Lecture Notes on Neural Network Scaling Limits

Mufan Li

Draft Version: September 5, 2026

Contents

Chapter 1

Introduction

These lecture notes grew out of STAT 946—Topics in Deep Learning Theory, taught at the University of Waterloo in Fall 2025. I taught the course from handwritten notes based on papers that I had read or written, including work that was then nearing completion. One motivation for preparing this document is a question I am often asked: what is a good place to start reading about deep learning theory? My aim is to offer a deliberately subjective guide to several directions in the field, with an emphasis on questions connected to my own research program.

The document began as a series of lecture notes scribed by dedicated students; many of the submissions far exceeded what I had expected from an ungraded task. GPT Pro and, later, Sol Ultra have been invaluable in combining and editing the scribed notes with the original papers and handwritten notes, substantially reducing the time required while improving the correctness and completeness of the resulting manuscript. Roughly one week of evening work produced a seemingly decent first draft, but very soon I realized the notes required much more careful editing and lots of time commitment. I should say I remain not completely happy with the current version, especially as the prose currently still reads quite AI-like in many places, although I'm pretty confident the mathematical content is correct and complete.

These lecture notes focus on *scaling limits* of neural networks. On the one hand, the mathematics of scaling limits is elegant and worth studying in its own right. On the other hand, many theorists, myself included, would like this work to have practical consequences. It is therefore worth motivating the selection of topics before turning to the technical details, including why some subjects that others may regard as foundational are not emphasized here.

1.1 Motivation

Modern deep learning is, in many ways, an empirical science: researchers propose a method, scale it up, and compare its performance on benchmarks. In recent years, increasing model, data, and compute scale has become a particularly reliable route to improved performance. As industrial research laboratories train increasingly large models, the gap between theory and practice continues to widen. Many mathematical obstacles make theoretical progress difficult, which raises a basic question:

> *What is the best way for theory to contribute to deep learning?*

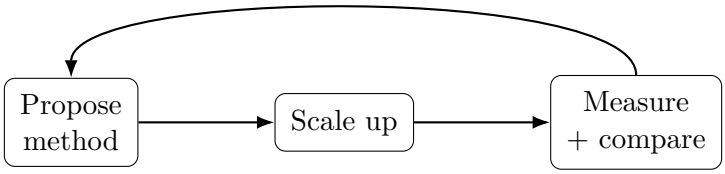

Figure 1.1: A recurring pattern in deep learning research: new ideas are often validated (or rejected) only after large-scale training runs.

It is instructive to examine empirical practice for clues; the cycle is summarized in Figure 1.1. First, the cycle of proposing a method, scaling it up, and measuring it on benchmarks has important limitations. In particular, benchmark comparisons do not always identify a uniformly better method. Conclusions can change with the architecture, data regime, tuning budget, or evaluation protocol, even when the empirical literature is extensive. When extensive experiments still do not yield a stable comparison, what can theory contribute?

Yet empirical practice also reveals a striking regularity. Across several deep learning tasks and over substantial measured ranges, test loss often decreases predictably as model size, dataset size, or training compute increases [23]. These empirical relations are called *scaling laws*: when the remaining resources are not limiting, test loss is often well approximated by a power law in the resource being varied [29]. A power law appears approximately as a straight line on log–log axes, as illustrated in Figure 1.2.

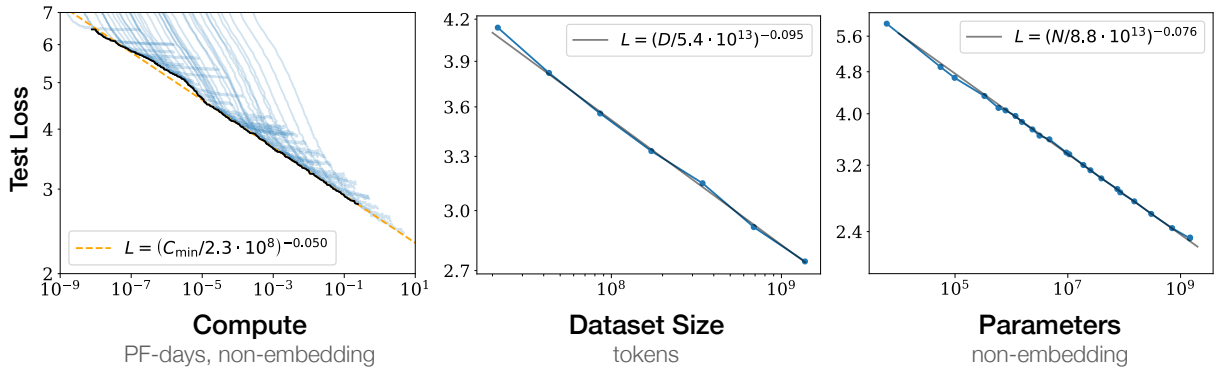

Figure 1.2: Empirical neural-network scaling laws for language modeling (Kaplan et al. [29], Figure 1). On log–log axes, test loss decreases approximately as a power law in (left) training compute C, (middle) dataset size D, and (right) non-embedding parameter count N, within regimes not bottlenecked by the remaining resources.

Within a fitted scaling regime, this predictability makes it possible to extrapolate larger-scale performance from smaller experiments [29]. At a fixed training-compute budget, compute-optimal prescriptions jointly scale model size and training data rather than increasing either one in isolation [29, 24]. The regimes of greatest practical interest, however, involve frontier models that can be prohibitively expensive to train. These observations raise two related questions: why should such simple power laws emerge from complex trained networks, and what can theory tell us about large models without paying their full computational cost?

One clue is that the same visual signature appears far beyond neural networks. Consider the

log–log plot in Figure 1.3.

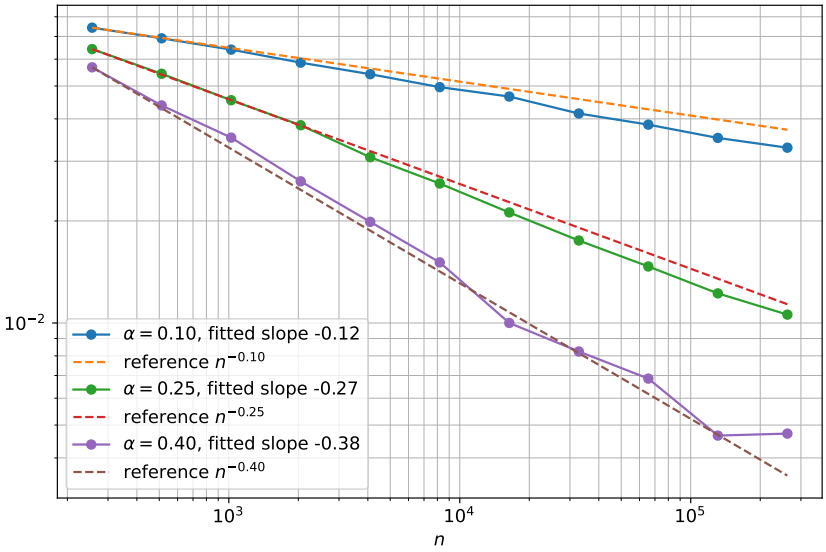

Figure 1.3: A numerical central-limit-theorem convergence experiment. Solid curves show empirical Kolmogorov errors for three values of α, while dashed lines show the corresponding $n^{-\alpha}$ reference rates.

The resemblance is the point: this figure was not produced by training neural networks, but by simulating convergence toward the Gaussian limit for deliberately heavy-tailed summands. Related quantitative central limit theorems turn moment assumptions into power-law error bounds. For example, for i.i.d. centered, unit-variance summands, if $\mathbb{E}[|X_1|^{2+2\alpha}] < \infty$ for some $\alpha \in (0, 1/2]$, then a generalized Berry–Esseen bound gives a Gaussian approximation error of order $O(n^{-\alpha})$ in Kolmogorov distance [43, Chapter V]. The broader point is that a straight line on log–log axes can record the rate at which a finite system approaches its scaling limit.

This comparison suggests viewing a neural-network scaling law as a finite-size correction to an underlying scaling limit. A scaling law describes how an observable changes with scale, whereas a scaling limit describes how the underlying finite system approaches a limiting object. Although trained neural networks of different sizes and architectures have different parameter spaces, we may nevertheless regard:

- $\mathcal{X}$ as a common space of trained neural networks for a fixed task and their possible scaling limits. Unlike a predictor space, $\mathcal{X}$ may retain architecture- and parameter-level information while respecting hidden-unit permutation symmetries;
- $x_n \in \mathcal{X}$ as the trained neural network obtained with compute budget n, with the number of parameters, data points, and training iterations allowed to vary with n;
- $L : \mathcal{X} \to \mathbb{R}$ as a test function, or observable, on this space, such as expected test loss; and
- x_∞ as the limiting object approached by x_n, with $L_\infty := L(x_\infty)$.

Suppose, heuristically, that $x_n \to x_\infty$ in $\mathcal{X}$ and that L detects the leading finite-compute correction nondegenerately. If this correction is of order $n^{-\alpha}$, then

$$L(x_n) = L_\infty + cn^{-\alpha} + o(n^{-\alpha}), \qquad c \neq 0. \tag{1.1}$$

In this sense, the convergence rate seen through the test function L appears as a scaling law. This conclusion is not automatic for every observable: regularity is needed, and cancellations or

degeneracies can change the leading exponent. Because n indexes compute, (1.1) already has the familiar form of a loss–compute scaling law.

This is presently a research program rather than a derivation of the empirical loss laws observed in frontier models. Most available results vary only part of the system, for example width and depth while holding the dataset and training horizon fixed. Nevertheless, the classical example above makes the proposed mechanism plausible: when a quantitative scaling limit has an algebraic convergence rate, suitable observables exhibit power-law finite-size corrections, and empirical neural-network scaling laws may reflect the same general mechanism.

This viewpoint is also particularly well suited to the current era of ever larger models. Frontier systems are increasingly expensive to study by direct experimentation and increasingly unwieldy to describe exactly at finite size. Scaling limits are designed for precisely this situation: they replace a growing family of complicated systems by a tractable asymptotic description while retaining information about how the finite systems approach the limit. They therefore promise both a possible explanation of scaling laws and a practical language for reasoning about systems that cannot be studied exhaustively at their largest scale.

These considerations lead to two recurring goals:

1. use quantitative scaling relations to understand finite-scale behavior and determine when small networks are informative about larger networks; and
2. replace complicated finite networks by simpler limiting objects.

The success of hyperparameter transfer illustrates the first point: under a suitable parameterization, hyperparameters tuned on small networks can transfer to much larger ones [51]. When the limiting approximation is stable across the candidate hyperparameters, small networks can therefore serve as useful proxies for selecting hyperparameters in larger networks. In particular, the mechanisms supporting hyperparameter transfer need not be confined to a single scaling regime: Section 4.4 develops the μP width scaling, while Section 6.3 studies feature learning in a distinct proportional depth–width limit. The following examples emphasize the second point by showing how scaling limits replace complicated finite systems with simpler limiting objects; quantitative versions additionally describe how the finite systems approach those limits.

1.2 Examples of Scaling Limits

We now return to the Central Limit Theorem (CLT), whose quantitative form supplied the motivating plot in Figure 1.3. Its qualitative statement already has the structure of a scaling limit: normalized sums with increasingly many terms converge in distribution to a simple, universal Gaussian law. Let $(X_i)_{i\geq 1}$ be i.i.d. with $\mathbb{E}[X_i] = 0$ and $\mathrm{Var}(X_i) = \sigma^2 \in (0, \infty)$. The CLT states

$$\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1). \tag{1.2}$$

The remarkable aspect is universality: only the first two moments matter for the limit. The scaling $1/\sqrt{n}$ is not cosmetic: it simultaneously keeps the variance $O(1)$ and ensures each individual term contributes $O(1/\sqrt{n})$. Quantitative versions of the CLT additionally control how quickly this Gaussian approximation occurs; Figure 1.3 illustrates the power-law rates discussed above.

The CLT has a functional analogue: a rescaled random walk converges to Brownian motion.

Let $(X_i)_{i\geq 1}$ be i.i.d. real random variables with $\mathbb{E}[X_i] = 0$ and $\mathrm{Var}(X_i) = \sigma^2 \in (0, \infty)$. Define the normalized partial sums

$$Y_k^{(n)} := \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{k} X_i, \qquad k = 0, 1, 2, \ldots. \tag{1.3}$$

To obtain a continuous path, linearly interpolate between consecutive partial sums:

$$Z_t^{(n)} := Y_{\lfloor nt \rfloor}^{(n)} + (nt - \lfloor nt \rfloor)(Y_{\lfloor nt \rfloor + 1}^{(n)} - Y_{\lfloor nt \rfloor}^{(n)}), \qquad t \in [0, T]. \tag{1.4}$$

For each fixed t, the CLT implies $Z_t^{(n)} \xrightarrow{d} \mathcal{N}(0, t)$.

