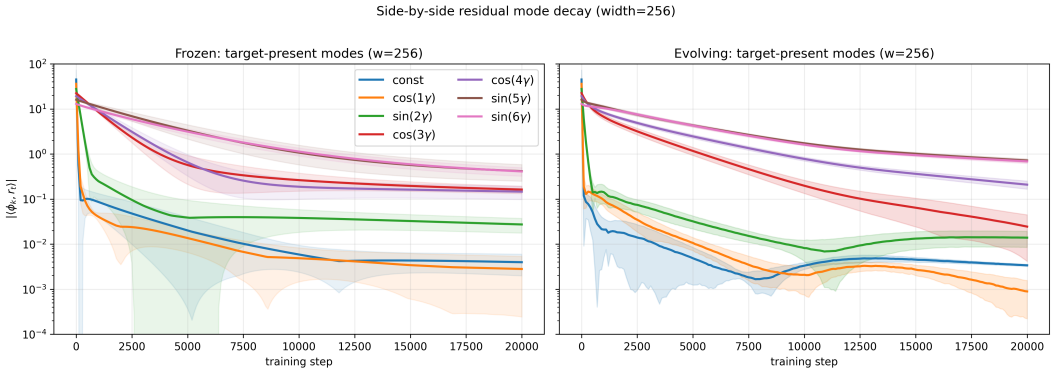


(b) Subspace cosine similarity.

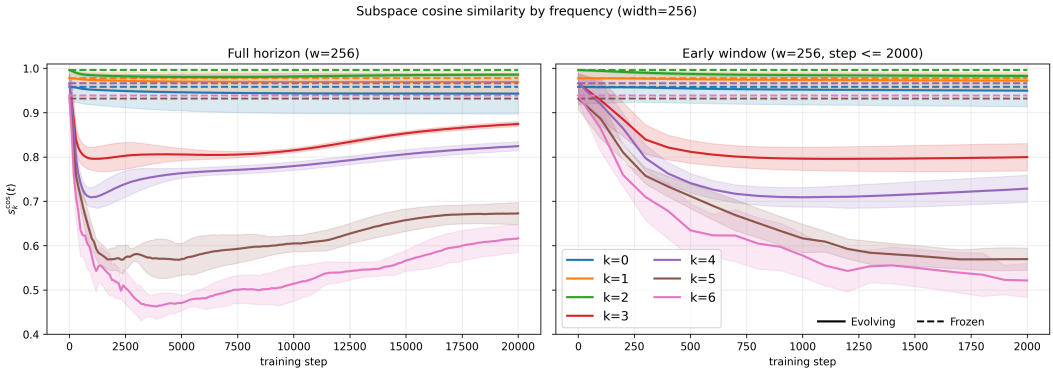


(c) Per-frequency spectral strength.

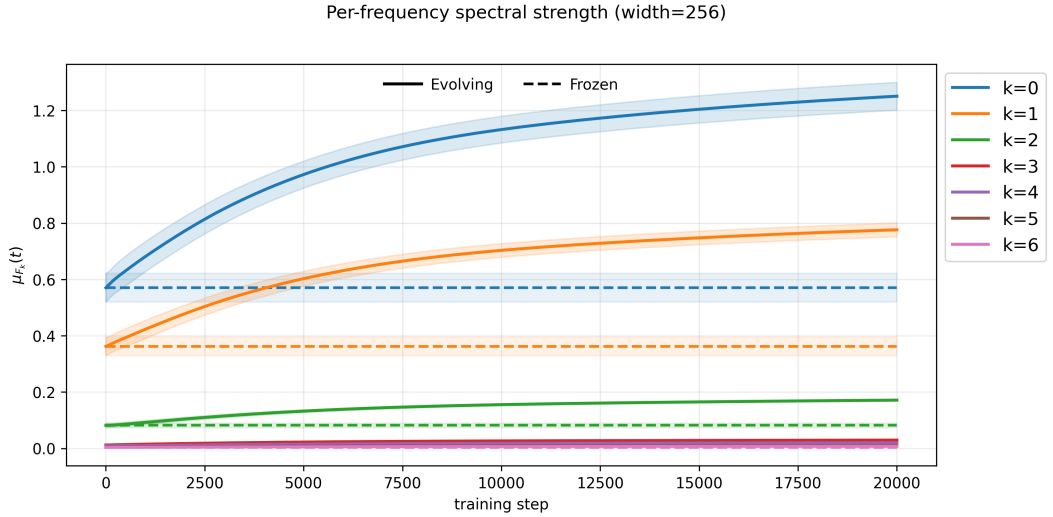
Figure 5.17: Joint diagnostics for the evolving finite-width tangent kernel at width 128. The top panel compares frozen and evolving residual decay for the target-present modes. The middle panel measures Fourier-subspace alignment. The bottom panel measures the average tangent-kernel strength inside each Fourier block.



(a) Mode-wise residual decay.

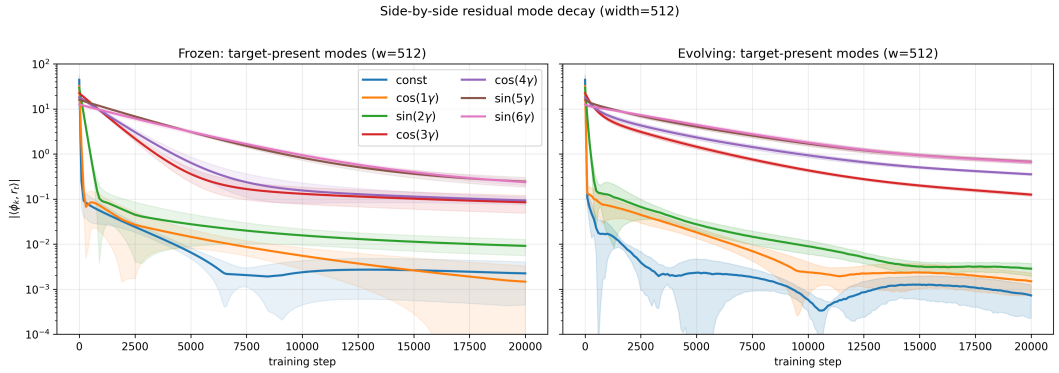


(b) Subspace cosine similarity.

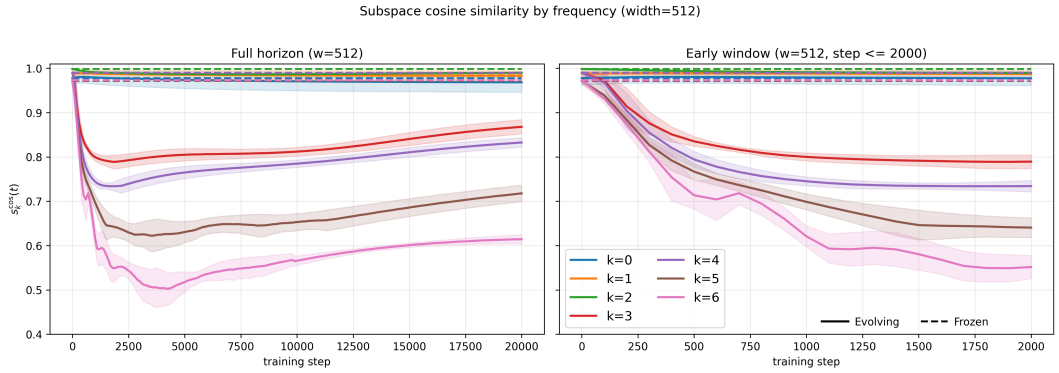


(c) Per-frequency spectral strength.

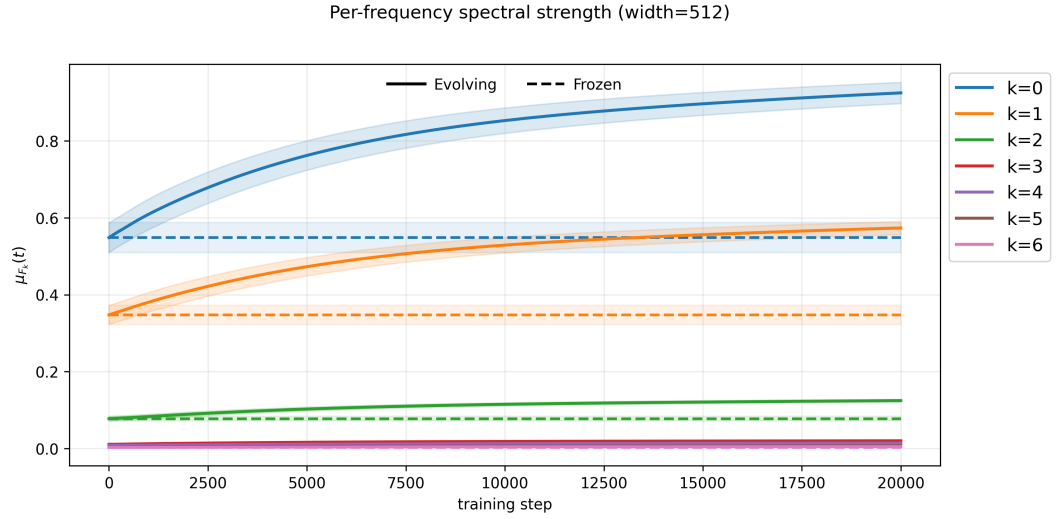
Figure 5.18: Joint diagnostics for the evolving finite-width tangent kernel at width 256.



(a) Mode-wise residual decay.

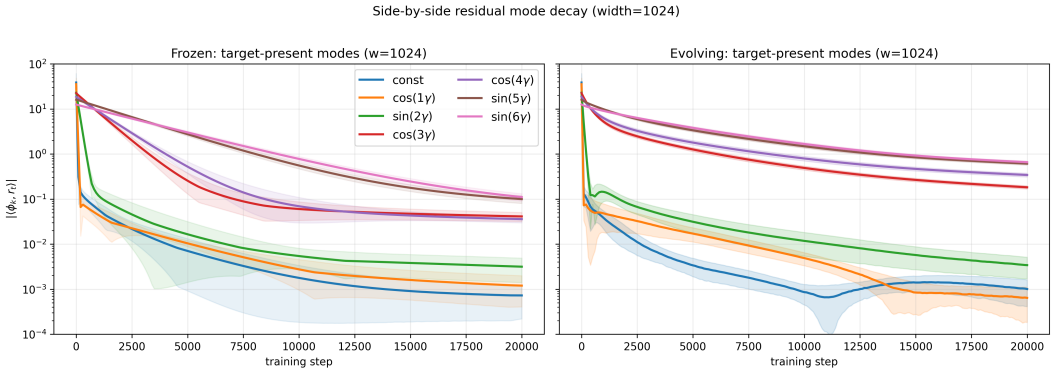


(b) Subspace cosine similarity.

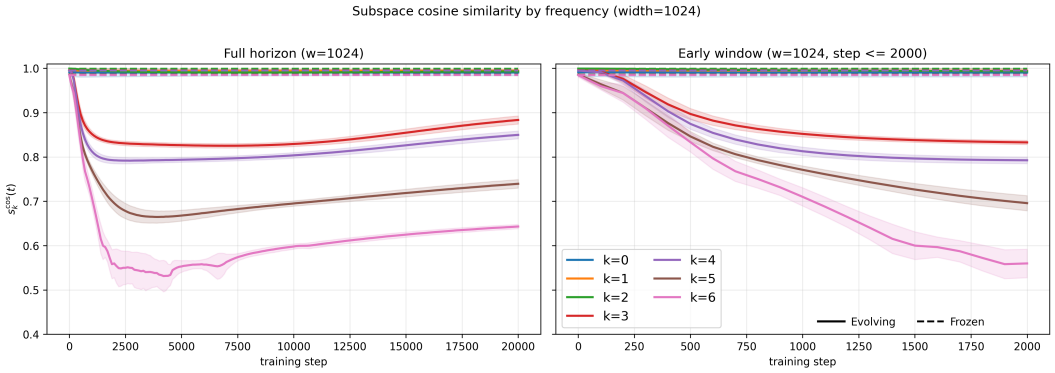


(c) Per-frequency spectral strength.

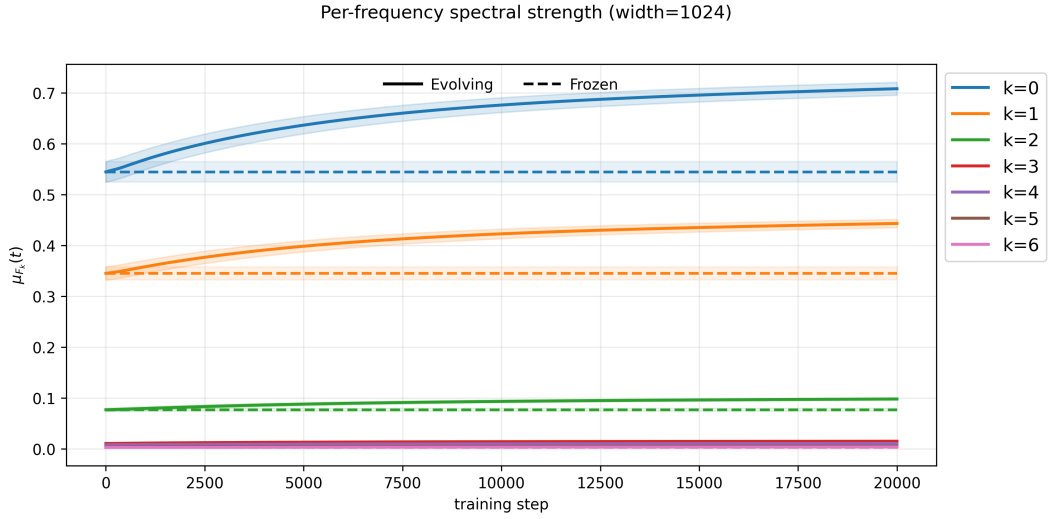
Figure 5.19: Joint diagnostics for the evolving finite-width tangent kernel at width 512.



(a) Mode-wise residual decay.



(b) Subspace cosine similarity.



(c) Per-frequency spectral strength.

Figure 5.20: Joint diagnostics for the evolving finite-width tangent kernel at width 1024.

Taken together, these plots suggest that the small-width benefit of kernel evolution is not caused by better preservation of Fourier geometry. The alignment either stays similar for the lowest frequencies or worsens for higher frequencies. The more visible effect is that kernel evolution increases the operator strength in the low-frequency blocks that are already well aligned with the Fourier subspaces. At small width, this extra low-frequency strength can outweigh the loss of alignment elsewhere. At large width, the frozen kernel already has good Fourier structure, and the additional strength gained from kernel evolution is smaller, so the evolving dynamics no longer gives a clear advantage.