Definition 1.1 (Brownian Motion)**.** A stochastic process $(B_t)_{t\in[0,T]}$ is a (standard) Brownian motion if:

1. $B_0 = 0$ almost surely;
2. for $0 \leq s \leq t$, the increment $B_t - B_s \sim \mathcal{N}(0, t - s)$ and disjoint increments are independent;
3. $t \mapsto B_t$ is almost surely continuous.

For random elements $Z^{(n)}$ and Z in a metric space $(\mathcal{X}, d)$, we write $Z^{(n)} \xrightarrow{d} Z$ when $\mathbb{E}[\psi(Z^{(n)})] \to \mathbb{E}[\psi(Z)]$ for every bounded continuous $\psi : \mathcal{X} \to \mathbb{R}$. Thus convergence in distribution applies directly to random paths once their function space and metric are specified.

Theorem 1.2 (Donsker's Invariance Principle [13])**.** *Under the assumptions above, the process $(Z_t^{(n)})_{t\in[0,T]}$ converges in distribution to standard Brownian motion in the function space $C([0, T])$ equipped with the sup norm:*

$$Z^{(n)} \xrightarrow{d} B \quad \text{in} \quad (C([0, T]), \|\cdot\|_\infty). \tag{1.5}$$

Remark 1.3. This is the prototype for a diffusion approximation: after identifying the correct local drift, variance, and time scale, a discrete process may converge to an SDE. The same template reappears in the depth-scaling arguments of Chapter 5 and is formalized in Appendix A.3.

A different type of scaling limit appears in random matrix theory, where a high-dimensional random matrix has a deterministic limiting spectral distribution. Let $X \in \mathbb{R}^{n\times n}$ be symmetric, with the entries on and above the diagonal independent and identically distributed, centered, of variance one, and with finite fourth moment. Consider the scaled matrix $n^{-1/2}X$, with eigenvalues $\lambda_1^{(n)} \leq \cdots \leq \lambda_n^{(n)}$. Define the *empirical spectral distribution* (ESD)

$$\rho^{(n)}(dx) := \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i^{(n)}}(dx). \tag{1.6}$$

Theorem 1.4 (Wigner Semicircle Law [49])**.** *As $n \to \infty$, the random measure $\rho^{(n)}$ converges weakly almost surely to the deterministic probability measure with density*

$$\rho_{\mathrm{sc}}(x) = \frac{\sqrt{(4 - x^2)_+}}{2\pi}, \qquad \textit{supported on } [-2, 2]. \tag{1.7}$$

Remark 1.5. The emergence of a deterministic limiting spectrum is an example of self-averaging: macroscopic observables concentrate as dimension grows. In the random-projection model of Chapter 3, limiting spectra of design and Gram matrices determine explicit asymptotic risk formulas.

The final example replaces a large interacting particle system by a deterministic evolution of its empirical measure. Let $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ be a bounded Lipschitz interaction kernel. Consider n particles $(X_i(t))_{i=1}^n$ in $\mathbb{R}$ evolving according to the mean-field system

$$\partial_t X_i(t) = \frac{1}{n} \sum_{j=1}^{n} \mathcal{K}\big(X_i(t) - X_j(t)\big). \tag{1.8}$$

Define the empirical measure

$$\rho_t^{(n)}(dx) := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(t)}(dx). \tag{1.9}$$

For a probability measure ρ on $\mathbb{R}$, define the law-dependent velocity field

$$b(x, \rho) := \int_{\mathbb{R}} \mathcal{K}(x - y)\, \rho(dy). \tag{1.10}$$

Then the particle equation can be written exactly as $\partial_t X_i(t) = b(X_i(t), \rho_t^{(n)})$. With i.i.d. initial conditions of law ρ_0, standard mean-field theory shows that the empirical measure approaches a deterministic evolution (ρ_t) governed by

$$\partial_t \rho_t + \partial_x\big(b(\cdot, \rho_t)\rho_t\big) = 0, \qquad \rho_{t=0} = \rho_0, \tag{1.11}$$

The equation is interpreted in the weak sense; see Appendix A.1.

Although the particles interact at finite n, the independence present at initialization survives in the large-n limit: any fixed collection behaves like independent samples from ρ_t. This phenomenon is called *propagation of chaos* [48]. Quantitative versions bound finite-n errors for suitable observables; a leading error of order $n^{-\alpha}$ is a scaling law for the approach to the mean-field limit.

In mean-field analyses of neural networks, neurons or parameters play the role of particles. Their empirical distribution evolves under training and, at large width, is described by a McKean–Vlasov transport equation of this type. We return to this connection in Chapter 4.

Remark 1.6 (A Checklist for Scaling Limits). The three examples have the same four ingredients:

1. a sequence of finite random objects;
2. a normalization and a declaration of what scales and what remains fixed;
3. a topology and mode of convergence; and
4. a limiting object with a tractable description.

For the random walk the limit is a continuous path, for the Wigner matrix it is a probability measure, and for the particle system it is a deterministic evolution of probability measures. Later chapters repeat this template with network outputs, kernels, parameter distributions, and covariance processes. A quantitative scaling limit adds a further ingredient: a convergence rate or, more sharply, a leading finite-size correction for the observable of interest.

1.3 Roadmap

The remainder of the lecture notes is organized as follows.

Chapter 2 treats two related but logically distinct infinite-width limits for wide fully-connected networks. At random initialization, the outputs on any fixed finite collection of inputs

converge in distribution to a centered Gaussian vector characterized by the neural network Gaussian process (NNGP) kernel. During gradient-flow training in the NTK, or lazy-training, regime, the tangent kernel at each fixed training time converges to a deterministic kernel K that does not depend on time, reducing the limiting output dynamics to kernel gradient flow.

Chapter 3 turns from fitting the training data to prediction on new data. Using linear regression with random projections, it explains why test error can spike when the model first becomes able to fit the training data exactly and then decrease as more features are added.

Chapter 4 introduces *feature-learning* scalings in which representations evolve nontrivially during training, including shallow mean-field limits, the maximal-update (μP) parameterization, and a sketch of DMFT for deep networks.

Chapter 5 studies depth scalings and joint depth–width limits at initialization, including proportional limits and the neural covariance SDE viewpoint.

Chapter 6 discusses several applications of the depth scaling theory, including shaped attention/transformer models, limiting spectrum of the feature covariance, and a one-step feature-learning mechanism in the proportional limit.

Across these chapters, the principal regimes differ in what is held fixed and what is scaled:

- in Chapter 2, width $n \to \infty$ while depth d, input dimension n_0, and dataset size m remain fixed;
- in Chapter 3, $(m, n_0, n) \to \infty$ proportionally, with $n_0/m \to \gamma$ and $n/m \to \delta$;
- in Chapter 4, width tends to infinity under feature-learning parameterizations and associated learning-rate scalings, including both the shallow mean-field and deep maximal-update (μP) regimes; and
- in Chapters 5 and 6, depth and width can grow jointly, typically with a fixed finite collection of inputs before any additional spectral limit is taken.

The technical appendix collects the weak-convergence, Gaussian-conditioning, random-matrix, and diffusion-approximation tools used across these regimes.

Notation

Symbol	Meaning
n	Width of a hidden layer (number of neurons / channels).
d	Depth, or number of hidden layers or blocks, excluding the input layer.
n_0	Input dimension.
m	Number of inputs (training examples, or tokens in attention).
x^α	Input (data point) indexed by $\alpha \in \{1, \dots, m\}$.
y^α	Label / target associated with x^α.
θ	Network parameters (weights).
W_0, W_ℓ, W_d	For scalar-output MLPs, $W_0 \in \mathbb{R}^{n \times n_0}$, $W_\ell \in \mathbb{R}^{n \times n}$ for $1 \le \ell \le d-1$, and $W_d \in \mathbb{R}^{1 \times n}$ is the readout row vector.
$f(x^\alpha; \theta) = f^\alpha(\theta)$	Network output at input x^α.
φ	Activation function.
c	Variance-preserving normalization constant, $c := \left(\mathbb{E}[\varphi(w)^2]\right)^{-1}$ for $w \sim \mathcal{N}(0,1)$.
h_ℓ^α	Hidden pre-activations (or hidden states) at layer ℓ for input x^α.
$\varphi_\ell^\alpha = \varphi(h_\ell^\alpha)$, $\varphi_\ell = [\varphi_\ell^1 \ \cdots \ \varphi_\ell^m]$	Hidden post-activations at layer ℓ and their dataset-column matrix.
g_ℓ^α	Backward post-activation neurons at layer ℓ for input x^α.
z_ℓ^α	Backward pre-activation neurons at layer ℓ for input x^α.
$\Phi^{(n)}, \Phi_\ell^{(n)}, \Phi_\ell'^{(n)}$	Finite-width empirical feature and derivative kernels in Chapter 2, with hidden-layer normalization $1/n$ and input normalization $1/n_0$. The superscript (n) records width dependence.
$\Phi, \Phi_\ell, \Phi_\ell'$	Corresponding infinite-width kernels in Chapters 2 and 4. In Chapters 5 and 6, Φ_ℓ also denotes finite-width layer covariances.
$K_t^{(n)}, K_t, K$	Finite-width empirical NTK and limiting NTKs. Subscripts record time and superscripts in parentheses record width.
$L(\theta)$	Training objective (empirical risk); $L(\theta) = \frac{1}{m} \sum_{\alpha=1}^m \ell(f^\alpha(\theta), y^\alpha)$.
$\mathbb{E}, \mathbb{P}$	Expectation and probability.
$\|\cdot\|$, $\langle \cdot, \cdot \rangle$	Euclidean norm and inner product (unless specified otherwise).

Chapter 2

Kernel Limits for Wide Networks: NNGP and NTK

The first limit considered in this chapter is the infinite-width initialization limit. Neal showed that a one-hidden-layer network converges to a Gaussian process at random initialization [38]; Lee et al. and Matthews et al. extended this NNGP description to finite-depth networks [30, 36]. We then turn to a logically distinct training-time limit. In the NTK, or lazy-training, regime, the tangent kernel at each fixed training time converges to a deterministic kernel K that does not depend on time, so the limiting output dynamics reduce to kernel gradient flow [26, 31].

2.1 Neural Network Gaussian Processes (NNGP)

We first consider a one-hidden-layer network. For each hidden neuron, write $h_i(x) := n_0^{-1/2} \langle w_i, x \rangle$. At width n, its scalar output is

$$f(x;\theta) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} a_i \, \varphi(h_i(x)) \,, \qquad x \in \mathbb{R}^{n_0}, \tag{2.1}$$

where parameters are $\theta = \{(a_i, w_i)\}_{i=1}^{n}$. At initialization we assume

$$a_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{n_0}), \qquad \text{independent across } i \text{ and between } a_i \text{ and } w_i. \tag{2.2}$$

For fixed inputs $x^1, \ldots, x^m \in \mathbb{R}^{n_0}$, let $\mathcal{F} := \sigma(h_i(x^\alpha) : i \in [n],\ \alpha \in [m])$. Thus $\mathcal{F}$ is generated by the hidden neurons at the inputs under consideration and does not include the output weights. Since the Gaussian output weights are independent of $\mathcal{F}$, the output vector is exactly conditionally Gaussian:

$$\big(f(x^\alpha;\theta)\big)_{\alpha=1}^{m} \,|\, \mathcal{F} \sim \mathcal{N}(0, \Phi^{(n)}), \qquad \Phi^{(n),\alpha\beta} := \frac{1}{n} \sum_{i=1}^{n} \varphi(h_i(x^\alpha)) \, \varphi\Big(h_i(x^\beta)\Big) \,. \tag{2.3}$$

Thus no conditional central limit theorem is needed in the Gaussian-output-weight setting. This output-layer identity uses only Gaussian readout weights independent of the hidden neurons and will be reused below. If the relevant second moments are finite, the strong law of large numbers

gives

$$\Phi^{(n),\alpha\beta} \xrightarrow{\text{a.s.}} \Phi(x^\alpha, x^\beta), \qquad \Phi(x, x') := \mathbb{E}\left[\varphi\left(\frac{\langle w, x\rangle}{\sqrt{n_0}}\right)\varphi\left(\frac{\langle w, x'\rangle}{\sqrt{n_0}}\right)\right], \quad w \sim \mathcal{N}(0, I_{n_0}). \tag{2.4}$$

Taking conditional characteristic functions in (2.3) and using bounded convergence therefore yields

$$(f(x^1), \ldots, f(x^m)) \xrightarrow{d} \mathcal{N}\big(0, [\Phi(x^\alpha, x^\beta)]_{\alpha,\beta=1}^m\big). \tag{2.5}$$

For centered, unit-variance non-Gaussian output weights, the analogous conclusion instead follows from a multivariate central limit theorem.

Exercise 2.1 (Gaussianity of Linear Combinations)**.** Show that for any fixed m and any coefficients $u \in \mathbb{R}^m$, $\sum_{\alpha=1}^m u_\alpha f(x^\alpha) \xrightarrow{d} \mathcal{N}(0, u^\top \Phi u)$ with $\Phi = [\Phi(x^\alpha, x^\beta)]_{\alpha,\beta}$.

The calculation above establishes finite-dimensional convergence, which identifies the limiting Gaussian process on every fixed collection of inputs. Convergence of the random functions themselves additionally requires a function-space topology and tightness.

Definition 2.2 (Gaussian Process)**.** A random function $f : \mathbb{R}^{n_0} \to \mathbb{R}$ is a *Gaussian process* with mean μ and covariance kernel Φ (notation: $f \sim \mathrm{GP}(\mu, \Phi)$) if

(i) Φ is symmetric positive semidefinite, i.e. for any r and any $x^1, \ldots, x^r$, the matrix $[\Phi(x^\alpha, x^\beta)]_{\alpha,\beta=1}^r$ is PSD.

(ii) For any r and any $x^1, \ldots, x^r$,

$$(f(x^1), \ldots, f(x^r)) \sim \mathcal{N}\big((\mu(x^1), \ldots, \mu(x^r)), [\Phi(x^\alpha, x^\beta)]_{\alpha,\beta=1}^r\big). \tag{2.6}$$

Theorem 2.3 (Finite-Dimensional NNGP Limit at Initialization [38])**.** *Assume φ is Borel measurable and, for every $x \in \mathbb{R}^{n_0}$,*

$$\mathbb{E}\left[\varphi\left(\frac{\langle w, x\rangle}{\sqrt{n_0}}\right)^2\right] < \infty, \qquad w \sim \mathcal{N}(0, I_{n_0}). \tag{2.7}$$

Then, for every fixed finite collection $x^1, \ldots, x^m \in \mathbb{R}^{n_0}$,

$$\big(f(x^\alpha; \theta)\big)_{\alpha=1}^m \xrightarrow{d} \mathcal{N}\big(0, [\Phi(x^\alpha, x^\beta)]_{\alpha,\beta=1}^m\big), \tag{2.8}$$

where Φ is the kernel in (2.4). Equivalently, the finite-dimensional distributions converge to those of a centered Gaussian process with covariance kernel Φ.

Proof. The conditional Gaussian identity (2.3), the strong law in (2.4), and bounded convergence of the conditional characteristic functions give the asserted finite-dimensional convergence. Moreover, for every $u \in \mathbb{R}^m$,

$$\sum_{\alpha,\beta=1}^m u_\alpha u_\beta \Phi(x^\alpha, x^\beta) = \mathbb{E}\left[\left(\sum_{\alpha=1}^m u_\alpha \varphi\left(\frac{\langle w, x^\alpha\rangle}{\sqrt{n_0}}\right)\right)^2\right] \geq 0. \tag{2.9}$$

Thus Φ is positive semidefinite and defines the stated centered Gaussian process. □

Remark 2.4 (Function-Space Strengthening). The theorem asserts finite-dimensional convergence, which is the notion needed for a fixed finite dataset. Weak convergence in $C(K)$ for a compact $K \subset \mathbb{R}^{n_0}$ additionally requires tightness and a continuous version of the limit. For example, a globally Hölder activation yields the dimension-dependent increment bound proved in Appendix A.2; Kolmogorov–Chentsov then upgrades the finite-dimensional convergence to weak convergence in $C(K)$.

For some activations the kernel Φ admits a closed form. For ReLU, $\varphi(u) = \max\{u, 0\}$, the relevant Gaussian expectations have explicit expressions (the Cho–Saul / *arc-cosine* kernel).

Proposition 2.5 (Cho–Saul Kernel for ReLU)**.** *Let (h^α, h^β) be a centered Gaussian pair with*

$$\mathbb{E}[(h^\alpha)^2] = \Phi^{\alpha\alpha}, \qquad \mathbb{E}[(h^\beta)^2] = \Phi^{\beta\beta}, \qquad \mathbb{E}[h^\alpha h^\beta] = \Phi^{\alpha\beta}. \tag{2.10}$$

Assume $\Phi^{\alpha\alpha}, \Phi^{\beta\beta} > 0$ and define the correlation

$$\rho \coloneqq \frac{\Phi^{\alpha\beta}}{\sqrt{\Phi^{\alpha\alpha}\Phi^{\beta\beta}}} \in [-1, 1]. \tag{2.11}$$

Then

$$\mathbb{E}[\varphi(h^\alpha)\varphi(h^\beta)] = \frac{\sqrt{\Phi^{\alpha\alpha}\Phi^{\beta\beta}}}{2\pi}\left(\sqrt{1-\rho^2} + \rho\left(\frac{\pi}{2} + \arcsin(\rho)\right)\right), \tag{2.12}$$

$$\mathbb{E}[\varphi'(h^\alpha)\varphi'(h^\beta)] = \mathbb{P}(h^\alpha > 0,\ h^\beta > 0) = \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho). \tag{2.13}$$

Proof. By positive homogeneity of ReLU, it is enough to work with a jointly standard Gaussian pair (z_1, z_2) with correlation ρ and restore the marginal variances at the end. First suppose $|\rho| < 1$. Let w_1, w_2 be independent standard Gaussians and realize $z_1 = w_1$ and $z_2 = \rho w_1 + \sqrt{1-\rho^2}\, w_2$. Let f and F denote the standard Gaussian density and distribution function, and write $p(\rho) := \mathbb{P}(z_1 > 0, z_2 > 0)$. Conditioning on w_1 gives

$$p(\rho) = \int_0^\infty f(u) F\left(\frac{\rho u}{\sqrt{1-\rho^2}}\right) \mathrm{d}u. \tag{2.14}$$

Differentiation under the integral is justified on compact subintervals of $(-1, 1)$ by Gaussian domination, and yields

$$p'(\rho) = \frac{1}{2\pi(1-\rho^2)^{3/2}} \int_0^\infty u \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) \mathrm{d}u = \frac{1}{2\pi\sqrt{1-\rho^2}}. \tag{2.15}$$

Since $p(0) = 1/4$, integration gives the orthant probability stated in (2.13). Because $\varphi'(z) = \mathbb{1}\{z > 0\}$ almost everywhere, this proves the derivative-kernel formula for $|\rho| < 1$.

We now compute the expectation in (2.12). Let f_ρ be the joint density of (z_1, z_2). For dummy coordinates u, v, direct differentiation gives $u f_\rho(u, v) = -\partial_u f_\rho(u, v) - \rho\, \partial_v f_\rho(u, v)$. Multiply this identity by v and integrate over the positive quadrant. Gaussian decay removes the boundary terms at infinity, and the factor v removes the boundary term at $v = 0$. The u-integration leaves the boundary at $u = 0$; integration by parts in v leaves $p(\rho)$. Consequently,

$$\mathbb{E}[\varphi(z_1)\varphi(z_2)] = \int_0^\infty v f_\rho(0, v)\, \mathrm{d}v + \rho p(\rho) = \frac{\sqrt{1-\rho^2}}{2\pi} + \rho p(\rho). \tag{2.16}$$

The boundary integral is an elementary one-dimensional Gaussian integral. Substituting the orthant probability already obtained gives the bracketed expression in (2.12) for the standard pair.

Finally, $(h^\alpha, h^\beta) \overset{d}{=} (\sqrt{\Phi^{\alpha\alpha}}\, z_1, \sqrt{\Phi^{\beta\beta}}\, z_2)$. Positive homogeneity therefore multiplies the ReLU expectation by $\sqrt{\Phi^{\alpha\alpha}\Phi^{\beta\beta}}$, while positive rescaling leaves the signs unchanged. This proves both stated formulas for $|\rho| < 1$. The cases $\rho = \pm 1$ follow from the same coupling: bounded convergence applies to the indicator product, and the Lipschitz property of ReLU together with Gaussian second moments gives continuity of the ReLU-product expectation. □

Remark 2.6. Formulas (2.12)–(2.13) go back to Cho and Saul's *arc-cosine* kernels [11]. They are also the standard closed forms used to compute the limiting NNGP/NTK kernels for ReLU networks.

Remark 2.7 (Bayesian View). In the infinite-width limit, initialization induces a GP prior $f \sim \mathrm{GP}(0, \Phi)$. This motivates GP-style Bayesian inference as an approximation to Bayesian neural nets.

2.2 Extension to Deep Networks: Kernel Recursion

To extend the NNGP construction beyond one hidden layer, fix a depth $d \geq 2$ and width n for each hidden layer. For data points $\{x^\alpha\}_{\alpha=1}^m \subset \mathbb{R}^{n_0}$ define a multilayer perceptron (MLP) by

$$h_1^\alpha := \frac{1}{\sqrt{n_0}} W_0 x^\alpha, \qquad W_0 \in \mathbb{R}^{n \times n_0}, \tag{2.17}$$

$$h_{\ell+1}^\alpha := \frac{1}{\sqrt{n}} W_\ell\, \varphi(h_\ell^\alpha), \qquad W_\ell \in \mathbb{R}^{n \times n},\ \ell = 1, \ldots, d-1, \tag{2.18}$$

$$f(x^\alpha; \theta) := \frac{1}{\sqrt{n}} W_d\, \varphi(h_d^\alpha), \qquad W_d \in \mathbb{R}^{1 \times n}. \tag{2.19}$$

All weights are i.i.d. $\mathcal{N}(0,1)$ and independent across layers.

We will repeatedly use the following facts.

Proposition 2.8 (Linear Images of Gaussians). *If $g \sim \mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$, then $Ag \sim \mathcal{N}(A\mu, A\Sigma A^\top)$.*

Proposition 2.9 (Inner Products with a Standard Gaussian). *If $g \sim \mathcal{N}(0, I_n)$ and $u, v \in \mathbb{R}^n$, then*

$$\begin{pmatrix} \langle g, u \rangle \\ \langle g, v \rangle \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \|u\|^2 & \langle u, v \rangle \\ \langle u, v \rangle & \|v\|^2 \end{pmatrix}\right). \tag{2.20}$$

Proposition 2.10 (Multiplying by a Gaussian Matrix). *Let $W \in \mathbb{R}^{n \times n}$ have i.i.d. entries $W_{ij} \sim \mathcal{N}(0,1)$ and let $u, v \in \mathbb{R}^n$ be deterministic. Then*

$$\begin{pmatrix} Wu \\ Wv \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \|u\|^2 I_n & \langle u, v \rangle I_n \\ \langle u, v \rangle I_n & \|v\|^2 I_n \end{pmatrix}\right) = \mathcal{N}\left(0, \begin{pmatrix} \|u\|^2 & \langle u, v \rangle \\ \langle u, v \rangle & \|v\|^2 \end{pmatrix} \otimes I_n\right). \tag{2.21}$$

Definition 2.11 (Kronecker Product). If $A = [a_{ij}] \in \mathbb{R}^{p \times q}$ and $B \in \mathbb{R}^{r \times s}$, the *Kronecker product* $A \otimes B \in \mathbb{R}^{pr \times qs}$ is defined blockwise by

$$A \otimes B := [a_{ij} B]_{i,j}. \tag{2.22}$$

Define the deterministic input Gram matrix

$$\Phi_0^{\alpha\beta} := \frac{1}{n_0}\left\langle x^\alpha, x^\beta \right\rangle, \tag{2.23}$$

and set $\Phi_0^{(n)} := \Phi_0$. For $\ell \geq 1$, define the empirical feature covariance matrix by

$$\Phi_\ell^{(n),\alpha\beta} := \frac{1}{n}\left\langle \varphi(h_\ell^\alpha), \varphi(h_\ell^\beta) \right\rangle. \tag{2.24}$$