5.5.3 Width dependence of tangent-kernel drift

The previous plots compare the consequences of frozen and evolving dynamics. We now measure the kernel movement directly. Recall from Section 4.6 that

$$C_t^{(K)} = P_K T_t^{(m)} P_K$$

denotes the low-frequency block of the finite-width tangent operator. We measure its drift from initialization by

$$\|C_t^{(K)} - C_0^{(K)}\|_F.$$

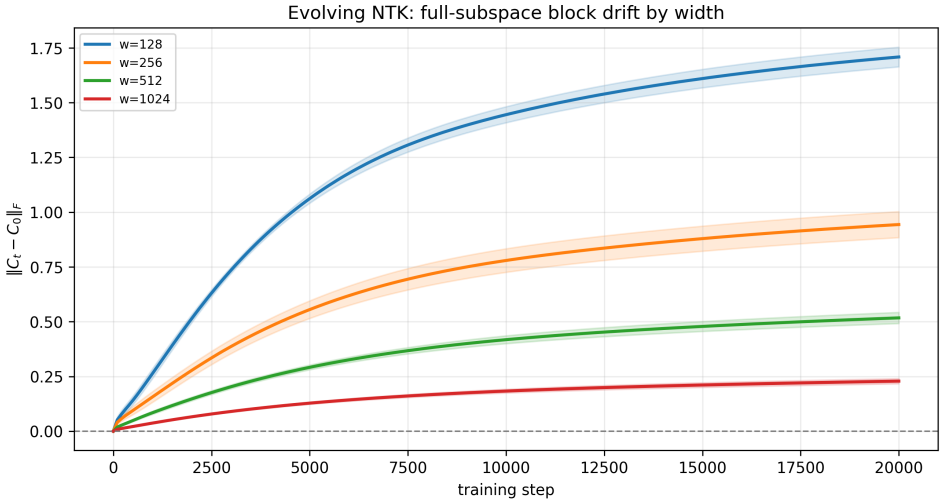


Figure 5.21: Drift of the projected evolving tangent-kernel block across widths, measured by $\|C_t^{(K)} - C_0^{(K)}\|_F$. The low-frequency block changes substantially at small width and much less at larger width.

Figure 5.21 shows the drift of the projected low-frequency tangent block during training. The drift is largest at widths 128 and 256, and becomes much smaller at widths 512 and 1024. Thus, in these experiments, the finite-width tangent kernel moves less as the width increases.

This is consistent with the NTK regime, where increasing width makes the training dynamics closer to the linearized dynamics around initialization and the tangent kernel

changes less during training (Jacot et al., 2020; Lee et al., 2020; Chizat et al., 2019). In the present experiments, this also explains why the difference between frozen and evolving dynamics is most visible at small width.

5.5.4 Effect of target amplitudes on spectral strength

The spectral-strength plots above suggest that the evolving tangent kernel increases its strength mainly on the low-frequency blocks. One possible concern is that this effect could be caused by the particular amplitudes in the target function. In the original target, the lower-frequency components have larger amplitudes than the higher-frequency components. To check whether the observed spectral reweighting is tied to this choice, we repeat the evolving-kernel experiment with two modified targets.

The three targets are

$$f_{\text{orig}}(\gamma) = 1.0 + 0.9\cos(\gamma) - 0.8\sin(2\gamma) + 0.7\cos(3\gamma) + 0.6\cos(4\gamma) - 0.5\sin(5\gamma) - 0.4\sin(6\gamma), \tag{5.23}$$

$$f_{\text{eq}}(\gamma) = 1.0 + \cos(\gamma) - \sin(2\gamma) + \cos(3\gamma) + \cos(4\gamma) - \sin(5\gamma) - \sin(6\gamma), \tag{5.24}$$

$$f_{\text{rev}}(\gamma) = 0.4 + 0.5\cos(\gamma) - 0.6\sin(2\gamma) + 0.7\cos(3\gamma) + 0.8\cos(4\gamma) - 0.9\sin(5\gamma) - \sin(6\gamma). \tag{5.25}$$

The equal-amplitude target gives all active non-constant modes the same amplitude. The reversed-amplitude target gives the largest amplitudes to the higher-frequency modes.

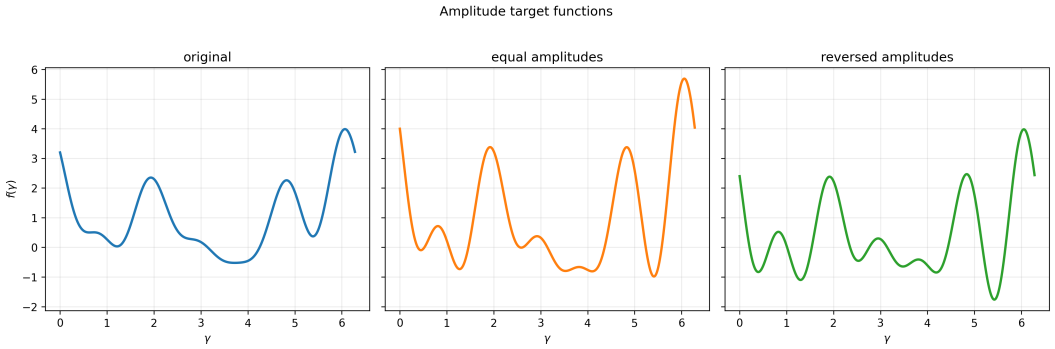
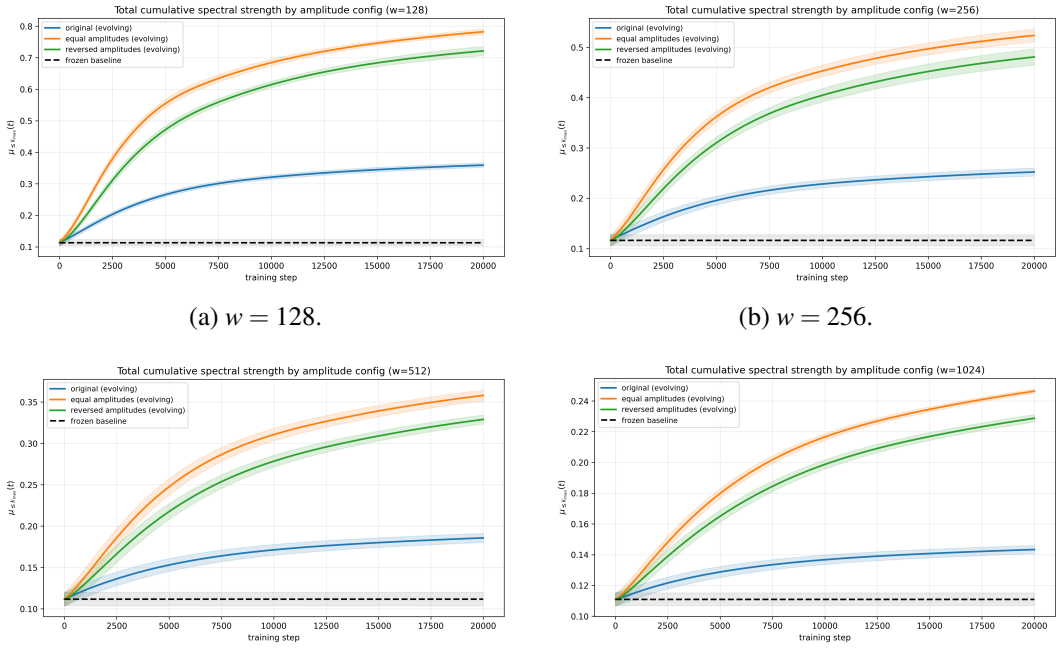


Figure 5.22: Target functions used in the amplitude-variant experiments. The original target has larger amplitudes on lower frequencies, the equal-amplitude target assigns the same amplitude to all active non-constant modes, and the reversed-amplitude target gives larger amplitudes to higher frequencies.

Figure 5.23 shows that the increase in cumulative spectral strength is not specific to the original target amplitudes. For all widths, the evolving tangent kernel increases the average strength on the retained low-frequency block relative to the frozen baseline. This increase is largest for the equal-amplitude target, smaller but still clear for the reversed-amplitude target, and weakest for the original target.

The effect also decreases with width. At width 128, all evolving curves move well above the frozen baseline. At width 1024, the same ordering is still visible, but the increase is smaller. Thus, the low-frequency strength increase persists across all three amplitude choices. It is not only a consequence of the original target assigning larger amplitudes to



(a) $w = 128$.

(b) $w = 256$.

(c) $w = 512$.

(d) $w = 1024$.

Figure 5.23: Cumulative low-frequency spectral strength under different target amplitude choices. The dashed line shows the frozen baseline, while the solid curves show the evolving tangent kernel for the original, equal-amplitude, and reversed-amplitude targets.

lower frequencies: the increase is also present when all non-constant modes have equal amplitudes, and when the larger amplitudes are assigned to higher frequencies.

The ordering across targets is also worth noting. In these experiments, the equal-amplitude target produces the largest increase, the reversed-amplitude target produces a smaller but still clear increase, and the original target produces the weakest increase. In particular, the reversed-amplitude target still produces a larger low-frequency strength increase than the original target, despite putting larger amplitudes on higher-frequency modes.

5.6 Summary of results

This chapter tested how the idealized infinite-width, infinite-data spectral bias picture changes under realistic constraints in a controlled fashion. Throughout these analyses, the primary comparisons are made on the retained low-frequency subspace $\mathcal{H}_K$, which contains the constant mode and the first K non-zero Fourier frequencies.

The core findings are:

- **Ideal continuum setting:** Under the limiting NTK, each Fourier component evolves independently. Lower frequencies possess larger eigenvalues and decay faster. Crucially, the architecture dictates the spectrum: odd modes $k \geq 3$ only receive non-zero eigenvalues if the network includes a trainable hidden bias.

- **Finite sampling:** Replacing the continuous measure with n discrete samples destroys exact Fourier diagonalization. However, on $\mathcal{H}_K$, the sampled operator deviates from the ideal Fourier prediction by an error of just $O(n^{-1/2})$ (with fixed confidence, up to logarithmic factors). The corrected finite-sample eigenvalues (see Proposition 5.4) accurately predict the dominant early-time residual decay.
- **Frozen finite width:** When fixing a finite-width (m) tangent kernel at initialization, it acts as a random perturbation of the continuum operator. On $\mathcal{H}_K$, the projected finite-width operator concentrates around the infinite-width baseline at a rate of $O(m^{-1/2})$.
- **Evolving finite width:** Allowing the kernel to train empirically breaks the fixed-kernel assumption. At small widths, training alters the operator to actively increase low-frequency spectral strength (particularly for $k = 0$ and $k = 1$). This advantage diminishes at larger widths, where the frozen initialization is already tightly concentrated around the ideal baseline.

Chapter 6

Discussion and Conclusion

On the circle S^1, functions can be decomposed into sine and cosine waves, i.e., the Fourier basis. This makes the circle a useful setting for studying spectral bias. Low frequencies describe slowly varying structure, while high frequencies describe more oscillatory structure. The question is whether the training dynamics treat these components differently, and whether this distinction remains meaningful after moving away from the ideal infinite-width and infinite-data setting.

In the continuum infinite-width limit, the NTK is exactly rotation-invariant, meaning its associated operator is perfectly diagonalized by the Fourier basis. Consequently, each Fourier component of the residual evolves independently:

$$\alpha_p(t) = \exp(-\lambda_p t)\alpha_p(0), \tag{6.1}$$

where $\alpha_p(t)$ is the residual coefficient of the p-th Fourier basis function, and λ_p is the corresponding kernel eigenvalue. For the ReLU NTK studied here, the low-frequency eigenvalues are larger, dictating that low-frequency residual components decay strictly faster. This constitutes the baseline spectral-bias mechanism (Jacot et al., 2020; Lee et al., 2020; Rahaman et al., 2019; Ronen et al., 2019; Bietti and Mairal, 2019; Cao et al., 2020).

The continuum calculation also shows the importance of including a trainable bias in the hidden layer of the network. With the hidden-bias contribution, odd frequencies $k \geq 3$ have nonzero eigenvalues. Without this contribution, those frequencies are null directions of the fixed-kernel dynamics. Thus, the architecture itself affects whether some frequencies are learned at all by the fixed-kernel operator.

Restricting to $\mathcal{H}_K$ lets us ask how much of the idealized Fourier description holds once the continuum measure and infinite width are replaced by finite ones. The frozen and sampled cases turn out to be controlled perturbations of the continuum picture; the more interesting departure is the evolving kernel at small width, which does not merely approximate the fixed-kernel dynamics but reshapes them in a way that strengthens the low-frequency bias.

6.1 Answers to the research questions

The overarching inquiry of this thesis asked to what extent the Fourier decomposition of the infinite-width NTK on S^1 persists under finite-sample and finite-width perturbations. We find that the idealized description survives as a controlled approximation for frozen and sampled networks on low-frequency blocks, with an error that decreases in the sample size and width. When the kernel is allowed to evolve, this description becomes incomplete: in our experiments, training improves the loss mainly by reweighting spectral strength toward low frequencies rather than by bringing the eigenspaces into closer alignment with the Fourier modes, and the alignment can even degrade. The sub-questions are answered as follows.

Finite-sample perturbation (RQ1). Replacing the uniform continuum measure on S^1 with n i.i.d. samples destroys exact Fourier diagonalization. On the low-frequency subspace $\mathcal{H}_K$, however, the Fourier structure persists approximately: the sampled operator deviates from the diagonal Fourier prediction by an error of order $O(n^{-1/2})$ on this block, up to logarithmic factors. The frequency components therefore no longer evolve independently, but the corrected finite-sample eigenvalues (Proposition 5.4) still govern the dominant early-time decay of the retained components. As the results chapter showed, deviations from this prediction can become visible at later times, once some components have already decayed.

Finite-width perturbation (RQ2). The effect of finite width depends on whether the tangent kernel is held fixed or allowed to evolve. With the kernel frozen at initialization—so that training reduces to kernel regression on the initial tangent features—finite width acts as a controlled random perturbation of the continuum operator. On $\mathcal{H}_K$, the frozen finite-width block concentrates around the infinite-width Fourier block at the $O(m^{-1/2})$ rate. When the network is trained fully, the kernel evolves during training. At small widths, this evolution reweights the operator toward larger eigenvalues on the low-frequency blocks, so the effective per-frequency learning rates change during training beyond what the initial spectrum predicts. This advantage diminishes with width, where the frozen kernel is already close to the continuum baseline.

6.2 The robustness of the fixed-kernel approximation

The infinite-width NTK is sometimes treated as a purely asymptotic object whose relevance to networks of finite width and finite data is unclear, and recent work has examined finite-width corrections and limitations of the kernel picture (Hanin and Nica, 2019; Bordelon and Pehlevan, 2023; Vyas et al., 2022). The frozen-kernel results in this thesis speak to a narrower question. On a fixed low-frequency block $\mathcal{H}_K$, and with the tangent kernel held fixed, both finite sampling and finite width perturb the continuum Fourier operator in a controlled way: the deviation concentrates around the continuum block and decreases at the $O(n^{-1/2})$ and $O(m^{-1/2})$ rates. Within this restricted setting, the low-frequency spectral

structure that underlies spectral bias is not an artifact of the infinite limit but survives as an approximation whose error we can bound.

The computed continuum eigenvalues decrease quickly with frequency. From Theorem 5.2, $\lambda_k = O(k^{-2})$, and numerically the first few are $\lambda_0 \approx 0.55$, $\lambda_1 \approx 0.35$, $\lambda_2 \approx 0.079$, $\lambda_3 \approx 0.011$, with the higher frequencies smaller still. This sets the scale against which the finite-sample and finite-width perturbations must be read: the bounds control an absolute error on $\mathcal{H}_K$, so a given perturbation leaves a mode's dynamics intact only if it is small compared to that mode's eigenvalue. The large low-frequency eigenvalues should therefore tolerate more sampling or width noise than the small high-frequency ones. For the lowest frequencies we can be reasonably confident that the continuum prediction still describes the dynamics at modest sample size or width. At high frequency the eigenvalues are small and close in size, and the alignment is poor enough that we cannot tell whether an empirical mode is still following its own eigenvalue or picking up contributions from neighbouring frequencies. We have not tested this magnitude argument as such, but it is consistent with the frozen alignment experiments, where low frequencies are well aligned at small width and higher frequencies need larger widths, and with the broader finding that high-frequency components are learned later and are more sensitive to model and data distributions (Rahaman et al., 2019; Ronen et al., 2019; Basri et al., 2020).

6.3 Kernel evolution and the limits of fixed dynamics

Once the kernel evolves, the fixed-kernel description no longer accounts for the full dynamics, and our networks sit between the lazy regime, where the tangent kernel barely moves, and feature learning, where it moves substantially (Chizat et al., 2019; Mei et al., 2018; Chizat and Bach, 2018). The evolving kernel helps where it moves: it improves on the frozen kernel in exactly the small-width range where the low-frequency block drifts most (Figure 5.21).

The improvement does not come from better Fourier alignment, which often degrades at higher frequencies, but from added spectral strength in the lowest blocks, and the amplitude variants show this is not just tracking where the target places its mass. Read alongside the fixed-kernel picture, this looks like a second-order form of the same bias: gradient descent not only learns low frequencies faster under a fixed spectrum, it also moves the kernel to strengthen those frequencies further. We have shown this as a correlation with lower loss, not as its established cause.

Why the added strength favours $k = 0$ and $k = 1$ is where the architecture and choice of input parameterisation suggests an explanation. With inputs $x(\varphi) = (\cos\varphi, \sin\varphi)$, the pre-activation of hidden neuron j is

$$w_j^\top x(\varphi) + b_j = w_{j_1}\cos\varphi + w_{j_2}\sin\varphi + b_j = \|w_j\|\cos(\varphi - \psi_j) + b_j. \tag{6.2}$$

Frequencies $k = 0$ and $k = 1$ are therefore present in the parameterisation itself, through the bias and the input embedding, while the higher harmonics (including the relatively well-aligned $k = 2$) appear only indirectly, once the ReLU acts on this frequency-one pre-activation. The network may then strengthen and keep aligned exactly the directions it represents most directly, since these need the least movement to amplify.

This is testable with the present machinery. A target with no constant or frequency-one component would separate the two readings: if the strength increase still concentrates on $k = 0$ and $k = 1$, the parameterisation account is favoured; if it follows the lowest frequency present in the target, then gradient descent is instead amplifying wherever the residual is largest. These observations come from a single architecture on S^1 and do not separate cause from correlation. What they show is that at small width, training reshapes the kernel in a low-frequency-favouring direction that the fixed-kernel theory does not capture, with the parameterisation a likely source. A formal account is left to future work.

6.4 Beyond the uniform circle setting

On the uniform circle two things line up: the limiting NTK we derived depends only on the angle between two points, and the uniform distribution treats every angle the same. When both hold, the kernel acts as a convolution, and the Fourier modes are its eigenfunctions, each scaled by its own eigenvalue. Both pieces matter, the kernel's rotation-invariance and the uniformity of the distribution.

A non-uniform input density removes the second piece. As shown in Section 3.2.3, the operator is then no longer a convolution, and the Fourier modes need not be its eigenfunctions. The components learned first are the leading eigenfunctions of this density-weighted operator rather than the lowest Fourier frequencies, and Basri et al. (2020) show that learning also speeds up where inputs are more concentrated. Spectral bias persists as a preference for the leading eigenfunctions of the operator, but these need not be the familiar low frequencies (Basri et al., 2020; Bowman and Montufar, 2022).

A more realistic input distribution makes the change starker. Real inputs are not confined to a circle but spread through the input space, and the standard model of this is a Gaussian on $\mathbb{R}^d$. The kernel is then no longer a convolution, and its eigenfunctions are not Fourier modes but Hermite polynomials, the Gaussian counterpart of sines and cosines (Mei et al., 2022). These form a family of increasing degree, with degree playing the role frequency did: low-degree polynomials are simple and slowly varying, a constant, then a line, then a parabola, and the coarse, low-degree shape of the target is learned first (Cao et al., 2020). We expect the spectral-bias mechanism to be unchanged, since eigenvalue size sets the decay rates regardless of the basis. Our proofs, though, would not transfer directly. They assume a bounded kernel and use Hoeffding, Bernstein, and McDiarmid inequalities, all of which rely on bounded or controlled-tail summands, available to us because the circle is compact. A Gaussian measure has unbounded support, so neither the boundedness nor these inequalities apply directly, and the bounds would have to be re-derived with tools suited to unbounded inputs.

We do not expect a smooth activation such as sigmoid or tanh to take us outside the framework used here. The Fourier basis is the eigenbasis because the operator is a convolution, and the derivation in Appendix B shows this follows from the Gaussian distribution of the parameters rather than from the activation being ReLU. For any pointwise activation the NTK still depends only on the angle between two inputs, so the Fourier modes remain its eigenfunctions and the same concentration inequalities and eigenspace diagnostics

should apply. What changes is the eigenvalues. ReLU is continuous but not differentiable at the origin, and this non-smoothness is what limits the eigenvalues to a polynomial decay rate; our derivation gives $\lambda_k = O(k^{-2})$ in our setting. Bietti and Mairal (2019) trace this polynomial decay to the non-smooth part of the ReLU kernel, and Murray et al. (2023) show more generally that smoother activations have faster-decaying NTK spectra. An activation without this non-differentiability, such as sigmoid or tanh, would therefore give faster-decaying eigenvalues and sharpen the bias toward low frequencies. The fixed-kernel analysis therefore carries over with only the spectrum recomputed. We are less sure about the evolving-kernel results. These depend on how the kernel moves during training, which we did not analyse for other activations and would not expect to behave the same way, so we leave that case open.

6.5 Limitations and future directions

The concentration results are fixed-block results. They control the prescribed low-frequency space $\mathcal{H}_K$ for fixed K, so they do not give a uniform statement over all Fourier modes, nor are they designed to stay sharp when K grows with samples (n) or width (m). This matters because higher frequencies have smaller eigenvalues that sit closer together, where the same perturbation has more effect. The proofs also use entrywise concentration and union bounds, which capture the correct decreasing trends but discard matrix structure and variance information, so the bounds are conservative in absolute magnitude. Sharpening them with operator-norm bounds (e.g., Tropp, 2015), variance-sensitive concentration, or tools for U- and V-statistics (Serfling, 1980) would bring them closer to the observed errors.

The finite-sample theory explains the leading behaviour of the sampled fixed-kernel dynamics but does not give a sharp time-uniform description of every component. Once a component has decayed close to zero, small operator errors can dominate what is plotted, so the late-time deviations we observe are expected but not quantified per mode.

The largest gap is the evolving kernel. The finite-width theorem controls the frozen tangent kernel at initialization but not the kernel as it changes during training, and our results for that regime are entirely empirical. The fixed-kernel limit is also tied to NTK scaling: in feature-learning or mean-field regimes the representation moves substantially during training, and a fixed initial spectrum cannot describe the dynamics (Mei et al., 2018; Chizat and Bach, 2018). Controlling how the low-frequency block drifts from initialization, and what that drift does to the residual dynamics, is the most direct next step.

The alignment diagnostics measure how close the empirical eigenspaces are to the Fourier subspaces, but alignment is not the same as loss: a well-aligned block can still be learned slowly if its eigenvalue is small, so these diagnostics describe the geometry of the operator rather than the convergence directly. A full description of practical training would need finite sampling, finite width, and training time treated together, which this work does not attempt. Our results cover the first two with the kernel frozen; the training-time effect is only studied empirically, and it is not simply additional noise, since at small width the kernel drift changes the dynamics in a structured, low-frequency-favouring way. Combining all three is the broadest open problem left.

The analysis also points to a concrete intervention. The fixed-kernel dynamics are a linear system whose per-mode convergence rate is set by the kernel eigenvalues, and preconditioning is the natural way to flatten that spectrum: rescaling each Fourier mode by its inverse eigenvalue clusters the effective rates and removes the frequency-dependent slowdown, as the theory preconditioner does in Section 5.3. This is available to us because the operator is known in closed form and diagonalised by a fixed basis, so the preconditioner can be applied directly to the residual in function space.

A finite-width trained network does not meet either condition. The operator that sets the rates is the tangent kernel, which has no closed form at finite width and drifts during training, and gradient descent acts on the parameters rather than on the function values, so the correction would have to be applied to the parameter-space gradient through this same evolving, implicitly known kernel. Preconditioning the rates in that setting, with an operator that is neither fixed nor explicitly available, is the real difficulty and is left to future work.

6.6 Reflection

This thesis sits within the broader line of work that tries to replace empirical intuition about deep learning with analysable mathematical models. The Neural Tangent Kernel is one of the most useful instruments in that line: it turns a non-convex training problem into a linear one whose behaviour can be characterised exactly. The value of the present work is as a controlled case study of where that reduction holds and where it stops: the fixed kernel is a useful baseline, but not a full account of training once the network learns features.

Implications, applicability, and stakeholders. Spectral bias has practical consequences wherever neural networks represent functions with fine detail, through the same NTK spectrum studied here. Wang et al. (2022) trace the failure of physics-informed neural networks to fit high-frequency parts of a PDE solution to a gap in the convergence rates of the different loss terms, and correct it by rescaling the dynamics with the kernel eigenvalues. This is the same eigenvalue-based correction used by the preconditioner in Section 5.3. In computer vision and graphics, Tancik et al. (2020) and Sitzmann et al. (2020) show that coordinate-based networks (networks that map a raw input coordinate, such as a pixel location or a 3D point, directly to the signal value) blur fine detail for the same reason, and that Fourier-feature mappings and periodic activations fix this by reshaping the effective kernel spectrum so high frequencies are no longer suppressed. The mechanism this thesis analyses on the circle is the one these methods modify in practice.

These methods are built and used by scientific-machine-learning practitioners working on physical simulation, by people building implicit neural representations for images, shapes, and audio, and by researchers designing ways to counteract spectral bias. Knowing why and when a network learns some components before others makes its behavior more predictable, which matters when these models inform downstream decisions. This work does not bear directly on the safety, fairness, or societal deployment of machine-learning systems, but we hope it adds to the understanding of one specific mechanism. However, a result on the uniform circle is still far from a guarantee about a deployed model.

6.7 Conclusion

This thesis studied spectral bias in a setting where the ideal dynamics can be written down explicitly. On the uniform circle the infinite-width NTK is diagonal in the Fourier basis, each frequency has its own learning rate, and low frequencies are learned faster because their eigenvalues are larger.

This picture survives finite sampling and frozen finite width on the fixed low-frequency block: the ideal operator is perturbed, but the perturbation decreases with sample size and width at the $O(n^{-1/2})$ and $O(m^{-1/2})$ rates. The evolving finite-width experiments mark where the fixed-kernel account stops. Once the kernel moves during training, the Fourier basis remains useful for measurement, but the dynamics are also shaped by how the operator itself changes, and at small width that change reinforces the low-frequency bias rather than following it passively.

The Fourier basis is therefore exact in the continuum model, approximately valid on the low-frequency block under finite sampling and frozen finite width, and a measurement tool for the evolving kernel. Turning that last regime into theory is the main problem left open by this work.

Chapter 7

Use of Generative AI

In accordance with the TU Delft guidelines on the use of generative AI in end projects, I disclose here the tools used in preparing this thesis, their purpose, and the scope of their application. All ideas, research directions, calculations, and theorems in this work are my own: the derivations were carried out manually and double-checked with professors who are experts in the relevant fields, and generative AI was used only as a supporting aid that I reviewed and verified at every stage. Concretely, ChatGPT 5.3 Thinking was used for brainstorming and for organizing my thoughts into actionable plans, and ChatGPT Codex 5.3 assisted with coding tasks such as setting up and running experiments and debugging. Generative AI also assisted in writing the scripts used to produce the plots in this thesis; these scripts were subsequently read and adjusted manually to ensure correctness. For the written text, large language model tools were used to improve grammar, spelling, and clarity, to smooth transitions between sections, and to make the document more coherent. My handwritten derivations were converted into LaTeX with the help of tools such as Mathpix and LLM-based assistants (ChatGPT, Gemini), which were also used to improve the consistency of notation throughout; the underlying mathematics is my own. Finally, NotebookLM was used as a retrieval system to quickly locate specific claims within the literature for citation; I read all cited papers myself, and this tool served purely for convenient lookup. In all cases I retain full creative and intellectual responsibility for the content, results, and academic integrity of this thesis, and I have verified the accuracy, validity, and originality of all material included. The cover illustration was generated using OpenAI's ChatGPT image-generation tool from prompts written and refined by the author and later manually edited. The image is used as a conceptual visualisation of neural-network training dynamics and Fourier/spectral structure.

Bibliography

Allen-Zhu, Z., Li, Y., and Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pages 242–252. PMLR.

Araújo, D., Oliveira, R. I., and Yukimura, D. (2019). A mean-field limit for certain deep neural networks.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks.

Bartolucci, F., Vito, E. D., Rosasco, L., and Vigogna, S. (2021). Understanding neural networks with reproducing kernel banach spaces.

Basri, R., Galun, M., Geifman, A., Jacobs, D., Kasten, Y., and Kritchman, S. (2020). Frequency bias in neural networks for input of non-uniform density. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 685–694. PMLR.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Bietti, A. and Mairal, J. (2019). On the inductive bias of neural tangent kernels.

Björck, A. and Golub, G. H. (1973). Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594.

Bordelon, B. and Pehlevan, C. (2023). Dynamics of finite width kernel and prediction fluctuations in mean field neural networks.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311.

Bowman, B. and Montufar, G. (2022). Spectral bias outside the training set for deep networks in the kernel regime.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Cao, Y., Fang, Z., Wu, Y., Zhou, D.-X., and Gu, Q. (2020). Towards understanding the spectral bias of deep learning.

Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport.

Chizat, L., Colombo, M., Fernández-Real, X., and Figalli, A. (2024). Infinite-width limit of deep linear neural networks. *Communications on Pure and Applied Mathematics*, 77(10):3958–4007.

Chizat, L., Oyallon, E., and Bach, F. (2019). On lazy training in differentiable programming. *Advances in neural information processing systems*, 32.