Let $\mathcal{F}_0$ be the trivial sigma-field, and for $\ell \geq 1$ define the filtration (history)

$$\mathcal{F}_\ell := \sigma\big(\{h_k^\alpha : \alpha \in [m],\ k \leq \ell\}\big). \tag{2.25}$$

Thus $\mathcal{F}_\ell$ records the hidden-neuron history through layer ℓ; by independence across layers, $W_\ell, \ldots, W_d$ are independent of $\mathcal{F}_\ell$. Conditioned on $\mathcal{F}_\ell$, the next-layer preactivations are Gaussian because $h_{\ell+1}^\alpha$ is a linear image of the Gaussian matrix W_ℓ. More precisely, for $\ell = 0, \ldots, d-1$, the m-tuple $(h_{\ell+1}^\alpha)_{\alpha=1}^m$ (each in $\mathbb{R}^n$) is conditionally Gaussian with covariance of the form

$$(h_{\ell+1}^\alpha)_{\alpha=1}^m \,|\, \mathcal{F}_\ell \ \sim\ \mathcal{N}(0,\, \Phi_\ell^{(n)} \otimes I_n). \tag{2.26}$$

The key observation is that $\Phi_{\ell+1}^{(n)}$ is, up to fluctuations that vanish with width, a deterministic function of $\Phi_\ell^{(n)}$. Indeed, by definition

$$\Phi_{\ell+1}^{(n),\alpha\beta} = \frac{1}{n}\left\langle \varphi(h_{\ell+1}^\alpha), \varphi(h_{\ell+1}^\beta) \right\rangle = \frac{1}{n}\sum_{j=1}^n \varphi(h_{\ell+1,j}^\alpha)\, \varphi(h_{\ell+1,j}^\beta). \tag{2.27}$$

Conditioned on $\mathcal{F}_\ell$, the coordinates $\{(h_{\ell+1,j}^\alpha)_{\alpha=1}^m\}_{j=1}^n$ are i.i.d. Gaussians with covariance $\Phi_\ell^{(n)}$ (this is the meaning of (2.26)). Under the assumptions of Theorem 2.13, for each fixed pair (α, β),

$$\Phi_{\ell+1}^{(n),\alpha\beta} - \mathcal{C}_\varphi(\Phi_\ell^{(n)})^{\alpha\beta} \xrightarrow{\mathbb{P}} 0, \tag{2.28}$$

where the deterministic covariance map $\mathcal{C}_\varphi$ is given entrywise by a Gaussian expectation:

$$\mathcal{C}_\varphi(\Phi)^{\alpha\beta} := \mathbb{E}\big[\varphi(g^\alpha)\, \varphi(g^\beta)\big], \qquad (g^1, \ldots, g^m) \sim \mathcal{N}(0, \Phi). \tag{2.29}$$

Define the deterministic forward kernels recursively by

$$\Phi_{\ell+1} := \mathcal{C}_\varphi(\Phi_\ell), \qquad \ell = 0, \ldots, d-1. \tag{2.30}$$

The theorem below makes precise that $\Phi_\ell^{(n)} \xrightarrow{\mathbb{P}} \Phi_\ell$ for each fixed $\ell \leq d$.

Remark 2.12 (Order of Limits). For deep networks, statements like "$n \to \infty$" implicitly hold depth fixed. If depth grows too, the order of limits matters: taking $n \to \infty$ first and then $d \to \infty$ can differ from a joint limit.

Theorem 2.13 (Deep NNGP Recursion [30, 36])**.** *Assume φ is continuous and has at most polynomial growth. Fix the depth d, the number of inputs m, and the inputs themselves. Then, as the width $n \to \infty$,*

$$\Phi_\ell^{(n)} \xrightarrow{\mathbb{P}} \Phi_\ell \qquad \textit{for each } \ell = 0, 1, \ldots, d, \tag{2.31}$$

where the deterministic matrices Φ_ℓ are defined by (2.30). Consequently, conditioned on the inputs, the output vector $\big(f(x^\alpha;\theta)\big)_{\alpha=1}^m$ converges in distribution to a centered Gaussian with covariance Φ_d.

Proof Sketch. The conditional law (2.26) implies that, given $\mathcal{F}_\ell$, the coordinates $\{(h^\alpha_{\ell+1,j})_{\alpha=1}^m\}_{j=1}^n$ are i.i.d. copies of a centered Gaussian vector in $\mathbb{R}^m$ with covariance $\Phi_\ell^{(n)}$. Therefore each entry of $\Phi_{\ell+1}^{(n)}$ is an empirical average of conditionally i.i.d. random variables,

$$\Phi_{\ell+1}^{(n),\alpha\beta} = \frac{1}{n}\sum_{j=1}^n \varphi(h^\alpha_{\ell+1,j})\varphi(h^\beta_{\ell+1,j}), \tag{2.32}$$

whose conditional expectation is $\mathcal{C}_\varphi(\Phi_\ell^{(n)})^{\alpha\beta}$. Its conditional centered second moment is

$$\mathbb{E}\left[\left|\Phi_{\ell+1}^{(n),\alpha\beta} - \mathcal{C}_\varphi(\Phi_\ell^{(n)})^{\alpha\beta}\right|^2 \,\middle|\, \mathcal{F}_\ell\right] = \frac{1}{n}\operatorname{Var}_{g\sim\mathcal{N}(0,\Phi_\ell^{(n)})}\left(\varphi(g^\alpha)\varphi(g^\beta)\right). \tag{2.33}$$

Polynomial growth and induction keep the conditional Gaussian moments on the right-hand side bounded in probability, so (2.28) follows. Continuity and polynomial growth also make $\mathcal{C}_\varphi$ continuous on bounded subsets of the positive semidefinite cone. Thus the continuous mapping theorem and induction over the fixed set of layers give $\Phi_\ell^{(n)} \xrightarrow{\mathbb{P}} \Phi_\ell$ entrywise and hence as an $m \times m$ matrix. Finally, since W_d is independent of $\mathcal{F}_d$, the output-layer identity (2.3) gives, conditional on $\mathcal{F}_d$, the exact Gaussian output law with covariance $\Phi_d^{(n)}$; convergence of this covariance gives the stated output limit. □

Remark 2.14 (Sequential versus Simultaneous Width Limits). The main derivation of Lee et al. [30] takes the hidden-layer widths to infinity successively, while their Appendix C gives an order-independent Gaussian-prior argument. Matthews et al. [36] give a rigorous simultaneous-width convergence theorem. In the Gaussian-weight, fixed-depth, finite-dataset setting used here, exact conditional Gaussianity leads directly to the common-width covariance recursion above.

2.3 Neural Tangent Kernel and Linearized Training

If we train the wide neural networks introduced above in the NTK regime, their output dynamics behave like those of *kernel methods*. The key object is the *Neural Tangent Kernel (NTK)*, a kernel induced by the network's tangent features. In the infinite-width limit, the NTK becomes (i) deterministic and (ii) constant in time along gradient flow, yielding closed-form linear training dynamics and, when the limiting kernel is positive definite, exponential convergence of the training loss. We then extend the construction from one-hidden-layer networks to finite-depth multilayer perceptrons (MLPs) and highlight the geometric interpretation as a preconditioned gradient flow in function space. We will reuse the forward-kernel notation and NNGP recursions from earlier sections.

The Neural Tangent Kernel was introduced by Jacot et al. [26]; Lee et al. [31] made explicit the associated linearization of wide-network training. A closely related line of work provides quantitative convergence guarantees for overparameterized deep networks; see, for instance, Allen-Zhu et al. [1], Du et al. [14], and Zou et al. [53]. In these notes we emphasize the "formal" infinite-width picture and use these works primarily as pointers to the rigorous theory.

Let the training dataset be

$$\{(x^\alpha, y^\alpha)\}_{\alpha=1}^m, \qquad x^\alpha \in \mathbb{R}^{n_0},\ y^\alpha \in \mathbb{R}. \tag{2.34}$$

Let $f(\cdot;\theta) : \mathbb{R}^{n_0} \to \mathbb{R}$ be a neural network with parameters θ. We write $f^\alpha(\theta) \coloneqq f(x^\alpha;\theta)$ and collect outputs into the vector

$$f(\theta) \coloneqq (f^1(\theta), \dots, f^m(\theta))^\top \in \mathbb{R}^m, \qquad y \coloneqq (y^1, \dots, y^m)^\top \in \mathbb{R}^m. \tag{2.35}$$

We will often use the empirical mean-squared error

$$L(\theta) \coloneqq \frac{1}{2m} \sum_{\alpha=1}^{m} \left(f^\alpha(\theta) - y^\alpha\right)^2 = \frac{1}{2m} \left\|f(\theta) - y\right\|_2^2. \tag{2.36}$$

Define the residual vector $r(\theta) \coloneqq f(\theta) - y$.

Gradient descent with step size η is

$$\theta_{k+1} = \theta_k - \eta \, \nabla_\theta L(\theta_k). \tag{2.37}$$

A convenient scaling limit is *gradient flow*:

$$\partial_t \theta(t) = -\nabla_\theta L(\theta(t)), \qquad \theta(0) = \theta_0, \tag{2.38}$$

where $\partial_t \theta(t) = \frac{\mathrm{d}}{\mathrm{d}t}\theta(t)$. In this chapter we analyze the induced dynamics of the network outputs $f(t) \coloneqq f(\theta(t))$.

We now encode the network's parameter sensitivities through its tangent feature map. For each input x, define the *tangent feature map* at parameters θ by

$$\nabla_\theta f(x;\theta) \in \mathbb{R}^{\dim(\theta)}. \tag{2.39}$$

For a width-n network, the (empirical) *Neural Tangent Kernel* at time t is the $m \times m$ matrix

$$K_t^{(n)} \in \mathbb{R}^{m\times m}, \qquad K_t^{(n),\alpha\beta} \coloneqq K_t^{(n)}(x^\alpha, x^\beta) \coloneqq \left\langle \nabla_\theta f^\alpha(\theta(t)), \nabla_\theta f^\beta(\theta(t)) \right\rangle. \tag{2.40}$$

By construction $K_t^{(n)}$ is symmetric positive semidefinite (PSD).

A chain-rule computation yields a closed-form ODE for $r(t)$ in terms of $K_t^{(n)}$.

Proposition 2.15 (Residual ODE)**.** *Under gradient flow for the squared loss, the residual vector $r(t) = f(t) - y$ satisfies*

$$\partial_t r(t) = -\frac{1}{m} K_t^{(n)} \, r(t). \tag{2.41}$$

Proof. For each α,

$$\partial_t f^\alpha(t) = \left\langle \nabla_\theta f^\alpha(\theta(t)), \partial_t \theta(t) \right\rangle = -\left\langle \nabla_\theta f^\alpha(\theta(t)), \nabla_\theta L(\theta(t)) \right\rangle. \tag{2.42}$$

For the squared loss, $\nabla_\theta L(\theta) = \frac{1}{m}\sum_{\beta=1}^m (f^\beta(\theta) - y^\beta)\, \nabla_\theta f^\beta(\theta)$. Substituting gives

$$\partial_t f^\alpha(t) = -\frac{1}{m} \sum_{\beta=1}^{m} \left\langle \nabla_\theta f^\alpha(\theta(t)), \nabla_\theta f^\beta(\theta(t)) \right\rangle \, (f^\beta(t) - y^\beta) = -\frac{1}{m} \sum_{\beta=1}^{m} K_t^{(n),\alpha\beta} \, r^\beta(t). \tag{2.43}$$

In vector form this is (2.41). □

Remark 2.16. Equation (2.41) shows that each residual coordinate $r^\alpha(t)$ mixes information from all residuals through the kernel matrix $K_t^{(n)}$. If $K_t^{(n)}$ were replaced by a *fixed* kernel matrix K, then training would become a linear ODE in r with an explicit solution via a matrix exponential. The main content of the NTK theory is that this replacement becomes accurate in the infinite-width limit.

This framework easily extends to general loss functions. For a differentiable pointwise loss $\ell(f, y)$, define the empirical risk

$$L(\theta) \coloneqq \frac{1}{m} \sum_{\alpha=1}^{m} \ell(f^\alpha(\theta), y^\alpha). \tag{2.44}$$

Along the training trajectory, define $r^\alpha(t) \coloneqq \partial_f \ell(f^\alpha(t), y^\alpha)$ and collect these entries in $r(t)$. For squared loss, this agrees with the residual $f(t) - y$ defined above.