Cho, Y. and Saul, L. (2009). Kernel methods for deep learning. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2015). The loss surfaces of multilayer networks.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR.

Fefferman, C., Mitter, S., and Narayanan, H. (2016). Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049.

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

Geifman, A., Barzilai, D., Basri, R., and Galun, M. (2024). Controlling the inductive bias of wide neural networks by modifying the kernel's spectrum. *Transactions on Machine Learning Research*.

Golikov, E., Pokonechnyy, E., and Korviakov, V. (2022). Neural tangent kernel: A survey.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix Computations*. Johns Hopkins University Press, 4 edition.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. (2018). Characterizing implicit bias in terms of optimization geometry. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841. PMLR.

Hanin, B. and Nica, M. (2019). Finite depth and width corrections to the neural tangent kernel.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jacot, A., Gabriel, F., and Hongler, C. (2020). Neural tangent kernel: Convergence and generalization in neural networks.

Jiang, S., Cyr, E. C., Southworth, B. S., and Voronin, A. (2026). On the convergence behavior of preconditioned gradient descent toward the rich learning regime. In *International Conference on Learning Representations*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2020). Wide neural networks of any depth evolve as linear models under gradient descent $*$. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124002.

Mei, S., Misiakiewicz, T., and Montanari, A. (2022). Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84.

Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Murray, M., Jin, H., Bowman, B., and Montufar, G. (2023). Characterizing the spectrum of the ntk via a power series expansion.

Nguyen, P.-M. (2019). Mean field limit of the learning dynamics of multilayer neural networks.

Nguyen, P.-M. and Pham, H. T. (2023). A rigorous framework for the mean field limit of multilayer neural networks.

Penrose, R. (1955). A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51:406–413.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F. A., Bengio, Y., and Courville, A. (2019). On the spectral bias of neural networks.

Ronen, B., Jacobs, D., Kasten, Y., and Kritchman, S. (2019). The convergence rate of neural networks for learned functions of different frequencies. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Rotskoff, G. and Vanden-Eijnden, E. (2022). Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. Wiley.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Sirignano, J. and Spiliopoulos, K. (2019). Mean field analysis of neural networks: A law of large numbers.

Sitzmann, V., Martel, J. N. P., Bergman, A. W., Lindell, D. B., and Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33.

Tancik, M., Srinivasan, P. P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J. T., and Ng, R. (2020). Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33.

Trefethen, L. N. and Bau, D. (2022). *Numerical linear algebra*. SIAM.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1–2):1–230.

Tu, Z., Aranguri, S., and Jacot, A. (2024). Mixed dynamics in linear networks: Unifying the lazy and active regimes.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999.

Varadarajan, V. S. (1958). On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

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

Vershynin, R. (2018b). *Concentration Without Independence*, page 98–126. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vyas, N., Bansal, Y., and Nakkiran, P. (2022). Limitations of the ntk for understanding generalization in deep learning.

Wang, S., Yu, X., and Perdikaris, P. (2022). When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768.

Xu, Z.-Q. J., Zhang, Y., and Xiao, Y. (2019). Training behavior of deep neural network in frequency domain. In *International Conference on Neural Information Processing*, pages 264–274. Springer.

Yang, Y. (2024). 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*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization.

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

Appendix A

Mean-Field Viewpoint

The main text uses the NTK viewpoint as its main analytical framework because it gives a fixed operator on functions. On S^1, this operator can be diagonalized in the Fourier basis, which makes it possible to discuss frequency-dependent learning in terms of eigenvalues and eigenspaces.

The mean-field viewpoint gives another way to take a large-width limit. It is not used in the main finite-sample or finite-width analysis, but it is useful context because it describes a different aspect of wide-network training. In the mean-field scaling, one does not primarily track a fixed tangent kernel at initialization. Instead, one tracks how the hidden neurons are distributed in parameter space, and how this distribution changes during training.

This appendix records the corresponding derivation for a two-layer network. The purpose is to keep the comparison available without interrupting the main line of the thesis.

A.1 Overview and relation to the NTK viewpoint

Consider a two-layer network with the mean-field scaling

$$f_\theta(x) = \frac{1}{m}\sum_{\alpha=1}^{m} v_\alpha \varphi(w_\alpha^\top x). \tag{A.1}$$

Here each neuron has parameters

$$\theta_\alpha = (v_\alpha, w_\alpha) \in \Omega := \mathbb{R}\times\mathbb{R}^d.$$

The key observation is that the network output depends on the collection of neurons only through their empirical distribution,

$$\mu_m := \frac{1}{m}\sum_{\alpha=1}^{m} \delta_{\theta_\alpha}.$$

Indeed,

$$f_\theta(x) = \int_\Omega v\,\varphi(w^\top x)\, d\mu_m(v,w).$$

Thus the state of the network can be described by a probability measure over neuron parameters.

In suitable large-width limits, the empirical measure μ_m converges to a deterministic time-dependent measure μ_t. Training is then described by the evolution of this measure. This viewpoint was developed for two-layer networks by Mei et al. (2018), and related interacting-particle descriptions were studied by Rotskoff and Vanden-Eijnden (2022). From an optimization perspective, Chizat and Bach (2018) studied the corresponding many-particle limit using tools from optimal transport. Extensions and rigorous frameworks for deeper networks have also been studied (Nguyen, 2019; Araújo et al., 2019; Sirignano and Spiliopoulos, 2019; Nguyen and Pham, 2023).

This viewpoint is useful because it can describe feature movement in the large-width limit. In contrast, the strict NTK limit freezes the tangent kernel: the features defined by the gradients at initialization remain fixed in the limit. For the spectral question studied in this thesis, however, the NTK viewpoint is more directly useful. The main object here is an operator on functions whose eigenvalues and eigenspaces can be compared with Fourier modes on S^1. The mean-field viewpoint instead describes the evolution of a probability measure over parameters, so it does not give a single fixed Fourier-diagonal operator in the same way.

A.2 Detailed derivation of the mean-field representation

A.2.1 Parameter space and empirical measure

Let $\Omega := \mathbb{R} \times \mathbb{R}^d$ denote the parameter space, and let $\mathcal{F}$ be its Borel σ-algebra. For each neuron α, define the Dirac measure δ_{θ_α} on $(\Omega, \mathcal{F})$, i.e. if $A \in \mathcal{F}$ then,

$$\delta_{\theta_\alpha}(A) = \begin{cases} 1, & \theta_\alpha \in A, \\ 0, & \text{otherwise.} \end{cases}$$

Define the empirical neuronal measure

$$\mu_m := \frac{1}{m} \sum_{\alpha=1}^{m} \delta_{\theta_\alpha}. \tag{A.2}$$

By construction, $\mu_m \geq 0$ and $\mu_m(\Omega) = 1$, hence $\mu_m \in \mathcal{P}(\Omega)$ where $\mathcal{P}(\Omega)$ is the space of Borel probability measures on Ω.

Integration against Dirac measures. Let χ_A denote the indicator function of a measurable set $A \subset \Omega$. Then, for any $\theta_\alpha \in \Omega$,

$$\int_\Omega \chi_A(\theta)\, d\delta_{\theta_\alpha}(\theta) = \delta_{\theta_\alpha}(A) = \chi_A(\theta_\alpha).$$

More generally, if $s(\theta) = \sum_{k=1}^{r} c_k \chi_{A_k}(\theta)$ is a simple function, then

$$\int_\Omega s(\theta)\, d\delta_{\theta_\alpha}(\theta) = \sum_{k=1}^{r} c_k \chi_{A_k}(\theta_\alpha) = s(\theta_\alpha).$$

Any nonnegative measurable function $g : \Omega \to \mathbb{R}$ can be approximated by a sequence of simple functions, $\{s_n\}_{n\geq 0}$. If $s_n \uparrow g$ and each $s_n \geq 0$ then by the Monotone Convergence Theorem,

$$\int_\Omega g\, d\delta_{\theta_\alpha} = \int_\Omega \lim_{n\to\infty} s_n\, d\delta_{\theta_\alpha} = \lim_{n\to\infty} \int_\Omega s_n\, d\delta_{\theta_\alpha} = \lim_{n\to\infty} s_n(\theta_\alpha) = g(\theta_\alpha). \tag{A.3}$$

Integral representation of the network. Consider now the integral of g with respect to the empirical measure μ_m:

$$\int_\Omega g\, d\mu_m = \int_\Omega g\, d\left(\frac{1}{m}\sum_{\alpha=1}^m \delta_{\theta_\alpha}\right) = \frac{1}{m}\sum_{\alpha=1}^m g(\theta_\alpha).$$

Define

$$g(\theta_\alpha) = g(v_\alpha, w_\alpha) := v_\alpha\, \varphi(w_\alpha^\top x).$$

Then

$$\int_\Omega g\, d\mu_m = \frac{1}{m}\sum_{\alpha=1}^m v_\alpha\, \varphi(w_\alpha^\top x) = f_\theta(x), \tag{A.4}$$

recovering the finite-width network.

A.2.2 Initialization and weak convergence

Assume that the parameters $\{\theta_\alpha(0)\}_{\alpha=1}^m$ are initialized i.i.d. according to a probability measure μ_0 on Ω. Define the empirical measure at initialization

$$\mu_{m,0} := \frac{1}{m}\sum_{\alpha=1}^m \delta_{\theta_\alpha(0)}.$$

Then $\mu_{m,0}$ converges weakly to μ_0 almost surely as $m \to \infty$. Indeed, for any bounded continuous test function $\psi : \Omega \to \mathbb{R}$,

$$\int_\Omega \psi\, d\mu_{m,0} = \frac{1}{m}\sum_{\alpha=1}^m \psi(\theta_\alpha(0)) \xrightarrow[m\to\infty]{\text{a.s.}} \mathbb{E}_{\theta\sim\mu_0}[\psi(\theta)] = \int_\Omega \psi\, d\mu_0,$$

where the convergence follows from the strong law of large numbers applied to the i.i.d. random variables $\psi(\theta_\alpha(0))$. Since this holds for all bounded continuous ψ, we conclude that $\mu_{m,0} \to \mu_0$ (Varadarajan, 1958). As a consequence, the network output at initialization converges pointwise to the deterministic limit

$$f_{\mu_0}(x) := \int_\Omega v\, \varphi(w^\top x)\, d\mu_0(v,w).$$

For instance, under i.i.d. Gaussian initialization of the parameters, i.e. $v_0 \sim \mathcal{N}(0, I_m)$ $W_0 \sim \mathcal{N}(0, I_m \otimes I_d)$, the limiting initial measure $\mu_0 = \mathcal{N}(0, I_{d+1})$ is a Gaussian measure on a $d+1$-dimensional space.

A.2.3 Training dynamics and evolution of the empirical measure

Let $X = (x_1, \dots, x_n) \in \mathbb{R}^{d \times n}$ and $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ denote the training set, and consider the squared loss

$$L(v, W) = \frac{1}{2} \sum_{i=1}^{n} \big(f_{v,W}(x_i) - y_i\big)^2 = \frac{1}{2} \|f_{v,W}(X) - y\|_2^2,$$

where $f_{v,W}(X) := (f_{v,W}(x_i))_{i=1}^n$.

We study continuous-time gradient flow for a finite-width network with m neurons and learning rate $\eta > 0$,

$$\dot{v}_{t,\alpha} = -\eta \nabla_{v_\alpha} L(v_t, W_t), \qquad \dot{w}_{t,\alpha} = -\eta \nabla_{w_\alpha} L(v_t, W_t), \qquad \alpha \in \{1, \dots, m\}.$$

A direct computation yields, for each α,

$$\dot{v}_{t,\alpha} = \frac{\eta}{m} \varphi(X^\top w_{t,\alpha})^\top \big(y - f_{v_t,W_t}(X)\big), \tag{A.5}$$

$$\dot{w}_{t,\alpha} = \frac{\eta}{m} X \Big(\varphi'(X^\top w_{t,\alpha}) \odot \big(y - f_{v_t,W_t}(X)\big)\Big) v_{t,\alpha}, \tag{A.6}$$

where $\odot$ denotes elementwise (Hadamard) multiplication.

As seen in (A.5)–(A.6), evolution of each neuron depends on the parameters only through the current empirical measure $\mu_{m,t}$ via $f_{v_t,W_t}(X)$. Thus, the particle system (A.5)–(A.6) induces an evolution of the empirical measure in parameter space (Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2022).

Hence, the particle system (A.5)–(A.6) can be viewed as an interacting particle system driven by a measure-dependent velocity field. For any $\theta = (v, w) \in \Omega$ and any $\mu \in \mathcal{P}(\Omega)$, define

$$b((v, w); \mu) := \begin{pmatrix} \varphi(X^\top w)^\top \big(y - f_\mu(X)\big) \\ X \Big(\varphi'(X^\top w) \odot \big(y - f_\mu(X)\big)\Big) v \end{pmatrix},$$

where

$$f_\mu(x) := \int_\Omega v \varphi(w^\top x) \, d\mu(v, w), \qquad f_\mu(X) := \big(f_\mu(x_i)\big)_{i=1}^n.$$

Then the gradient-flow dynamics can be written compactly as

$$\dot{\theta}_{t,\alpha} = \frac{\eta}{m} b(\theta_{t,\alpha}; \mu_{m,t}), \qquad \alpha = 1, \dots, m.$$

In particular, each particle interacts with the rest of the system only through the current empirical measure $\mu_{m,t}$ (via $f_{\mu_{m,t}}(X)$).

A.2.4 Continuity equation and push-forward formulation

Equivalently, $\mu_{m,t}$ solves the continuity equation

$$\partial_t \mu_{m,t} + \nabla_\theta \cdot \Big(\mu_{m,t} \frac{\eta}{m} b(\cdot; \mu_{m,t})\Big) = 0, \tag{A.7}$$

in the sense of distributions, i.e. for every test function $\psi \in C_c^\infty(\Omega)$,

$$\frac{d}{dt}\int_\Omega \psi(\theta)\, d\mu_{m,t}(\theta) = \frac{\eta}{m}\int_\Omega \nabla_\theta \psi(\theta)\cdot b(\theta;\mu_{m,t})\, d\mu_{m,t}(\theta).$$

Choosing the mean-field scaling $\eta = m$ yields a nontrivial $O(1)$ evolution in time and, as $m \to \infty$, one expects $\mu_{m,t} \to \mu_t$, where the limit μ_t satisfies

$$\partial_t \mu_t + \nabla_\theta \cdot \big(\mu_t\, b(\cdot;\mu_t)\big) = 0, \qquad \mu_0 = \mathcal{N}(0, I_{d+1}). \tag{A.8}$$

Push-forward formulation. Recall that if $g : \mathcal{X} \to \mathcal{Z}$ is measurable and $\mu \in \mathcal{P}(\mathcal{X})$, then the *push-forward* of μ by g is the measure $g_*\mu \in \mathcal{P}(\mathcal{Z})$ defined by

$$(g_*\mu)(A) := \mu\big(g^{-1}(A)\big), \qquad A \subset \mathcal{Z} \text{ measurable.}$$

Let $\Phi_t : \Omega \to \Omega$ denote the flow map associated with the velocity field $\frac{\eta}{m}\, b(\cdot;\mu_{m,t})$. Then the empirical measure is transported by this flow:

$$\mu_{m,t} = (\Phi_t)_*\mu_{m,0}.$$

In particular, for any measurable observable $h : \Omega \to \mathbb{R}$,

$$\int_\Omega h(\theta)\, d\mu_{m,t}(\theta) = \int_\Omega h(\Phi_t(\theta))\, d\mu_{m,0}(\theta),$$

so evaluating the network at time t is obtained by taking $h_x(\theta) := v\varphi(w^\top x)$ and writing

$$f_{\mu_{m,t}}(x) = \int_\Omega v\varphi(w^\top x)\, d\mu_{m,t}(v,w) = \int_\Omega v\varphi(w^\top x)\, d\big((\Phi_t)_*\mu_{m,0}\big)(v,w).$$

A.3 Comparison with the NTK scaling

The distinction between the mean-field and NTK viewpoints can also be seen from the tangent kernel. Recall that, in the NTK parameterization used in the main text, the two-layer network is scaled by $1/\sqrt{m}$. This scaling gives a nontrivial limiting tangent kernel at initialization, and in the infinite-width NTK regime this kernel remains fixed during training.

In the mean-field parameterization,

$$f_\theta(x) = \frac{1}{m}\sum_{\alpha=1}^m v_\alpha \varphi(w_\alpha^\top x),$$

the empirical tangent kernel is instead

$$\Theta_t^{(m)}(x,x') = \frac{1}{m^2}\sum_{\alpha=1}^m \Big(\varphi(w_{t,\alpha}^\top x)\,\varphi(w_{t,\alpha}^\top x') + v_{t,\alpha}^2 \varphi'(w_{t,\alpha}^\top x)\varphi'(w_{t,\alpha}^\top x') x^\top x'\Big). \tag{A.9}$$

This kernel is of order m^{-1}. Under the mean-field time scaling, the effective kernel driving the output dynamics is the scaled quantity $m\Theta_t^{(m)}$. As $m \to \infty$, this has the formal limit

$$\lim_{m\to\infty} m\Theta_t^{(m)}(x,x') = \int_\Omega \Big(\varphi(w^\top x)\varphi(w^\top x') + v^2\varphi'(w^\top x)\varphi'(w^\top x')x^\top x'\Big)\, d\mu_t(v,w).$$

Because the measure μ_t changes with time, this limiting kernel also changes with time. This is the sense in which the mean-field viewpoint can retain feature movement in the large-width limit.

This is useful for studying representation change, but it is less convenient for the spectral analysis in the main text. There, the central object is a fixed operator on $L^2(S^1)$, whose eigenfunctions and eigenvalues can be compared with Fourier modes. In the mean-field viewpoint, the limiting object is the evolving measure μ_t, and the associated kernel changes with time. Thus the same fixed Fourier-diagonal picture is not available directly.

There are also technical limitations. The long-time behavior of the limiting measure-valued dynamics is generally harder to characterize: one may need to ask whether μ_t converges as $t \to \infty$, and which limiting measure is selected. Moreover, the simple empirical-measure description above is most natural for two-layer networks. For deeper architectures, one needs more careful mean-field constructions, as studied for example by Nguyen (2019), Araújo et al. (2019), Sirignano and Spiliopoulos (2019), and Nguyen and Pham (2023).

Appendix B

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 Chapter 4, and computes the corresponding Fourier eigenvalues of the integral operator T. These derivations justify Proposition 5.1 and Theorem 5.2.

B.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

$$\Theta_m(x,y) = \left\langle \nabla_\theta f(x;\theta), \nabla_\theta f(y;\theta) \right\rangle. \tag{B.1}$$

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 (B.1) yields

$$\Theta_m(x,y) = \Theta_m^{(a)}(x,y) + \Theta_m^{(w)}(x,y) + \Theta_m^{(b)}(x,y), \tag{B.2}$$

where

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

B.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

$$\Theta(x,y) = \mathbb{E}\big[\sigma(u)\sigma(v)\big] + \big(x^\top y + 1\big)\mathbb{E}\big[\sigma'(u)\sigma'(v)\big], \tag{B.3}$$

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

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

B.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 δ. 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 B.1. *For* $\delta \in [0,\pi]$,

$$K_1(\delta) = \mathbb{P}(u > 0, v > 0) = \frac{\pi - \delta}{2\pi}. \tag{B.4}$$

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 B.2. *For* $\delta \in [0,\pi]$,

$$K_0(\delta) = \mathbb{E}[\sigma(u)\sigma(v)] = \frac{1}{2\pi}\Big(\sin\delta + (\pi-\delta)\cos\delta\Big). \tag{B.5}$$

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

Proof of Proposition 5.1. Using (B.3), (B.4), and (B.5), we obtain

$$\begin{aligned}\Theta(\delta) &= \frac{1}{2\pi}\Big(\sin\delta + (\pi-\delta)\cos\delta\Big) + \frac{1+\cos\delta}{2\pi}(\pi-\delta) \\ &= \frac{1}{2\pi}\Big(\sin\delta + 2(\pi-\delta)\cos\delta + (\pi-\delta)\Big).\end{aligned} \tag{B.6}$$

This is exactly (5.1). The convolution form of the operator follows from the fact that Θ depends only on the angular difference. □

B.4 Fourier coefficients of the limiting kernel

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

$$\lambda_k = \frac{1}{\pi}\int_0^\pi \Theta(\delta)\cos(k\delta)\,d\delta, \qquad k \geq 0. \tag{B.7}$$

Substituting (B.6) 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 B.3. *For every integer* $m \geq 1$,

$$I_m := \int_0^\pi (\pi-\delta)\cos(m\delta)\,d\delta = \frac{1-(-1)^m}{m^2}. \tag{B.8}$$

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 5.2. 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 B.3, 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

$$\lambda_k = \frac{1}{\pi^2 k^2}, \qquad k \geq 3 \text{ odd}. \tag{B.9}$$

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

$$\lambda_k = \frac{k^2+3}{\pi^2(k^2-1)^2}, \qquad k \geq 2 \text{ even.} \tag{B.10}$$

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 (5.3) and (5.4). □

B.5 Effect of the bias contribution

We now isolate the effect of the trainable hidden bias. From the NTK decomposition in (B.2), the limiting kernel can be written as

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

The first two terms come from differentiating with respect to the output weights and input weights. The last term comes from differentiating with respect to the hidden biases. Thus, if the trainable hidden-bias contribution is removed, the limiting kernel becomes

$$\Theta^{(\mathrm{nb})}(x,y) = K_0(x,y) + x^\top y K_1(x,y).$$

On the circle, $x^\top y = \cos\delta$, so by (B.5) and (B.4),

$$\begin{aligned}\Theta^{(\mathrm{nb})}(\delta) &= \frac{1}{2\pi}\Big(\sin\delta + (\pi-\delta)\cos\delta\Big) + \cos\delta\,\frac{\pi-\delta}{2\pi} \\ &= \frac{1}{2\pi}\Big(\sin\delta + 2(\pi-\delta)\cos\delta\Big).\end{aligned} \tag{B.11}$$

The Fourier eigenvalues of this no-bias kernel are

$$\lambda_k^{(\mathrm{nb})} = \frac{1}{\pi}\int_0^\pi \Theta^{(\mathrm{nb})}(\delta)\cos(k\delta)\, d\delta.$$

Using (B.11), this becomes

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

In the notation of the proof of Theorem 5.2,

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

whereas the full eigenvalue is

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

Thus the term C_k, which comes from

$$\frac{1}{2\pi}(\pi - \delta),$$

is precisely the trainable-bias contribution.

For $k = 0$, direct integration gives

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

Since

$$\int_0^\pi \sin\delta\, d\delta = 2, \qquad \int_0^\pi (\pi-\delta)\cos\delta\, d\delta = 2,$$

we obtain

$$\lambda_0^{(\mathrm{nb})} = \frac{1}{2\pi^2}(2+4) = \frac{3}{\pi^2}.$$

For $k = 1$,

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

The first integral is zero. For the second, using

$$\cos^2\delta = \frac{1+\cos(2\delta)}{2},$$

we get

$$2\int_0^\pi (\pi-\delta)\cos^2\delta\, d\delta = \int_0^\pi (\pi-\delta)\, d\delta + \int_0^\pi (\pi-\delta)\cos(2\delta)\, d\delta.$$

The second term is zero by Lemma B.3, since 2 is even. Hence

$$2\int_0^\pi (\pi-\delta)\cos^2\delta\, d\delta = \frac{\pi^2}{2},$$

and therefore

$$\lambda_1^{(\mathrm{nb})} = \frac{1}{4}.$$

For $k \geq 2$, we use the values of A_k and B_k computed in the proof of Theorem 5.2. If $k \geq 3$ is odd, then $k-1$ and $k+1$ are even, so

$$A_k = 0, \qquad B_k = 0.$$

Therefore

$$\lambda_k^{(\mathrm{nb})} = 0, \qquad k \geq 3 \text{ odd.} \tag{B.12}$$

If $k \geq 2$ is even, then

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

Thus

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

and hence

$$\lambda_k^{(\mathrm{nb})} = \frac{k^2+3}{\pi^2(k^2-1)^2}, \qquad k \geq 2 \text{ even.} \tag{B.13}$$

This proves Corollary 5.3. Comparing the no-bias eigenvalues with the full eigenvalues in Theorem 5.2, the trainable-bias contribution is

$$\lambda_0 - \lambda_0^{(\mathrm{nb})} = \frac{1}{4}, \qquad \lambda_1 - \lambda_1^{(\mathrm{nb})} = \frac{1}{\pi^2},$$

while for $k \geq 2$,

$$\lambda_k - \lambda_k^{(\mathrm{nb})} = \begin{cases} 0, & k \geq 2 \text{ even}, \\ \dfrac{1}{\pi^2 k^2}, & k \geq 3 \text{ odd}. \end{cases}$$

Thus the bias term does not change the even frequencies $k \geq 2$, but it is exactly what makes the odd frequencies $k \geq 3$ visible in the fixed-kernel dynamics. The frequency $k = 1$ is exceptional: it already has eigenvalue $1/4$ without the bias, and the bias adds the additional $1/\pi^2$ contribution.

Appendix C

Proofs of the finite-sample estimates

This appendix proves the finite-sample results from Section 5.2. We work with the truncated Fourier subspace $\mathcal{H}_K$, the feature matrix Φ, the sampled kernel matrix A, the Gram matrix G, and the matrix H defined in (4.20)–(4.23).

C.1 Preliminaries

For each basis function in (4.14)–(4.15),

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

We also write

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

For each basis index p, let λ_p denote the eigenvalue of the operator T associated with ϕ_p, so that

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

Define the finite-sample corrected eigenvalues by

$$\lambda_p^{(n)} = \left(1 - \frac{1}{n}\right)\lambda_p + \frac{\Theta(0)}{n}, \tag{C.3}$$

and let

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

C.2 Expectation of H

Proof of Proposition 5.4. By definition,

$$H_{pq} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi_p(\varphi_i)\Theta(\varphi_i - \varphi_j)\phi_q(\varphi_j). \tag{C.4}$$

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

C.3 Concentration of the Gram matrix

Proof of (5.13). 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 (C.1),

$$|Z_i| \leq 2.$$

If $p \neq q$, then $\delta_{pq} = 0$, so $|Y_i| = |Z_i| \leq 2$. 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 (C.1),

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

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 (5.13). □

C.4 Concentration of H

Proof of (5.14). 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 (C.1) and (C.2), each summand satisfies

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

Hence, when φ_s is replaced by φ_s', a single summand can change by at most 4κ.

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

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

A union bound over all d^2 entries gives

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

as claimed. □

Remark C.1. 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 in this thesis, we do not pursue this refinement here.

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

Proof of (5.15). 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 φ_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 φ_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 $Y_1, \ldots, Y_m$ 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 φ_i.

First, by (C.1) and (C.2),

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

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 (C.1) and (C.2),

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

Thus, conditionally on φ_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

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

Removing the conditioning and simplifying gives

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

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

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

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 (5.15). □

C.6 Proof of the simultaneous high-probability bound

Corollary C.2 (High-probability bounds)**.** *Let* $\delta \in (0,1)$*, and define*

$$u_{n,d,\delta}^{(G)} := \log\left(\frac{3d(d+1)}{\delta}\right), \tag{C.5}$$

$$u_{n,d,\delta}^{(H)} := \log\left(\frac{6d^2}{\delta}\right), \tag{C.6}$$

and

$$u_{n,d,\delta}^{(A)} := \log\left(\frac{6nd}{\delta}\right). \tag{C.7}$$

Now set

$$\alpha_{n,d,\delta} := 2\sqrt{\frac{u_{n,d,\delta}^{(G)}}{n}} + \frac{2}{3}\frac{u_{n,d,\delta}^{(G)}}{n}, \tag{C.8}$$

$$\beta_{n,d,\delta} := \kappa\sqrt{\frac{32}{n}u_{n,d,\delta}^{(H)}}, \tag{C.9}$$

and

$$\gamma_{n,d,\delta} := \kappa\sqrt{\frac{2}{n}u_{n,d,\delta}^{(A)}} + \kappa\frac{u_{n,d,\delta}^{(A)}}{n}. \tag{C.10}$$

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

$$\|G - I_d\|_{\max} \le \alpha_{n,d,\delta}, \qquad \|H - \Lambda^{(n)}\|_{\max} \le \beta_{n,d,\delta}, \qquad \|A\Phi - \Phi\Lambda^{(n)}\|_{\max} \le \gamma_{n,d,\delta}. \tag{C.11}$$

Proof of Corollary C.2. Set

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

Define

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

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

We first recall the standard Bernstein inversion: if

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

then choosing

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

ensures

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

For the Gram matrix bound (5.13), the Bernstein proof used

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

Thus, with

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

the choice

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

yields

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

For the compressed operator, the McDiarmid estimate (5.14) 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 (5.15), 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 (C.11). □

Appendix D

Proofs of the finite-width frozen-kernel estimates

This appendix proves the finite-width frozen-kernel results stated in Section 5.4. We use the notation introduced in Chapter 4. In particular, $T_0^{(m)}$ denotes the frozen finite-width tangent operator at initialization, T_∞ 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.