Proposition 2.17 (Risk Dissipation Identity)**.** *Along gradient flow for L,*

$$\frac{\mathrm{d}}{\mathrm{d}t} L(\theta(t)) = -\frac{1}{m^2}\, r(t)^\top K_t^{(n)} r(t). \tag{2.45}$$

Proof. The chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t} L(\theta(t)) = \frac{1}{m} \sum_{\alpha=1}^{m} r^\alpha(t)\, \partial_t f^\alpha(t). \tag{2.46}$$

Moreover,

$$\nabla_\theta L(\theta(t)) = \frac{1}{m} \sum_{\beta=1}^{m} r^\beta(t)\, \nabla_\theta f^\beta(\theta(t)). \tag{2.47}$$

Substituting $\partial_t f^\alpha(t) = -\left\langle \nabla_\theta f^\alpha(\theta(t)), \nabla_\theta L(\theta(t)) \right\rangle$ and using (2.40) yields (2.45). □

2.4 The One-Layer Case

The NTK $K_t^{(n)}$ is random at initialization (because the parameters are random) and evolves during training (because $\theta(t)$ changes). The *NTK regime* is an infinite-width limit with two related features:

(i) Determinism: for every fixed training time t, the empirical kernel $K_t^{(n)}(x, x')$ concentrates around a deterministic kernel $K(x, x')$.

(ii) Stationarity: the limiting kernel K does not depend on t.

We start with the simplest case to illustrate the mechanism.

Definition 2.18 (One-Hidden-Layer MLP (NTK Parameterization))**.** Fix width n and input dimension n_0. Let

$$f(x; \theta) = \frac{1}{\sqrt{n}}\, a^\top \varphi\left(\frac{Wx}{\sqrt{n_0}}\right), \qquad W \in \mathbb{R}^{n \times n_0},\ a \in \mathbb{R}^n, \tag{2.48}$$

where $\varphi : \mathbb{R} \to \mathbb{R}$ is applied elementwise to vectors. We take i.i.d. Gaussian initialization $W_{ij} \sim \mathcal{N}(0, 1)$ and $a_i \sim \mathcal{N}(0, 1)$ independently.

Remark 2.19 (Bias Parameters). For notational simplicity we omit biases throughout this set of notes. Including bias terms does not change any of the core ideas (concentration of kernels, near-stationarity of the tangent kernel, etc.); it mainly makes the algebra longer and more tedious in an uninformative way.

Theorem 2.20 (One-Hidden-Layer NTK Limit (Informal))**.** *Consider the width-n network and NTK parameterization in* (2.48)*, with the Gaussian initialization above, trained by gradient flow for squared loss on a fixed finite dataset. Write $K_t^{(n)}$ for its empirical NTK at time t. For sufficiently nice φ, the standard NTK limit [26, 31] gives a deterministic symmetric positive semidefinite matrix K, independent of t, such that, for every fixed $t \geq 0$,*

$$K_t^{(n)} \xrightarrow{\mathbb{P}} K \qquad \text{as } n \to \infty. \tag{2.49}$$

The limiting residual dynamics are

$$\partial_t r(t) = -\frac{1}{m} K\, r(t). \tag{2.50}$$

Proof Sketch. Fix a finite dataset $\{x^\alpha\}_{\alpha=1}^m$.

Step 1: Determinism. For any parameters θ (and in particular at θ_0), the neuronwise gradients for a single datum x^α are

$$\frac{\partial f^\alpha}{\partial a_i} = \frac{1}{\sqrt{n}}\varphi(h_i^\alpha), \qquad \frac{\partial f^\alpha}{\partial w_i} = \frac{1}{\sqrt{nn_0}} a_i\, \varphi'(h_i^\alpha)\, x^\alpha, \qquad h_i^\alpha := \frac{1}{\sqrt{n_0}} w_i^\top x^\alpha. \tag{2.51}$$

Plugging (2.51) into $K_0^{(n),\alpha\beta} = \left\langle \nabla_\theta f^\alpha(\theta_0), \nabla_\theta f^\beta(\theta_0) \right\rangle$ yields the exact decomposition

$$K_0^{(n),\alpha\beta} = \frac{1}{n}\sum_{i=1}^n \kappa_i^{\alpha\beta}, \qquad \kappa_i^{\alpha\beta} := \varphi(h_i^\alpha)\varphi(h_i^\beta) + a_i^2\, \varphi'(h_i^\alpha)\varphi'(h_i^\beta)\, \frac{1}{n_0}\left\langle x^\alpha, x^\beta \right\rangle, \tag{2.52}$$

where $\kappa_i^{\alpha\beta}$ is evaluated at initialization $\theta = \theta_0$. Because (a_i, w_i) are i.i.d. at Gaussian initialization, for each fixed pair (α, β) the random variables $\{\kappa_i^{\alpha\beta}\}_{i=1}^n$ are i.i.d., and hence, provided $\mathbb{E}|\kappa_1^{\alpha\beta}| < \infty$, the weak law of large numbers gives $K_0^{(n),\alpha\beta} \xrightarrow{\mathbb{P}} K^{\alpha\beta}$ for each fixed (α, β). Since m is fixed, this entrywise convergence implies convergence in any matrix norm. Moreover, K is positive semidefinite because each $K_0^{(n)}$ is positive semidefinite and the positive-semidefinite cone is closed. The neuronwise form in (2.52) is convenient for the law-of-large-numbers argument and will also be used in the stationarity estimate below. To display separately the contributions of the two parameter blocks, write $h^\alpha = (h_1^\alpha, \ldots, h_n^\alpha)^\top$. Then the same identity becomes

$$\begin{aligned} K_0^{(n),\alpha\beta} &= \frac{1}{n}\left\langle \varphi(h^\alpha), \varphi(h^\beta) \right\rangle + \Phi_0^{\alpha\beta}\frac{1}{n}\left\langle a \odot \varphi'(h^\alpha), a \odot \varphi'(h^\beta) \right\rangle \\ &\xrightarrow{\mathbb{P}} K^{\alpha\beta} = \Phi_1^{\alpha\beta} + \Phi_0^{\alpha\beta}\Phi_1'^{,\alpha\beta}. \end{aligned} \tag{2.53}$$

Here $\Phi_1'^{,\alpha\beta} = \mathbb{E}[\varphi'(g^\alpha)\varphi'(g^\beta)]$ for a centered Gaussian pair with covariance $[\Phi_0^{\gamma\delta}]_{\gamma,\delta\in\{\alpha,\beta\}}$.

Step 2: Stationarity on a fixed time interval. Fix $t \geq 0$. For the one-hidden-layer model, gradient flow gives, for $0 \leq s \leq t$,

$$\partial_s a_i(s) = -\frac{1}{m\sqrt{n}}\sum_{\gamma=1}^m r^\gamma(s)\varphi(h_i^\gamma(s)), \qquad \partial_s w_i(s) = -\frac{a_i(s)}{m\sqrt{nn_0}}\sum_{\gamma=1}^m r^\gamma(s)\varphi'(h_i^\gamma(s))x^\gamma. \tag{2.54}$$

Along gradient flow, the decrease in loss equals the time integral of the squared parameter velocity. Cauchy–Schwarz therefore gives the empirical displacement bound

$$\sup_{0\le s\le t}\frac{1}{n}\sum_{i=1}^{n}\Big[(a_i(s)-a_i(0))^2+\|w_i(s)-w_i(0)\|_2^2\Big]\le\frac{tL(\theta_0)}{n}=O_{\mathbb{P}}(n^{-1}),\tag{2.55}$$

where $L(\theta_0)=O_{\mathbb{P}}(1)$ under the Gaussian initialization on the fixed dataset. Let $\kappa_i^{\alpha\beta}(s)$ denote the expression in (2.52) evaluated along the flow. A first-order Taylor expansion at initialization gives the next-order correction

$$\begin{aligned}K_t^{(n),\alpha\beta}-K_0^{(n),\alpha\beta}&=\frac{1}{n}\sum_{i=1}^{n}\Big(\kappa_i^{\alpha\beta}(t)-\kappa_i^{\alpha\beta}(0)\Big)\\&=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial\kappa_i^{\alpha\beta}}{\partial a_i}(0)\,(a_i(t)-a_i(0))+\Big\langle\nabla_{w_i}\kappa_i^{\alpha\beta}(0),w_i(t)-w_i(0)\Big\rangle\right]+o_{\mathbb{P}}(n^{-1/2}).\end{aligned}\tag{2.56}$$

The derivatives in the leading term are

$$\frac{\partial\kappa_i^{\alpha\beta}}{\partial a_i}=2a_i\varphi'(h_i^\alpha)\varphi'(h_i^\beta)\frac{\langle x^\alpha,x^\beta\rangle}{n_0},\tag{2.57}$$

$$\begin{aligned}\nabla_{w_i}\kappa_i^{\alpha\beta}=&\frac{x^\alpha}{\sqrt{n_0}}\left(\varphi'(h_i^\alpha)\varphi(h_i^\beta)+a_i^2\varphi''(h_i^\alpha)\varphi'(h_i^\beta)\frac{\langle x^\alpha,x^\beta\rangle}{n_0}\right)\\&+\frac{x^\beta}{\sqrt{n_0}}\left(\varphi(h_i^\alpha)\varphi'(h_i^\beta)+a_i^2\varphi'(h_i^\alpha)\varphi''(h_i^\beta)\frac{\langle x^\alpha,x^\beta\rangle}{n_0}\right).\end{aligned}\tag{2.58}$$

For sufficiently nice φ, these derivatives have bounded empirical second moments at initialization. Cauchy–Schwarz and (2.55) therefore bound the averaged linear Taylor term by $O_{\mathbb{P}}(n^{-1/2})$, uniformly on the fixed interval. The smoothness and moment assumptions also control the Taylor remainder uniformly, giving the same bound for the full correction in (2.56). Combining this with Step 1 gives $K_s^{(n)}\xrightarrow{\mathbb{P}}K$ uniformly for $0\le s\le t$. The same estimate in the time-integral form of (2.41), together with the NNGP limit for the initial residual, gives the limiting residual dynamics stated above. □

Remark 2.21 (Scope of the Proof Sketch). The calculation above displays the one-hidden-layer mechanism. For multiple hidden layers, an additional layerwise argument is required; we omit those details here. Nonsmooth activations such as ReLU require a different argument.

We now study the limiting fixed-kernel dynamics, in which the time-dependent empirical kernel in (2.41) is replaced by a fixed positive-definite matrix K. The resulting linear ODE has an explicit solution and converges exponentially.

Proposition 2.22 (Exponential Convergence of the Fixed-Kernel Dynamics)**.** *Let $K\in\mathbb{R}^{m\times m}$ be symmetric positive definite, and suppose $\partial_t r(t)=-m^{-1}Kr(t)$. Then*

$$r(t)=\exp\left(-\frac{t}{m}K\right)r(0),\tag{2.59}$$

and the squared loss $L(t) \coloneqq (2m)^{-1} \|r(t)\|_2^2$ *satisfies*

$$L(t) \leq \exp\left(-\frac{2\lambda_{\min}(K)}{m} t\right) L(0). \tag{2.60}$$

Thus the loss converges to zero exponentially.

Proof. By the spectral theorem, $K = P \operatorname{diag}(\lambda_1, \ldots, \lambda_m) P^\top$ for an orthogonal matrix P and eigenvalues $\lambda_i > 0$. The matrix exponential in (2.59) solves the constant-coefficient residual ODE. Moreover,

$$\|r(t)\|_2^2 = \sum_{i=1}^{m} e^{-2\lambda_i t/m} \left((P^\top r(0))_i\right)^2 \leq e^{-2\lambda_{\min}(K) t/m} \|r(0)\|_2^2 . \tag{2.61}$$

Dividing by $2m$ proves (2.60). □

Remark 2.23 (When Is $\lambda_{\min}(K) > 0$?). For a fixed dataset, positive definiteness is a property of the deterministic limiting kernel K. If $\lambda_{\min}(K) > 0$, then $K_0^{(n)} \xrightarrow{\mathbb{P}} K$ implies

$$\mathbb{P}\left(K_0^{(n)} \text{ is positive definite}\right) \longrightarrow 1. \tag{2.62}$$

When the data are random, positivity and conditioning of K can themselves be studied probabilistically. See, for example, Nguyen et al. [39] for quantitative lower bounds on the smallest eigenvalue of empirical NTK matrices for deep ReLU networks under standard assumptions.