D.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

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

The corresponding one-neuron kernel is

$$\Theta^{(r)}(\theta,\eta) = \psi_r(\theta)^\top\psi_r(\eta). \tag{D.2}$$

Substituting (D.1) gives

$$\begin{aligned}\Theta^{(r)}(\theta,\eta) &= \sigma(u_r(\theta))\sigma(u_r(\eta)) \\ &\quad + a_r^2\mathbf{1}_{\{u_r(\theta)>0\}}\mathbf{1}_{\{u_r(\eta)>0\}}\big(\cos\theta\cos\eta + \sin\theta\sin\eta + 1\big).\end{aligned} \tag{D.3}$$

Using

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

we obtain

$$\Theta^{(r)}(\theta,\eta) = \sigma(u_r(\theta))\sigma(u_r(\eta)) + a_r^2\mathbf{1}_{\{u_r(\theta)>0\}}\mathbf{1}_{\{u_r(\eta)>0\}}\big(\cos(\theta-\eta)+1\big). \tag{D.4}$$

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

$$(T^{(r)}g)(\theta) := \int_0^{2\pi}\Theta^{(r)}(\theta,\eta)g(\eta)\,d\mu(\eta). \tag{D.5}$$

Then, by (4.40),

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

D.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

$$(B_m^{(K)})_{pq} = \langle \phi_p, T_0^{(m)} \phi_q \rangle_{L^2(\mu)}. \tag{D.6}$$

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

$$\xi_{r,pq} := \langle \phi_p, T^{(r)} \phi_q \rangle. \tag{D.7}$$

Then

$$(B_m^{(K)})_{pq} = \frac{1}{m} \sum_{r=1}^{m} \xi_{r,pq}. \tag{D.8}$$

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

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

then

$$B_m^{(K)} = \frac{1}{m} \sum_{r=1}^{m} Z_r. \tag{D.9}$$

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

D.3 Expectation of the projected operator

Proof of (5.19). From (D.9),

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

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_∞ is diagonal in the Fourier basis,

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

Therefore

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

□

D.4 Sub-exponential bound for one projected entry

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

$$\xi_{r,pq} = \int_0^{2\pi} \int_0^{2\pi} \phi_p(\theta)\Theta^{(r)}(\theta,\eta)\phi_q(\eta)\, d\mu(\eta)\, d\mu(\theta). \tag{D.10}$$

Taking absolute values gives

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

Since every retained Fourier basis function satisfies

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

we obtain

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

We now bound the kernel. Since

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

Cauchy–Schwarz gives

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

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 (D.4), we conclude that

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

Since μ is a probability measure,

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

Therefore

$$|\xi_{r,pq}| \le 2\|w_r\|^2 + 4a_r^2. \tag{D.11}$$

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

$$\mathbb{E}[e^{tY_r}] = (1-4t)^{-1}(1-8t)^{-1/2}, \qquad t < 1/8. \tag{D.12}$$

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 ψ_1-norm,

$$\|Y_r\|_{\psi_1} \le 16. \tag{D.13}$$

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

$$\|\xi_{r,pq} - \mathbb{E}[\xi_{r,pq}]\|_{\psi_1} \le 32. \tag{D.14}$$

D.5 Entrywise and blockwise concentration

Proof of (5.20). 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 (D.14),

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

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

$$\mathbb{P}\Big(\Big| (B_m^{(K)})_{pq} - (\Lambda_K)_{pq} \Big| \ge \varepsilon \Big) \le 2 \exp\left(-cm \min\left(\frac{\varepsilon^2}{32^2}, \frac{\varepsilon}{32} \right) \right). \tag{D.15}$$

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 (D.15),

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

Since $d = 2K+1$, this proves (5.20). □

Appendix 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

Figure E.1 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.

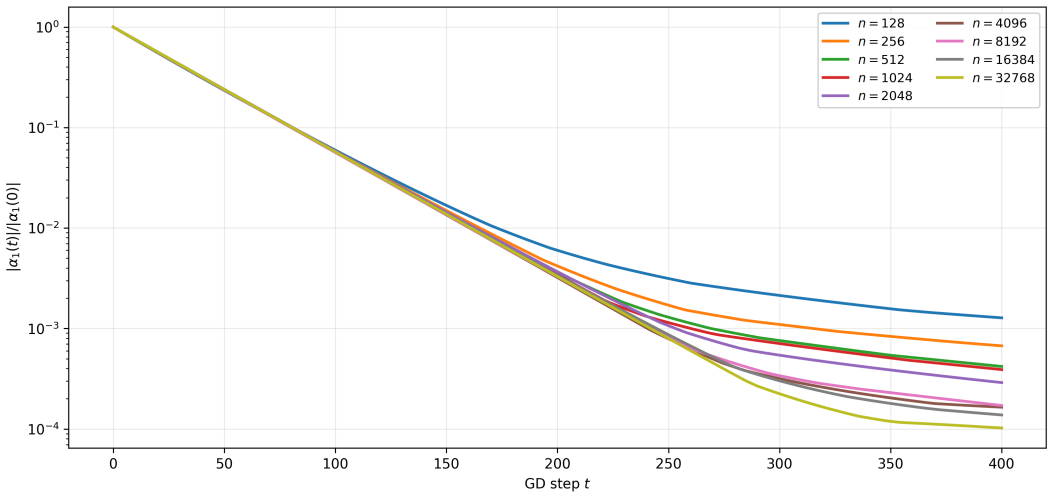


Figure E.1: Normalised residual for mode $k = 1$ over the first 400 gradient steps, for sample sizes $n \in \{128, \ldots, 32768\}$. Larger n gives faster effective decay, converging toward the continuum rate.

Figure E.2 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.

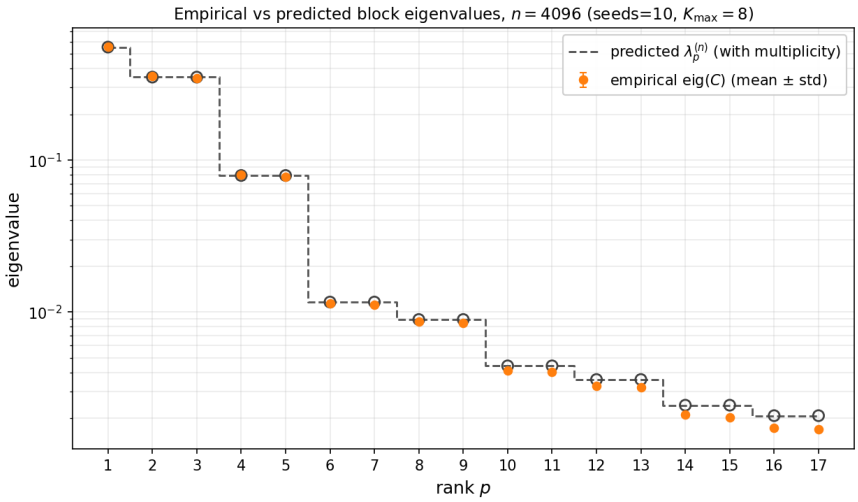


Figure E.2: Empirical versus predicted block eigenvalues at $n = 4096$, on a logarithmic scale, averaged over 10 seeds. The dashed step shows the predicted eigenvalues $\lambda_p^{(n)}$ with their correct multiplicity (single for $k = 0$, degenerate pairs for $k \geq 1$); markers show the empirical eigenvalues eig(C) (mean $\pm$ standard deviation).

Figure E.3 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.

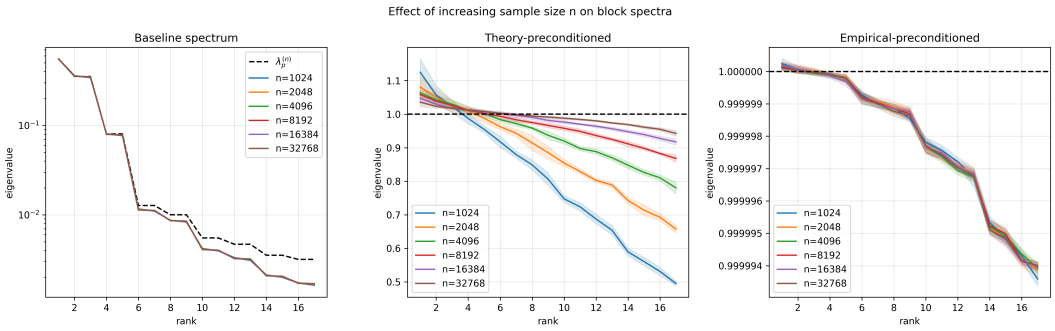


Figure E.3: Block spectra of the compressed Gram matrix for $n \in \{1024, \ldots, 32768\}$. Left: baseline (decaying) spectrum. Centre: theory-preconditioned spectrum approaching 1. Right: empirical-preconditioned spectrum, closely tracking the theory.

Figure E.4 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 $\kappa \approx 1$.

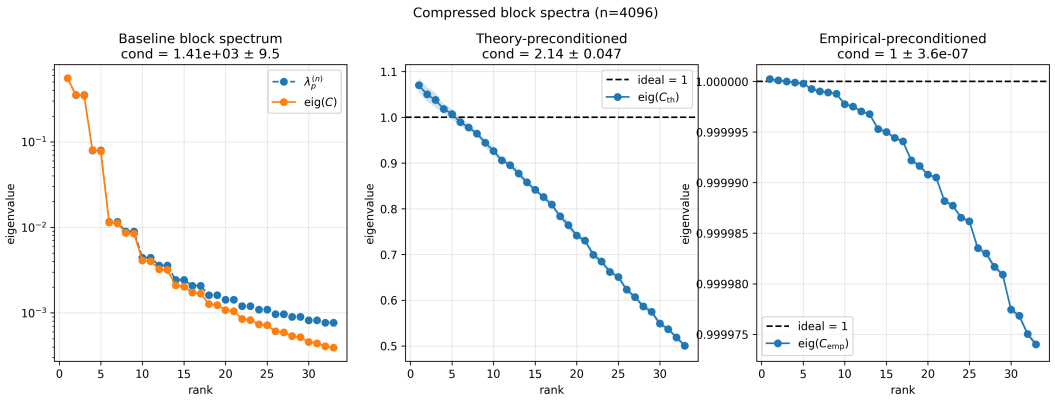


Figure E.4: Compressed block spectra at $n = 4096$ with all active modes inside the preconditioning block. Left: baseline spectrum with a wide spread. Centre: theory-preconditioned spectrum, reduced spread but a residual slope. Right: empirical-preconditioned spectrum, essentially flat at 1. Titles report the per-block condition number.

Figures E.5–E.6 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 5.4 and 5.5). Here we repeat the same experiment across the full range of sample sizes $n \in \{128, \ldots, 32768\}$. 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 $|\alpha_p(t)|/|\alpha_p(0)|$; 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.

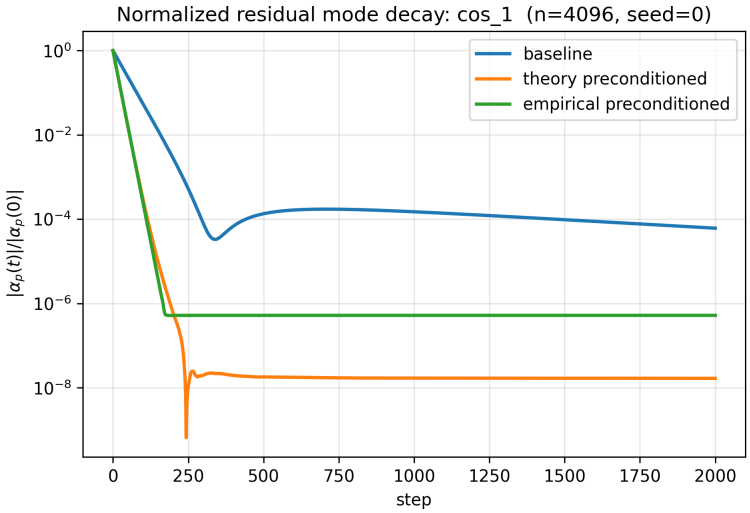


Figure E.5: Per-mode residual decay for $\cos_1$ ($n = 4096$, seed 0). Baseline (blue) is effectively flat; theory-preconditioned (orange) and empirical-preconditioned (green) converge rapidly.

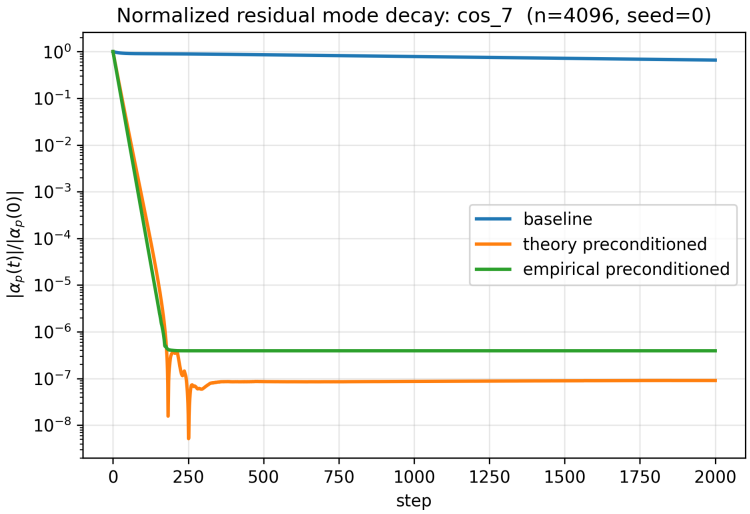
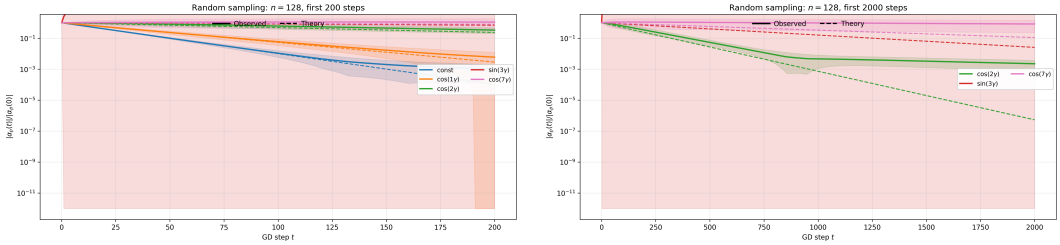


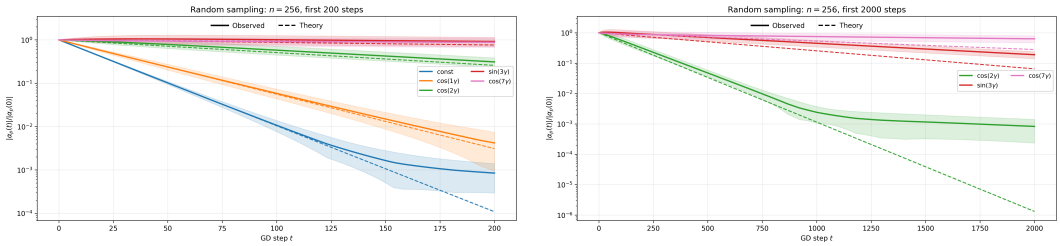
Figure E.6: Per-mode residual decay for $\cos_7$ ($n = 4096$, seed 0). The spectral gap for this higher mode makes the baseline decay even slower; preconditioning recovers roughly six orders of magnitude of relative residual within 2000 steps.



(a) All modes, first 200 steps.

(b) Higher-frequency modes, 2000 steps.

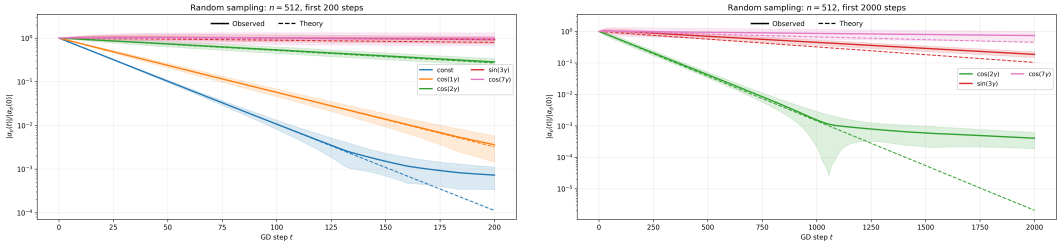
Figure E.7: Random sampling with $n = 128$: observed (solid) versus finite-sample eigenvalue prediction (dashed). The prediction bands are wide and the observed curves flatten early.



(a) All modes, first 200 steps.

(b) Higher-frequency modes, 2000 steps.

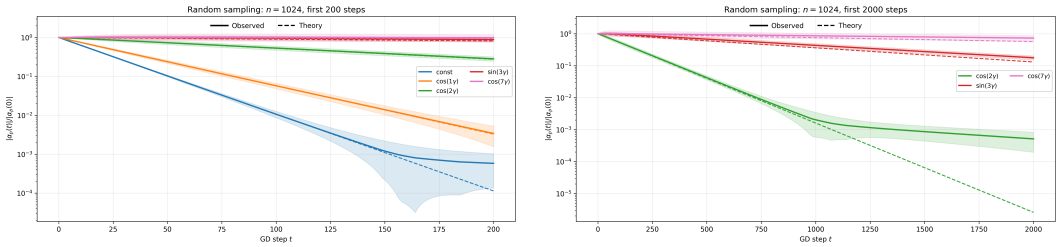
Figure E.8: Random sampling with $n = 256$: observed (solid) versus finite-sample eigenvalue prediction (dashed).



(a) All modes, first 200 steps.

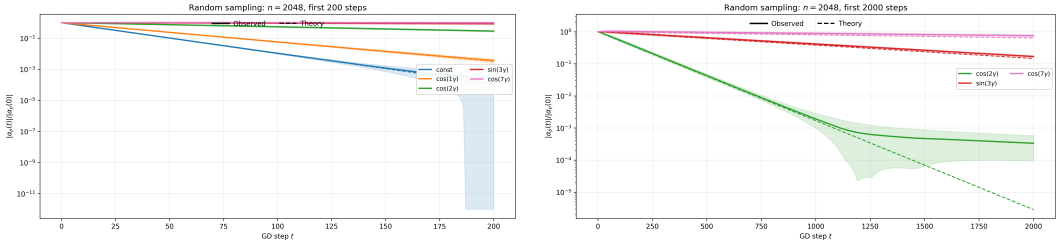
(b) Higher-frequency modes, 2000 steps.

Figure E.9: Random sampling with $n = 512$: observed (solid) versus finite-sample eigenvalue prediction (dashed).



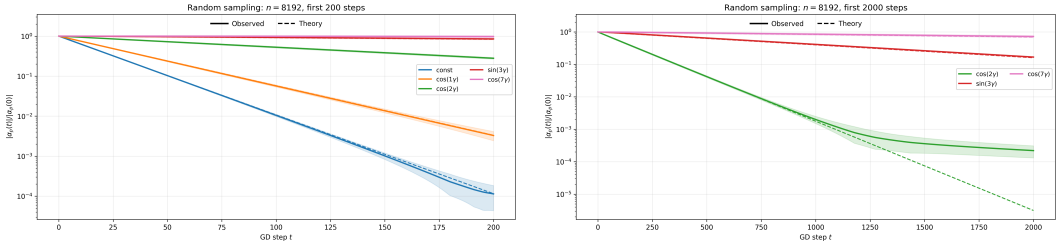
(a) All modes, first 200 steps. (b) Higher-frequency modes, 2000 steps.

Figure E.10: Random sampling with $n = 1024$: observed (solid) versus finite-sample eigenvalue prediction (dashed). The observed decay already tracks the prediction over most of the early window.



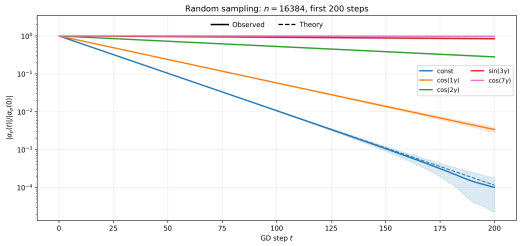
(a) All modes, first 200 steps. (b) Higher-frequency modes, 2000 steps.

Figure E.11: Random sampling with $n = 2048$: observed (solid) versus finite-sample eigenvalue prediction (dashed).

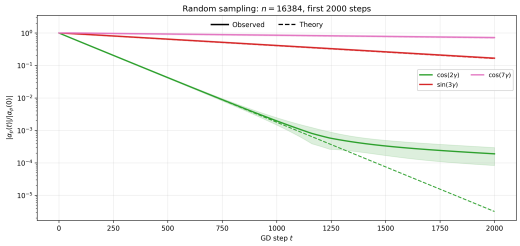


(a) All modes, first 200 steps. (b) Higher-frequency modes, 2000 steps.

Figure E.12: Random sampling with $n = 8192$: observed (solid) versus finite-sample eigenvalue prediction (dashed).

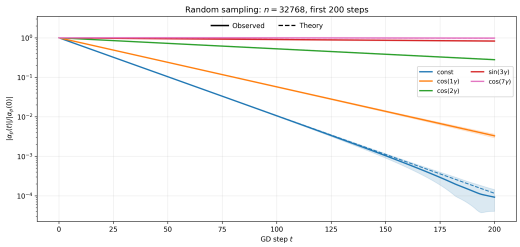


(a) All modes, first 200 steps.

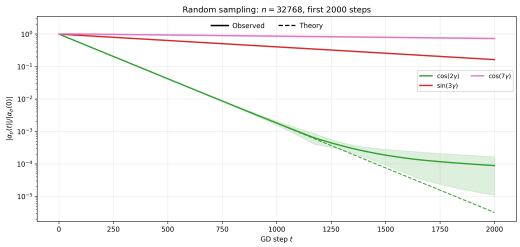


(b) Higher-frequency modes, 2000 steps.

Figure E.13: Random sampling with $n = 16384$: observed (solid) versus finite-sample eigenvalue prediction (dashed).



(a) All modes, first 200 steps.



(b) Higher-frequency modes, 2000 steps.

Figure E.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.

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 assumption 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. Figure E.15 shows that the eigenvalue alignment error and the empirical pair-splitting both decrease quickly with width, yet these quantities can be small at high frequencies while the corresponding eigenspaces are still poorly aligned.

The eigenspace diagnostics show a clear frequency hierarchy. Figure E.16 compares 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. Figure E.17 shows the same effect in the raw Fourier plane, where the low-frequency clouds lie near the unit circle and higher frequencies drift inward. Figure E.18 summarises this as a projector error

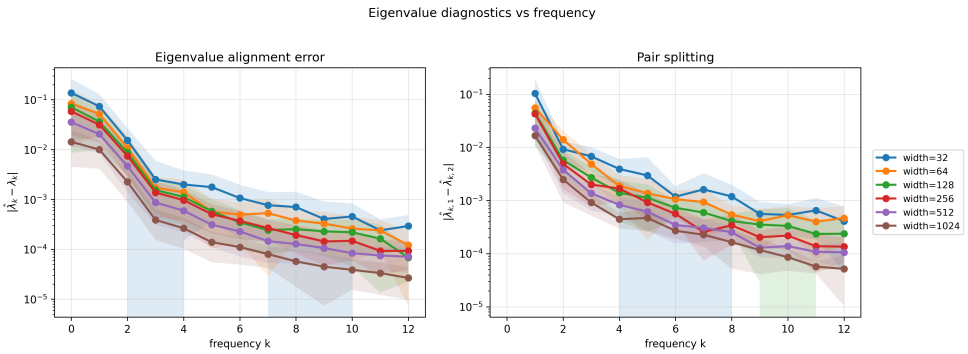


Figure E.15: Eigenvalue diagnostics for the frozen empirical kernel versus frequency, across widths. Left: absolute eigenvalue alignment error $|\widehat{\lambda}_k - \lambda_k|$. Right: empirical splitting of the degenerate k-pair $|\widehat{\lambda}_{k,1} - \widehat{\lambda}_{k,2}|$. Both shrink with width, but small eigenvalue error alone does not imply good eigenspace recovery.

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. Figure E.19 shows the principal angles increasing during training at width 256, with the largest drift at $k = 5$ and $k = 6$; Figure E.20 shows the same drift 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 E.21 confirms this directly: the absolute projection error $\|\hat{P}_k(t) - \hat{P}_k(0)\|_F$ stays close to zero for the frozen network at both widths, while it grows for the evolving network, especially at the smallest width.

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

Figure E.22 shows the full per-subspace story—alignment, projection error, and spectral strength—across all four widths and six target frequencies simultaneously. The pattern is consistent: alignment is high and grows with width; projection error is moderate and shrinks with width; spectral strength accumulates at a width-dependent rate.

Figure E.23 checks that the spectral ordering of mode decay is not an artefact of unequal target amplitudes. When the target is constructed with equal Fourier coefficients across all modes, the same low-to-high ordering of convergence is observed, confirming that the effect is driven by the kernel eigenvalue spectrum and not by initialisation.

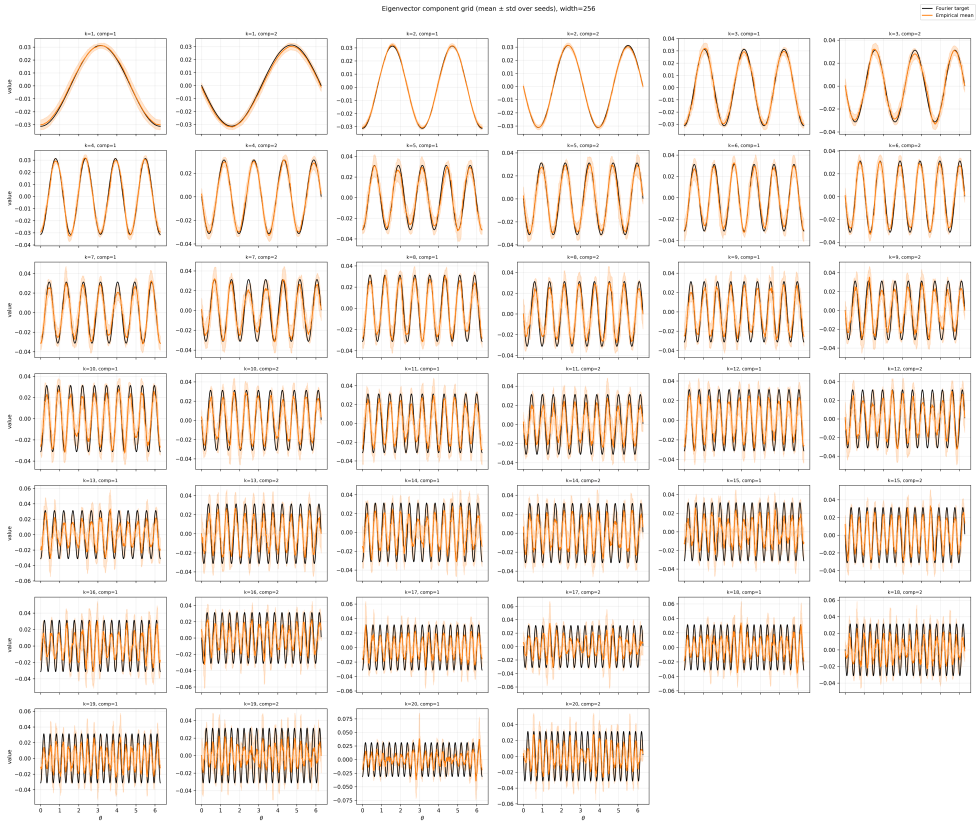


Figure E.16: Mean empirical eigenvector components (orange) versus the Fourier targets (black) at width 256, with one-standard-deviation bands over seeds. Low frequencies are tracked closely; distortion and variability grow with frequency.

Finally, Figure E.24 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.

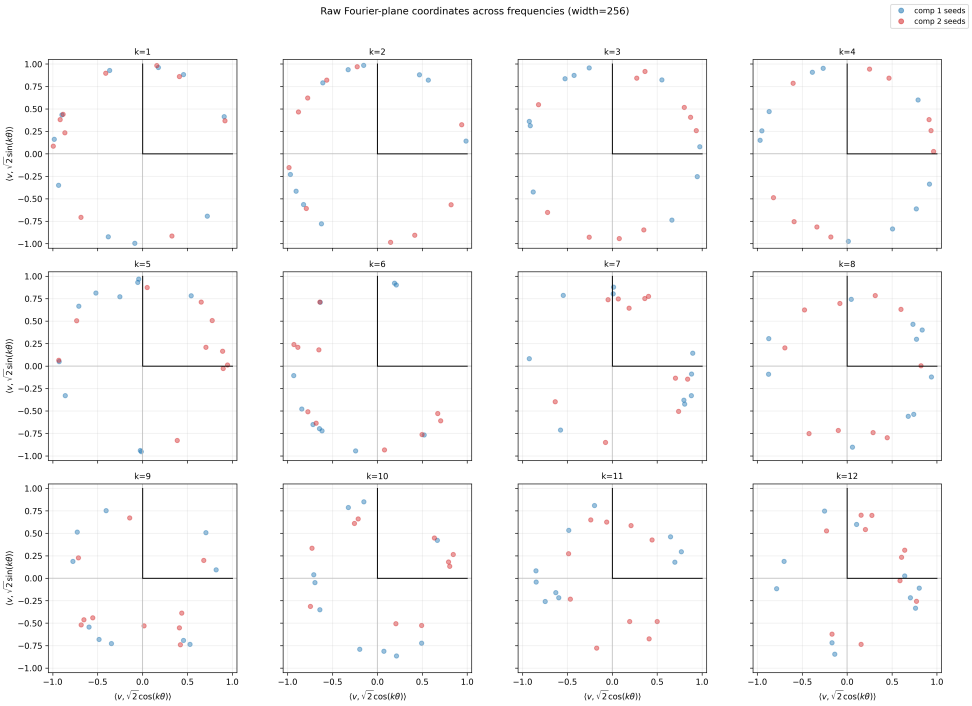


Figure E.17: Raw Fourier-plane coordinates of the empirical eigenvectors across frequencies at width 256. Points near the unit circle lie mostly inside $\mathcal{F}_k$; inward drift indicates leakage outside the Fourier plane.