Remark 2.24 (Linearization and the Kernel Regression Viewpoint). The NTK regime can be viewed heuristically as a linearization of the network around initialization. Consider gradient descent with step size η: $\theta_{k+1} = \theta_k - \eta \nabla_\theta L(\theta_k)$. Heuristically, the usual NTK width counting gives

$$f^\alpha(\theta_{k+1}) = f^\alpha(\theta_k) + \langle \nabla_\theta f^\alpha(\theta_k), \theta_{k+1} - \theta_k \rangle + \underbrace{\mathcal{O}(n^{-1/2})}_{\text{nonlinear correction}} . \tag{2.63}$$

Here the $\mathcal{O}(n^{-1/2})$ term is schematic; it summarizes the leading nonlinear correction. Discarding it and freezing the tangent features at initialization gives

$$f^\alpha(\theta) \approx f^\alpha(\theta_0) + \langle \nabla_\theta f^\alpha(\theta_0), \theta - \theta_0 \rangle . \tag{2.64}$$

Hence training becomes approximately linear regression in the tangent features $\nabla_\theta f(x^\alpha; \theta_0)$. Rigorous NTK results make this heuristic precise along the training trajectory [31].

A complementary geometric viewpoint is already visible in (2.41). When K is positive definite, equip the output space with the inner product

$$\langle u, v \rangle_{K^{-1}} \coloneqq u^\top K^{-1} v. \tag{2.65}$$

With respect to this inner product, the fixed-kernel dynamics become

$$\partial_t f = -\frac{1}{m} K(f - y) = -\operatorname{grad}_{K^{-1}} \left(\frac{1}{2m} \|f - y\|_2^2\right). \tag{2.66}$$

On the fixed dataset, the space of predictor values is simply the linear vector space $\mathbb{R}^m$. Nevertheless, K^{-1} is precisely a constant, and hence flat, Riemannian metric on this space; equivalently, K acts as its inverse metric or as a preconditioner.

2.5 Extension to Finite-Depth MLPs

We now describe the exact layerwise structure of the NTK for a depth-d fully-connected network and then compute its deterministic infinite-width limit at Gaussian initialization.

The finite-depth NTK is built from three layerwise quantities: the empirical forward kernels $\Phi_\ell^{(n)}$, the empirical derivative kernels $\Phi_\ell'^{(n)}$, and the backward covariances $G_\ell^{(n)}$. At Gaussian initialization, the deterministic forward- and derivative-kernel limits feed into a backward recursion for the limits of the backward covariances. Substituting these limits into the exact layerwise decomposition in Proposition 2.25 then yields the limiting NTK. Throughout this section, the superscript (n) denotes finite width, and common training-time arguments are suppressed.

To analyze the finite-depth NTK, we use the following network notation. Let n be the hidden width and n_0 the input dimension. For $\ell = 0, 1, \dots, d$, let W_ℓ be weight matrices with i.i.d. $\mathcal{N}(0,1)$ entries, where $W_0 \in \mathbb{R}^{n\times n_0}$, $W_\ell \in \mathbb{R}^{n\times n}$ for $1 \le \ell \le d-1$, and $W_d \in \mathbb{R}^{1\times n}$ (scalar output). Define for input x^α the pre-activations

$$h_1^\alpha = \frac{1}{\sqrt{n_0}} W_0 x^\alpha, \qquad h_{\ell+1}^\alpha = \frac{1}{\sqrt{n}} W_\ell \varphi(h_\ell^\alpha) \;\; (1 \le \ell \le d-1), \tag{2.67}$$

and output

$$f^\alpha = f(x^\alpha;\theta) = \frac{1}{\sqrt{n}} W_d\, \varphi(h_d^\alpha). \tag{2.68}$$

We first record the forward and derivative kernels, reusing the forward-kernel notation and deterministic recursion from Section 2.2. Define the *input (zeroth-layer) kernel*

$$\Phi_0^{\alpha\beta} \coloneqq \frac{1}{n_0} \left\langle x^\alpha, x^\beta \right\rangle, \tag{2.69}$$

Recall that $\Phi_0^{(n)} = \Phi_0$. For $\ell \ge 1$, define the *empirical forward kernels*

$$\Phi_\ell^{(n),\alpha\beta} \coloneqq \frac{1}{n} \left\langle \varphi(h_\ell^\alpha), \varphi(h_\ell^\beta) \right\rangle, \quad \Phi_\ell'^{(n),\alpha\beta} \coloneqq \frac{1}{n} \left\langle \varphi'(h_\ell^\alpha), \varphi'(h_\ell^\beta) \right\rangle. \tag{2.70}$$

Theorem 2.13 gives $\Phi_\ell^{(n)} \xrightarrow{\mathbb{P}} \Phi_\ell$. Suppose additionally that φ' exists almost everywhere, its Gaussian-product second moments are uniformly bounded for covariance matrices in a neighborhood of the deterministic recursion, and the corresponding expectations are continuous there. Then the same conditional second-moment argument gives $\Phi_\ell'^{(n)} \xrightarrow{\mathbb{P}} \Phi_\ell'$.

For $\ell \ge 0$, let (h^α, h^β) be the centered Gaussian pair whose covariance is determined by Φ_ℓ. The derivative-kernel limit obeys $\Phi_{\ell+1}'^{,\alpha\beta} = \mathbb{E}[\varphi'(h^\alpha)\varphi'(h^\beta)]$. We write this compactly as $\Phi_{\ell+1}' = \mathcal{C}_{\varphi'}(\Phi_\ell)$, where $\mathcal{C}_{\varphi'}$ denotes the companion Gaussian expectation rather than the derivative of $\mathcal{C}_\varphi$.

For ReLU, the Cho–Saul formulas in Proposition 2.5 make the forward and derivative-kernel recursions explicit. For fixed ℓ, α, β with $\Phi_\ell^{\alpha\alpha}, \Phi_\ell^{\beta\beta} > 0$, write $\rho \coloneqq \Phi_\ell^{\alpha\beta} / \sqrt{\Phi_\ell^{\alpha\alpha}\Phi_\ell^{\beta\beta}} \in [-1,1]$. Then

$$\begin{aligned} \Phi_{\ell+1}^{\alpha\beta} &= \frac{\sqrt{\Phi_\ell^{\alpha\alpha}\Phi_\ell^{\beta\beta}}}{2\pi} \left(\sqrt{1-\rho^2} + \rho \left(\frac{\pi}{2} + \arcsin(\rho) \right) \right), \\ \Phi_{\ell+1}'^{,\alpha\beta} &= \frac{1}{4} + \frac{1}{2\pi} \arcsin(\rho). \end{aligned} \tag{2.71}$$

Thus the same correlation determines both the next forward kernel and the derivative factor that enters the backward recursion.

We next introduce the backpropagated sensitivities whose covariances complete the NTK decomposition. Fix a training time t and evaluate all forward and backward quantities below at $\theta(t)$, suppressing their common time argument except on

$$K_t^{(n),\alpha\beta} \coloneqq \left\langle \nabla_\theta f^\alpha(\theta(t)), \nabla_\theta f^\beta(\theta(t)) \right\rangle. \tag{2.72}$$

The additional state variables in the decomposition are the empirical backward covariances $G_\ell^{(n),\alpha\beta}$, which encode inner products of backpropagated sensitivities. At Gaussian initialization, the empirical forward kernels $\Phi_\ell^{(n),\alpha\beta}$ converge to the deterministic NNGP kernels $\Phi_\ell^{\alpha\beta}$ defined earlier. Define the *backward sensitivity vectors*

$$g_\ell^\alpha \;\coloneqq\; \sqrt{n}\,\frac{\partial f^\alpha}{\partial h_\ell^\alpha} \in \mathbb{R}^n, \qquad 1 \le \ell \le d. \tag{2.73}$$

A direct backpropagation calculation gives the exact top-layer identity and hidden-layer recursion

$$g_d^\alpha = W_d^\top \odot \varphi'(h_d^\alpha), \tag{2.74}$$

$$g_\ell^\alpha = \frac{1}{\sqrt{n}}\,\mathrm{diag}(\varphi'(h_\ell^\alpha))\, W_\ell^\top g_{\ell+1}^\alpha, \qquad \ell = d-1, \dots, 1. \tag{2.75}$$

Define the *backward covariance* matrices

$$G_\ell^{(n),\alpha\beta} \coloneqq \frac{1}{n}\left\langle g_\ell^\alpha, g_\ell^\beta \right\rangle, \qquad \ell = 1, \dots, d, \qquad G_{d+1}^{(n),\alpha\beta} \coloneqq 1. \tag{2.76}$$

Proposition 2.25 (Layerwise NTK Decomposition)**.** *For the depth-d MLP defined above (with no biases), at every training time t the empirical NTK on the dataset $\{x^\alpha\}_{\alpha=1}^m$ decomposes exactly as*

$$K_t^{(n),\alpha\beta} = \sum_{\ell=0}^{d} G_{\ell+1}^{(n),\alpha\beta}\, \Phi_\ell^{(n),\alpha\beta}, \tag{2.77}$$

where $\Phi_0^{(n)} = \Phi_0$, and $G_{d+1}^{(n),\alpha\beta} = 1$ corresponds to the top (linear) output layer.

In matrix form, viewing $\Phi_\ell^{(n)}, G_{\ell+1}^{(n)}, K_t^{(n)} \in \mathbb{R}^{m\times m}$, (2.77) reads

$$K_t^{(n)} = \sum_{\ell=0}^{d} G_{\ell+1}^{(n)} \odot \Phi_\ell^{(n)}, \tag{2.78}$$

where $\odot$ denotes the Hadamard (entrywise) product.

Proof. The NTK is an inner product of parameter gradients: $K_t^{(n),\alpha\beta} = \left\langle \nabla_\theta f^\alpha(\theta(t)), \nabla_\theta f^\beta(\theta(t)) \right\rangle$. Split θ layer-by-layer and sum the resulting contributions. The definition of g_ℓ^α and the network scaling give the boundary-aware identities

$$\begin{aligned} \frac{\partial f^\alpha}{\partial W_0} &= \frac{1}{\sqrt{n n_0}}\, g_1^\alpha(x^\alpha)^\top, \\ \frac{\partial f^\alpha}{\partial W_\ell} &= \frac{1}{n}\, g_{\ell+1}^\alpha(\varphi(h_\ell^\alpha))^\top, \quad 1 \le \ell \le d-1, \\ \frac{\partial f^\alpha}{\partial W_d} &= \frac{1}{\sqrt{n}}\,(\varphi(h_d^\alpha))^\top. \end{aligned} \tag{2.79}$$

Their Frobenius inner products are therefore

$$\begin{aligned}\left\langle \frac{\partial f^\alpha}{\partial W_0}, \frac{\partial f^\beta}{\partial W_0} \right\rangle_F &= G_1^{(n),\alpha\beta} \Phi_0^{\alpha\beta}, \\ \left\langle \frac{\partial f^\alpha}{\partial W_\ell}, \frac{\partial f^\beta}{\partial W_\ell} \right\rangle_F &= G_{\ell+1}^{(n),\alpha\beta} \Phi_\ell^{(n),\alpha\beta}, \quad 1 \le \ell \le d-1, \\ \left\langle \frac{\partial f^\alpha}{\partial W_d}, \frac{\partial f^\beta}{\partial W_d} \right\rangle_F &= \Phi_d^{(n),\alpha\beta}.\end{aligned} \tag{2.80}$$

Since $\Phi_0^{(n)} = \Phi_0$ and $G_{d+1}^{(n),\alpha\beta} = 1$, summing these contributions yields (2.77). □

To pass this decomposition to the infinite-width limit, we must handle the dependence created by reusing the same weights in the forward and backward passes. Conditioning on the forward activations breaks the i.i.d. structure of those weights; the Gaussian conditioning identity below isolates a usable independent component.

Write $\varphi_\ell^\alpha \coloneqq \varphi(h_\ell^\alpha)$ and collect the dataset postactivations into

$$\varphi_\ell \coloneqq [\varphi_\ell^1 \ \cdots \ \varphi_\ell^m] \in \mathbb{R}^{n \times m}. \tag{2.81}$$

Since $h_{\ell+1}^\alpha = \frac{1}{\sqrt{n}} W_\ell \varphi_\ell^\alpha$, conditioning on the next-layer preactivations is (up to the deterministic scaling $1/\sqrt{n}$) the same as conditioning on $W_\ell \varphi_\ell$.

Lemma 2.26 (Gaussian Conditioning (One-Sided))**.** *Let $W \in \mathbb{R}^{n \times p}$ have i.i.d. $\mathcal{N}(0,1)$ entries and let $\varphi \in \mathbb{R}^{p \times q}$ be deterministic. Let $P_\varphi \coloneqq \varphi(\varphi^\top \varphi)^\dagger \varphi^\top$ be the orthogonal projector onto $\mathrm{col}(\varphi)$, and $P_\varphi^\perp \coloneqq I_p - P_\varphi$. Then, conditioning on the σ-field generated by $W\varphi$,*

$$W \,|\, \sigma(W\varphi) \ \overset{d}{=}\ W P_\varphi \ +\ \widetilde{W} P_\varphi^\perp, \tag{2.82}$$

where $\widetilde{W}$ is an independent copy of W.