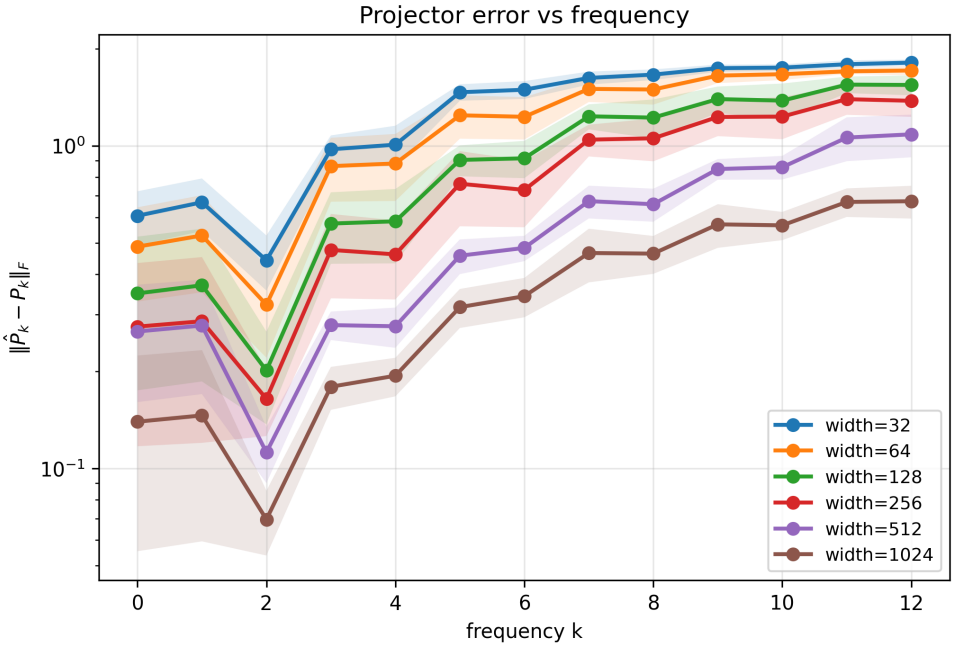


Figure E.18: Projector error $\|\widehat{P}_k - P_k\|_F$ versus frequency for several widths. The error grows rapidly with k: subspace recovery deteriorates at higher frequencies even when the lowest modes are well aligned.

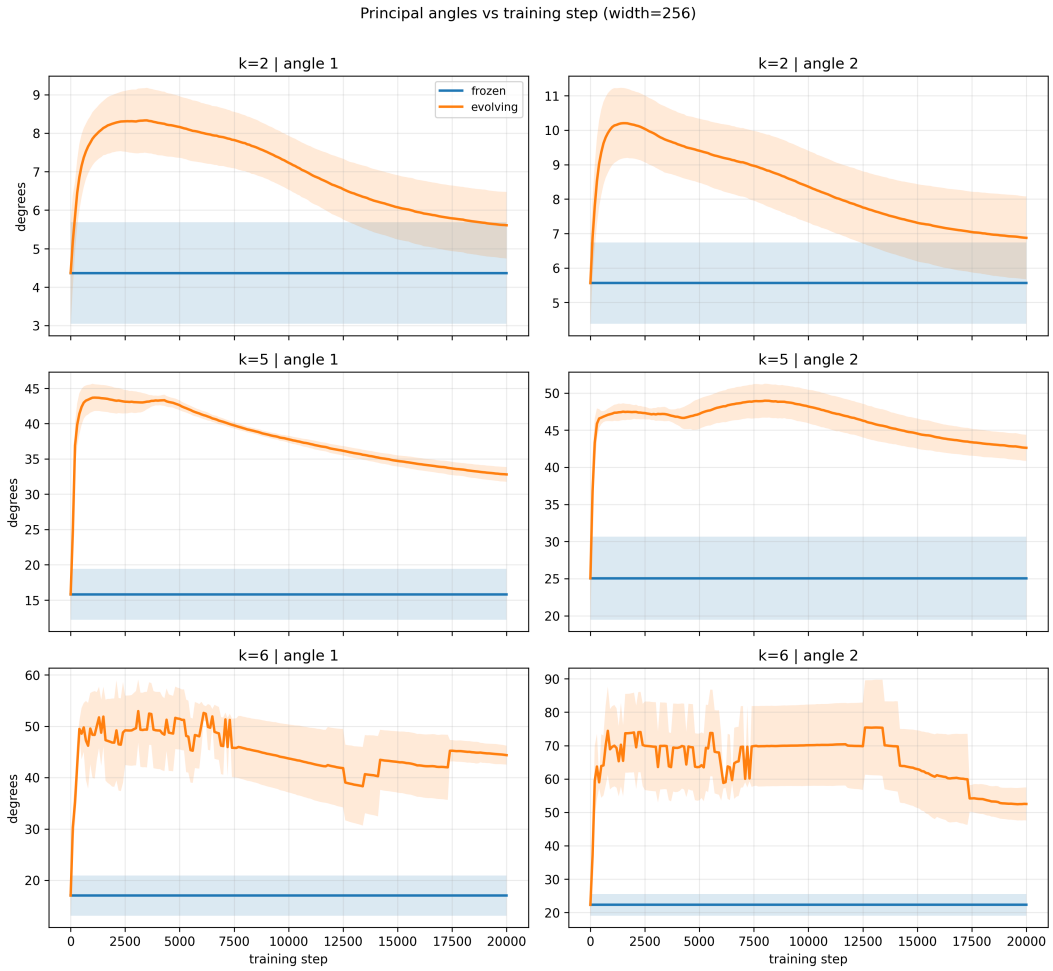


Figure E.19: Principal angles between the empirical eigenspaces and the Fourier planes during training at width 256. The frozen branch is constant by construction; the evolving branch departs immediately, with the largest drift at $k = 5$ and $k = 6$.

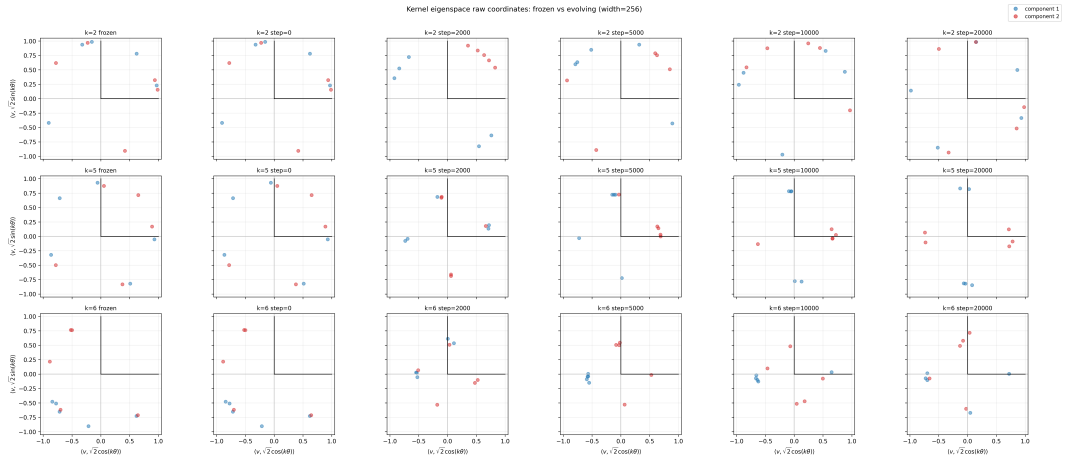
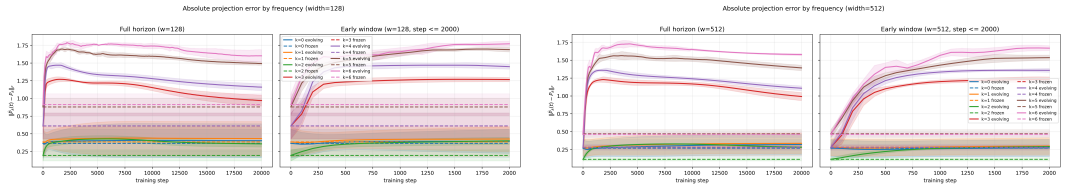


Figure E.20: Raw Fourier-plane coordinates of the empirical eigenspaces for $k = 2, 5, 6$, comparing the frozen baseline with the evolving branch at several training steps (width 256). The frozen and step-0 columns coincide; the $k = 5$ and $k = 6$ eigenspaces reorganise away from the initial Fourier pattern.



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

Figure E.21: 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.

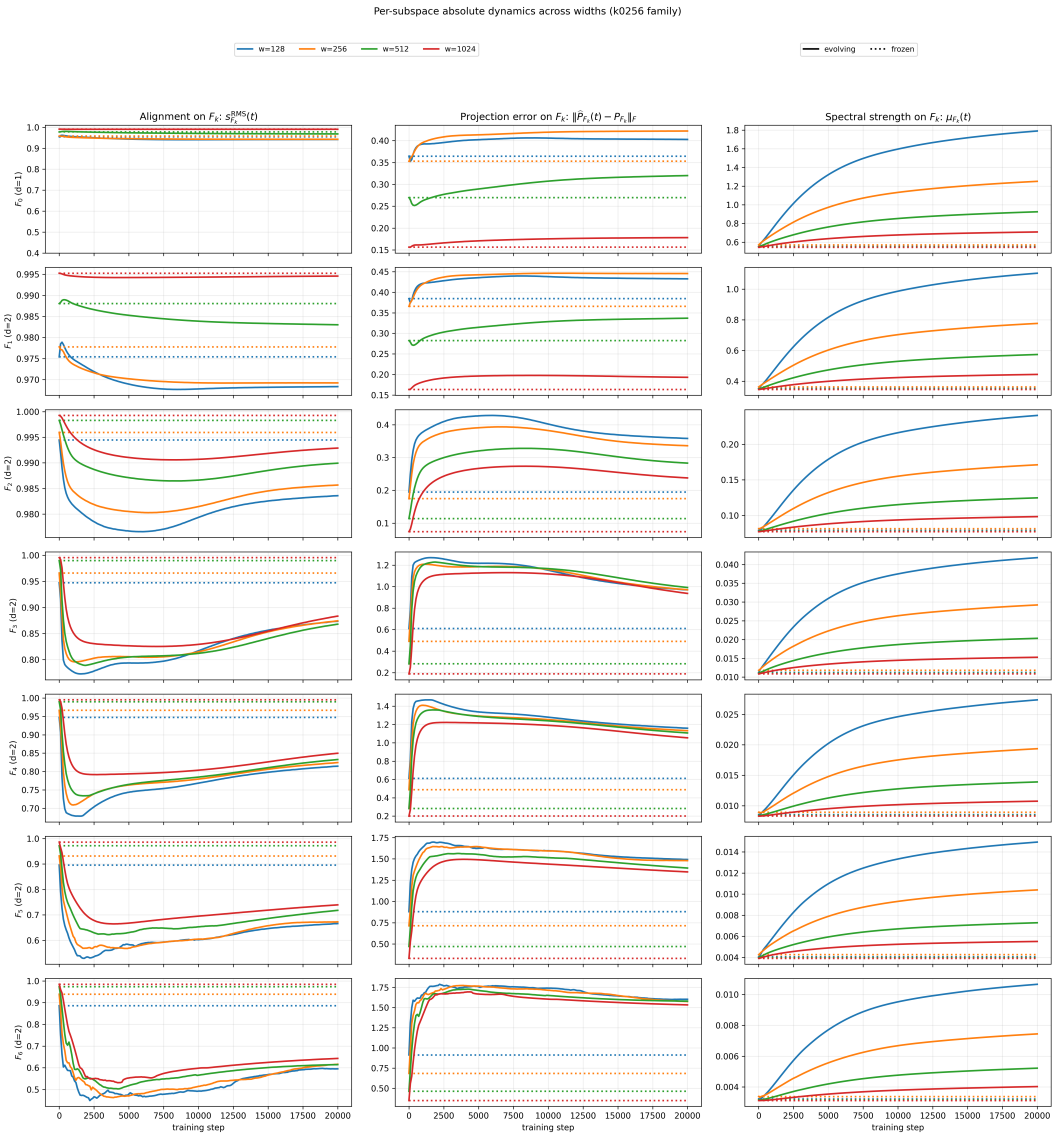


Figure E.22: Per-subspace dynamics across widths $w \in \{128, 256, 512, 1024\}$ (line colours) for each of the six target frequencies (rows). Columns show cosine alignment, absolute projection error, and spectral strength, for both frozen (dashed) and evolving (solid) networks.

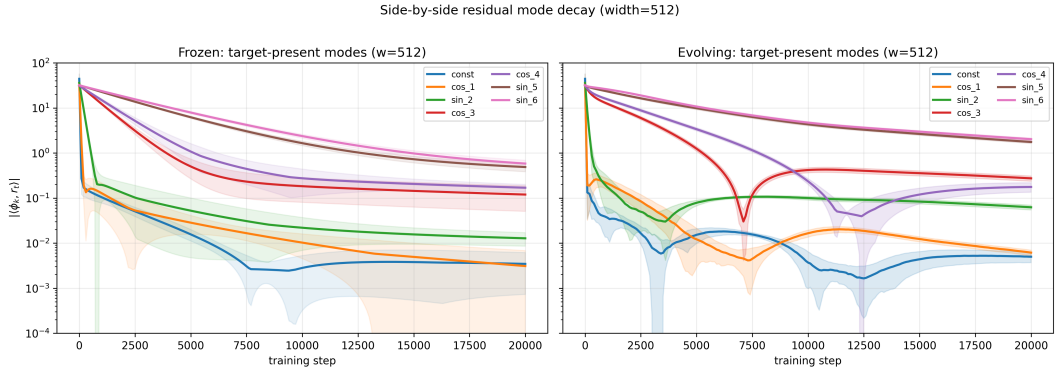
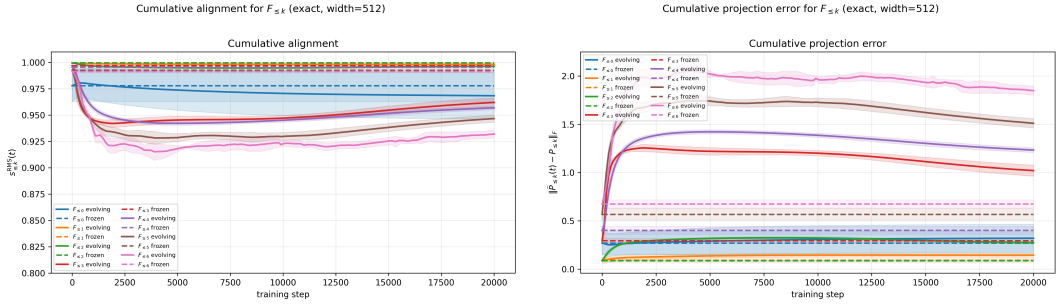


Figure E.23: Mode residual decay for frozen (left) and evolving (right) networks at $w = 512$, using a target with equal Fourier amplitudes across all modes. The low-frequency modes still converge first, ruling out amplitude bias as a confound.



(a) Cumulative RMS-cosine alignment.

(b) Cumulative projection error.

Figure E.24: 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.