Proof. See Appendix A.5.

Proposition 2.27 (Backward Covariance Recursion)**.** *Fix m, d, n_0 and the inputs, and evaluate all quantities at Gaussian initialization. For sufficiently nice φ, there are deterministic matrices $G_1, \ldots, G_d$ such that, entrywise,*

$$G_\ell^{(n)} \xrightarrow{\mathbb{P}} G_\ell, \qquad \ell = 1, \ldots, d. \tag{2.83}$$

With $G_{d+1}^{\alpha\beta} \coloneqq 1$, their entries satisfy the backward recursion

$$G_\ell = \Phi_\ell' \odot G_{\ell+1}, \qquad \ell = d, d-1, \ldots, 1, \tag{2.84}$$

where $\odot$ denotes the Hadamard (entrywise) product. Consequently,

$$K_0^{(n)} \xrightarrow{\mathbb{P}} K \coloneqq \sum_{\ell=0}^{d} G_{\ell+1} \odot \Phi_\ell \tag{2.85}$$

entrywise.

Proof Sketch. First consider the top hidden layer. The exact identity (2.74) gives

$$G_d^{(n),\alpha\beta} = \frac{1}{n}\sum_{i=1}^{n} W_{d,i}^2 \varphi'(h_{d,i}^{\alpha})\varphi'(h_{d,i}^{\beta}). \tag{2.86}$$

Since W_d is independent of $\mathcal{F}_d$ and $\mathbb{E}[W_{d,i}^2] = 1$, the conditional law-of-large-numbers calculation gives $G_d^{(n),\alpha\beta} \xrightarrow{\mathbb{P}} \Phi_d'^{,\alpha\beta}$. Thus $G_d = \Phi_d'$.

For $1 \le \ell \le d-1$, take inner products in (2.75) to obtain

$$G_\ell^{(n),\alpha\beta} = \frac{1}{n^2}\,(g_{\ell+1}^{\alpha})^\top W_\ell \mathrm{diag}\big(\varphi'(h_\ell^{\alpha}) \odot \varphi'(h_\ell^{\beta})\big) W_\ell^\top g_{\ell+1}^{\beta}. \tag{2.87}$$

At the formal level of this proof sketch, Gaussian conditioning suggests that the leading contribution is obtained by replacing W_ℓ in this quadratic form by a fresh Gaussian matrix $\widetilde{W}_\ell$. Conditional on the displayed g's and h's, the elementary Gaussian quadratic-form identity gives

$$\begin{aligned}&\mathbb{E}_{\widetilde{W}_\ell}\left[\frac{1}{n^2}\,(g_{\ell+1}^{\alpha})^\top \widetilde{W}_\ell \mathrm{diag}\big(\varphi'(h_\ell^{\alpha}) \odot \varphi'(h_\ell^{\beta})\big)\widetilde{W}_\ell^\top g_{\ell+1}^{\beta}\right] \\ &\qquad = G_{\ell+1}^{(n),\alpha\beta}\Phi_\ell'^{(n),\alpha\beta}.\end{aligned} \tag{2.88}$$

The normalized quadratic form concentrates around this value, so

$$G_\ell^{(n)} - \Phi_\ell'^{(n)} \odot G_{\ell+1}^{(n)} \xrightarrow{\mathbb{P}} 0 \tag{2.89}$$

entrywise. Passing to the limit and proceeding backward through the finitely many layers gives (2.84). Finally, substituting the forward and backward limits into the exact decomposition (2.77), with $t = 0$, gives (2.85). □

Remark 2.28 (Rigor of the Proof Sketch). The sketch suppresses the estimates needed to justify the fresh-Gaussian replacement and the concentration of the normalized quadratic forms. At fixed depth, these steps can be made rigorous by applying Gaussian conditioning recursively. The formal calculation above is the part that determines the limiting recursion.

Remark 2.29 (Other Architectures). The same strategy—identify the relevant "state" variables, use Gaussian conditioning (Lemma 2.26) to recover effective independence, then close a self-consistent recursion—extends beyond fully-connected networks (e.g. certain convolutional nets, residual networks in suitable limits, and some attention-like blocks). The details depend strongly on the architecture.

We have so far focused on optimization and training dynamics in the kernel regime. In the infinite-width limit, when the limiting kernel is positive definite, the squared training loss converges exponentially to zero. This does not address the statistical properties of the trained predictor, particularly its generalization error on unseen data. We turn to this question in the next chapter.

2.6 Exercises

1. Residual ODE. Derive Proposition 2.15 directly from the chain rule.

2. General Loss Dissipation. Prove Proposition 2.17. Then pass formally to the fixed-kernel limit and show that, for any differentiable pointwise loss (not necessarily squared) and any fixed symmetric positive semidefinite matrix K,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{m}\sum_{\alpha=1}^{m}\ell(f^{\alpha}(t),y^{\alpha})\right]=-\frac{1}{m^{2}}r(t)^{\top}Kr(t)\leq 0. \tag{2.90}$$

3. Fixed-Kernel Dynamics. Verify Proposition 2.22 directly from the spectral theorem.
4. One-Hidden-Layer NTK Formula. At time t, write $a_t := a(t)$, $h_t(x) := W(t)x/\sqrt{n_0}$, and $f_t(x) = n^{-1/2}a_t^{\top}\varphi(h_t(x))$. Show that the NTK admits the decomposition

$$\begin{aligned}K_t^{(n)}(x,x') = &\underbrace{\frac{1}{n}\left\langle\varphi(h_t(x)),\varphi(h_t(x'))\right\rangle}_{\text{output-weight contribution}}\\ &+\underbrace{\frac{\langle x,x'\rangle}{n_0}\frac{1}{n}\left\langle a_t\odot\varphi'(h_t(x)),a_t\odot\varphi'(h_t(x'))\right\rangle}_{\text{feature-gradient term}}.\end{aligned} \tag{2.91}$$

5. Deep NTK Recursion. From the definition of g_{ℓ}^{α}, derive the three boundary-aware identities in (2.79) and prove the exact decomposition (2.77). Verify the top-layer law-of-large-numbers calculation and the fresh-Gaussian expectation in the proof of Proposition 2.27. Use these calculations formally to obtain (2.89), and then complete the backward induction yielding $G_{\ell}^{(n)} \xrightarrow{\mathbb{P}} G_{\ell}$ and $K_0^{(n)} \xrightarrow{\mathbb{P}} K$.

Chapter 3

Double Descent with Random Projections

3.1 Motivation

One of the central puzzles in deep learning is the following: modern neural networks are often heavily overparameterized, so in principle they can interpolate (and even memorize) the training data, yet in practice they can still generalize well on unseen data. This tension between expressive power and test performance motivated a large line of work on the *interpolation regime*. Double descent is a first clean window into this phenomenon [4]: in overparameterized linear and random-feature models, it shows explicitly how prediction risk can worsen near interpolation and then improve again beyond it, offering a concrete baseline mechanism for why "more parameters" need not imply worse generalization.

An illustrative non-isotropic example is shown in Figure 3.1, where the risk first decreases, spikes near interpolation, and then decreases again.

A tractable way to study this behavior is through a random-feature model: the weights defining the hidden features are sampled once and then frozen, while only the readout coefficients are fitted by least squares. Conditional on the frozen features, training is therefore linear regression. We specialize to a linear activation, for which the feature matrix is a random projection of the input matrix.

Although simple, this model retains the rank mechanisms behind double descent. If the ambient dimension exceeds the sample size, increasing the number of random features crosses an interpolation threshold; if the ambient dimension is smaller, the random features eventually span the input space and further increases in width do not change the fitted function. In the isotropic Gaussian setting, rotational invariance gives explicit bias–variance formulas across these rank regimes. Bach's general-covariance analysis is our starting point [3].

3.2 Model: Linear Regression with Random Projections

We observe

$$y = X\theta_\star + \varepsilon, \qquad X \in \mathbb{R}^{m\times n_0}, \;\; X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_m), \tag{3.1}$$

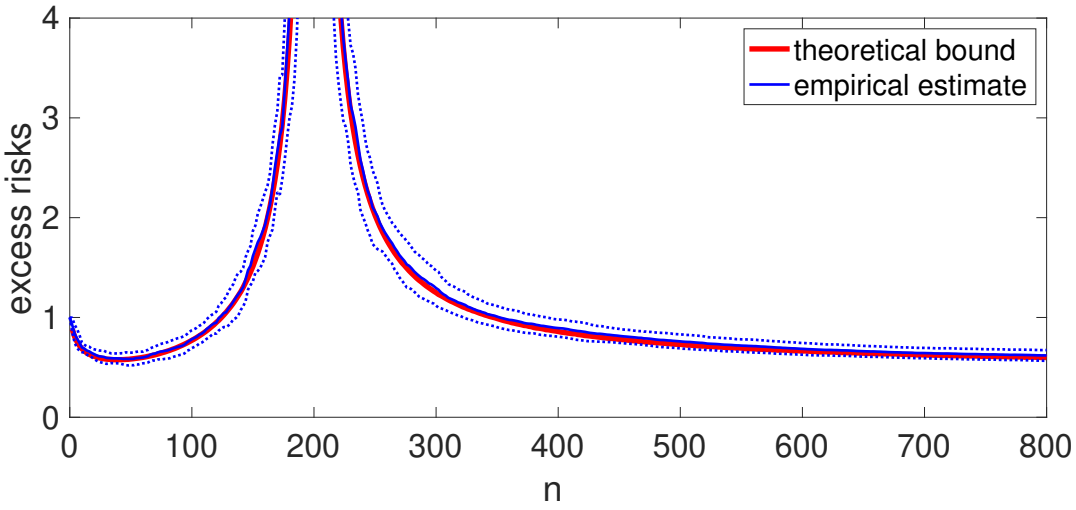

Figure 3.1: A non-isotropic double descent example for linear regression with random projections, with $m = 200$ observations and ambient dimension $n_0 = 400$. Adapted from [3, Figure 1]; the horizontal axis has been relabeled by n, our notation for the number of random projections. The isotropic model below exhibits the interpolation spike and second descent when the ambient dimension exceeds the sample size, but not the initial descent visible here.

where $\theta_\star \in \mathbb{R}^{n_0}$ is the unknown parameter, σ^2 is the noise variance, and X is independent of ε.

Let $S \in \mathbb{R}^{n_0 \times n}$ be a random projection matrix with i.i.d. entries

$$S_{jk} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad S \text{ independent of } (X, \varepsilon). \tag{3.2}$$

We fit a linear predictor in the *projected* feature space $x \mapsto S^\top x$. The projected design matrix, minimum-norm feature coefficient, and induced ambient coefficient are

$$\begin{aligned} Z &\coloneqq XS \in \mathbb{R}^{m \times n}, \\ \hat{\eta} &\coloneqq Z^\dagger y = \left(Z^\top Z\right)^\dagger Z^\top y \in \mathbb{R}^n, \qquad \hat{\theta} \coloneqq S\hat{\eta} \in \mathbb{R}^{n_0}. \end{aligned} \tag{3.3}$$

where $\dagger$ denotes the Moore–Penrose pseudoinverse.

Thus m, n_0, and n denote the number of observations, the ambient dimension, and the number of random projections, respectively. This differs from Bach's notation: there, n denotes the number of observations, m the number of random projections, and d the ambient dimension [3].

We measure parameter estimation error by

$$R(\hat{\theta}) \coloneqq \left\|\hat{\theta} - \theta_\star\right\|^2. \tag{3.4}$$

For an independent test covariate $x_{\text{new}} \sim \mathcal{N}(0, I_{n_0})$, this also equals the excess prediction error $\mathbb{E}_{x_{\text{new}}}[(x_{\text{new}}^\top(\hat{\theta} - \theta_\star))^2]$. For a general covariance $\Sigma = \mathbb{E}[xx^\top]$, prediction error instead equals $(\hat{\theta} - \theta_\star)^\top \Sigma (\hat{\theta} - \theta_\star)$ and need not coincide with $R(\hat{\theta})$. Since $\hat{\theta} = My$ for

$$M \coloneqq SZ^\dagger = S\left(Z^\top Z\right)^\dagger Z^\top, \tag{3.5}$$

substituting $y = X\theta_\star + \varepsilon$ gives the bias–variance decomposition

$$\mathbb{E}_\varepsilon\Big[R(\hat{\theta}) \mid X, S\Big] = \|(MX - I_{n_0})\theta_\star\|^2 + \mathbb{E}_\varepsilon \|M\varepsilon\|^2 = R^{(\text{bias})}(\hat{\theta}) + R^{(\text{var})}(\hat{\theta}), \tag{3.6}$$

where

$$R^{(\text{var})}(\hat{\theta}) = \sigma^2 \operatorname{Tr}(M^\top M). \tag{3.7}$$

3.3 High-Dimensional Risk and Double Descent

We now state the asymptotic behavior of the conditional risk $\mathbb{E}_\varepsilon\Big[R(\hat{\theta}) \mid X, S\Big]$ (equivalently the conditional bias–variance terms from (3.6)) in the proportional limit

$$m, n_0, n \to \infty \qquad \text{with} \qquad \frac{n_0}{m} \to \gamma \in (0, \infty), \quad \frac{n}{m} \to \delta \in (0, \infty). \tag{3.8}$$

The following theorem gives the three rank regimes for the Gaussian isotropic model. The feature- and sample-bottleneck formulas are the $\Sigma = I_{n_0}$ specialization of [3, Proposition 4]; the ambient-bottleneck formula follows directly from ordinary least squares once the random features span $\mathbb{R}^{n_0}$.

Theorem 3.1 (Asymptotic Bias and Variance for Random Projections (Isotropic Case))**.** *Assume the isotropic Gaussian model* (3.1)–(3.2) *and consider the ridgeless estimator* (3.3)*. In the proportional limit* (3.8)*, assume that* $\|\theta_\star\|^2$ *converges to a finite value as* $n_0 \to \infty$*. Then, in each of the following regimes, the conditional bias–variance terms in* (3.6) *converge in probability, with respect to the randomness of* (X, S)*, to the displayed deterministic limits.*

(a) Feature-bottleneck regime ($\delta < \min\{1, \gamma\}$*):*

$$R^{(\text{var})}(\hat{\theta}) \;\longrightarrow\; \sigma^2 \frac{\delta}{1-\delta}, \tag{3.9}$$

$$R^{(\text{bias})}(\hat{\theta}) \;\longrightarrow\; \frac{\gamma - \delta}{\gamma(1-\delta)} \|\theta_\star\|^2 . \tag{3.10}$$

(b) Ambient-bottleneck regime ($\gamma < 1$ *and* $\delta \geq \gamma$*):*

$$R^{(\text{var})}(\hat{\theta}) \;\longrightarrow\; \sigma^2 \frac{\gamma}{1-\gamma}, \tag{3.11}$$

$$R^{(\text{bias})}(\hat{\theta}) \;\longrightarrow\; 0. \tag{3.12}$$

(c) Sample-bottleneck regime ($\gamma > 1$ *and* $\delta > 1$*):*

$$R^{(\text{var})}(\hat{\theta}) \;\longrightarrow\; \sigma^2 \left(\frac{1}{\gamma - 1} + \frac{1}{\delta - 1}\right), \tag{3.13}$$

$$R^{(\text{bias})}(\hat{\theta}) \;\longrightarrow\; \left(1 - \frac{1}{\gamma}\right) \frac{\delta}{\delta - 1} \|\theta_\star\|^2 . \tag{3.14}$$

The theorem excludes the singular boundaries $\delta = 1$ for $\gamma > 1$ and $\gamma = 1$ for $\delta \geq 1$.

The theorem is most easily read by fixing γ and increasing the width ratio δ. When $\gamma > 1$, the feature-bottleneck formulas apply for $\delta < 1$, up to the boundary. Provided $\sigma^2 > 0$ or the limiting value of $\|\theta_\star\|^2$ is positive, the total risk is strictly increasing and diverges as $\delta \uparrow 1$. The closing

dimension gap $m - n$ makes the random design nearly square and produces the interpolation spike. For $\delta > 1$, the sample-bottleneck formulas apply, and the part of the risk above its infinite-width limit decreases proportionally to $(\delta - 1)^{-1}$ toward $\sigma^2/(\gamma - 1) + (1 - 1/\gamma)\|\theta_\star\|^2$ as $\delta \to \infty$. This is the second descent. Thus the isotropic model captures the interpolation spike and the post-interpolation descent, but not the initial descent in Figure 3.1.

When $\gamma < 1$, the width reaches the ambient dimension first, at $\delta = \gamma < 1$. The feature-bottleneck formulas then meet the ambient-bottleneck formulas continuously: the bias vanishes, and the risk remains at the ordinary least-squares value $\sigma^2\gamma/(1 - \gamma)$ for all $\delta \geq \gamma$. Consequently there is no singularity at $\delta = 1$; before ambient saturation, the risk may increase, decrease, or remain flat depending on the signal-to-noise ratio. In this way, γ determines whether increasing width first reaches the interpolation threshold or the ambient dimension, while δ determines the active bottleneck.

3.3.1 Gaussian Derivation of the Risk Theorem

We now give the formal calculation behind each branch of Theorem 3.1. The linear-algebra identities and conditional Gaussian expectations are written out explicitly. We leave implicit only the standard concentration estimates for inverse-Wishart traces, Gaussian quadratic forms, and projections onto Haar subspaces. In the strict proportional regimes, these estimates upgrade the displayed conditional-mean calculations to convergence of the corresponding random quantities.

Lemma 3.2 (Gaussian Gram-Matrix Inputs)**.** *Let $G \in \mathbb{R}^{p\times q}$ have i.i.d. $\mathcal{N}(0,1)$ entries. The inverse-Wishart identities are*

$$\begin{aligned} \mathbb{E}\Big[(G^\top G)^{-1}\Big] &= \frac{I_q}{p-q-1}, \quad p > q+1, \\ \mathbb{E}\Big[(GG^\top)^{-1}\Big] &= \frac{I_p}{q-p-1}, \quad q > p+1. \end{aligned} \tag{3.15}$$

Consequently, as $p, q \to \infty$, away from the square boundary,

$$\begin{aligned} \mathrm{Tr}\Big((G^\top G)^{-1}\Big) &\xrightarrow{\mathbb{P}} \frac{\rho}{1-\rho}, \quad \frac{q}{p} \to \rho < 1, \\ \mathrm{Tr}\Big((GG^\top)^{-1}\Big) &\xrightarrow{\mathbb{P}} \frac{1}{\rho-1}, \quad \frac{q}{p} \to \rho > 1. \end{aligned} \tag{3.16}$$

Proof. The identities in (3.15) are the standard first inverse moments of a Wishart matrix, and their proportional trace limits follow from the corresponding inverse-Wishart concentration. □

Lemma 3.3 (Directional Inverse-Wishart Law)**.** *If $q > p$, $G \in \mathbb{R}^{p\times q}$ has i.i.d. $\mathcal{N}(0,1)$ entries, and $a \in \mathbb{R}^p$ is deterministic, then*

$$a^\top (GG^\top)^{-1} a \overset{d}{=} \frac{\|a\|^2}{\chi^2_{q-p+1}}, \tag{3.17}$$

with the zero-vector case interpreted directly.

Proof. For $a \neq 0$, rotational invariance reduces to $a = \|a\| e_1$. Writing the first row of G as g_1 and the remaining rows as G_2, the Schur-complement identity gives

$$\Big[(GG^\top)^{-1}\Big]^{-1}_{11} = g_1 \Big(I_q - G_2^\top (G_2 G_2^\top)^{-1} G_2\Big) g_1^\top. \tag{3.18}$$

The matrix in parentheses is an orthogonal projector of rank $q-p+1$, independent of g_1, so the right-hand side has the χ^2_{q-p+1} distribution. □

Lemma 3.4 (Haar Projection Lengths)**.** *Let P be the orthogonal projector onto a Haar-distributed r-dimensional subspace of $\mathbb{R}^d$. For every deterministic $v \in \mathbb{R}^d$,*

$$\mathbb{E}\Big[\|Pv\|^2\Big] = \frac{r}{d}\|v\|^2 . \tag{3.19}$$

If $d \to \infty$ and $r/d \to \rho$, then $\|Pv\|^2 - (r/d)\|v\|^2 \xrightarrow{\mathbb{P}} 0$ for every bounded-norm deterministic sequence v.

Proof. For $v = 0$, both conclusions are immediate, so suppose $v \neq 0$. Rotating v to the first coordinate and representing a uniform direction by a normalized standard Gaussian vector shows that $\|Pv\|^2 / \|v\|^2$ has the same law as $\chi^2_r/(\chi^2_r + \widetilde{\chi}^2_{d-r})$, where the two chi-squared variables are independent. The mean follows by symmetry. If $\rho \in (0,1)$, the convergence follows from the law of large numbers for the two chi-squared variables. If $\rho = 0$, it follows from the mean identity and Markov's inequality; if $\rho = 1$, apply the same argument to $I_d - P$. □

All $o_{\mathbb{P}}(1)$ terms below refer to the joint randomness of (X, S). The intermediate conditional expectations are proof devices and are distinct from the label-noise expectation in (3.6).

Proposition 3.5 (Feature-Bottleneck Calculation)**.** *Under the assumptions of Theorem 3.1, if $\delta < \min\{1, \gamma\}$, then, in probability,*

$$\begin{aligned} R^{(\mathrm{var})}(\hat{\theta}) &\longrightarrow \sigma^2 \frac{\delta}{1-\delta}, \\ R^{(\mathrm{bias})}(\hat{\theta}) &\longrightarrow \frac{\gamma-\delta}{\gamma(1-\delta)}\|\theta_\star\|^2 . \end{aligned} \tag{3.20}$$

Proof. For all sufficiently large dimensions, $n < \min\{m, n_0\}$. Write the thin QR factorization

$$S = QT, \qquad Q^\top Q = I_n, \tag{3.21}$$

where $T \in \mathbb{R}^{n\times n}$ is invertible almost surely. Since $Z = (XQ)T$ has full column rank, the least-squares coefficient and its mapped ambient parameter satisfy

$$\begin{aligned} \hat{\eta} &= \Big(T^\top (XQ)^\top (XQ) T\Big)^{-1} T^\top (XQ)^\top y \\ &= T^{-1}\Big((XQ)^\top XQ\Big)^{-1}(XQ)^\top y, \\ \hat{\theta} &= Q\Big((XQ)^\top XQ\Big)^{-1}(XQ)^\top y. \end{aligned} \tag{3.22}$$

Thus the singular values of S cancel: in this regime the mapped estimator $\hat{\theta}$ depends on S only through its column space.

Set

$$P_S \coloneqq QQ^\top, \qquad \theta_\perp \coloneqq (I_{n_0} - P_S)\theta_\star. \tag{3.23}$$

The response decomposes as

$$y = (XQ)Q^\top \theta_\star + X\theta_\perp + \varepsilon, \tag{3.24}$$

Conditional on Q, the matrix XQ is standard Gaussian and $X\theta_\perp$ is independent of XQ, with distribution $\mathcal{N}(0, \|\theta_\perp\|^2 I_m)$. Substituting (3.24) into (3.22) gives

$$\hat{\theta} = P_S\theta_\star + Q\left((XQ)^\top XQ\right)^{-1}(XQ)^\top X\theta_\perp + Q\left((XQ)^\top XQ\right)^{-1}(XQ)^\top\varepsilon. \tag{3.25}$$

The last term is the label-noise contribution. Using $Q^\top Q = I_n$ and cyclicity of trace, its conditional variance is

$$R^{(\mathrm{var})}(\hat{\theta}) = \sigma^2 \operatorname{Tr}\Big((XQ)\left((XQ)^\top XQ\right)^{-2}(XQ)^\top\Big) = \sigma^2 \operatorname{Tr}\Big(\left((XQ)^\top XQ\right)^{-1}\Big). \tag{3.26}$$

After averaging only over the label noise and subtracting $\theta_\star$, the bias vector is $-\theta_\perp + Q((XQ)^\top XQ)^{-1}(XQ)^\top X\theta_\perp$. These two terms lie in the orthogonal subspaces $\mathrm{span}(Q)^\perp$ and $\mathrm{span}(Q)$, respectively, so the conditional squared bias is exactly

$$R^{(\mathrm{bias})}(\hat{\theta}) = \|\theta_\perp\|^2 + (X\theta_\perp)^\top(XQ)\left((XQ)^\top XQ\right)^{-2}(XQ)^\top X\theta_\perp. \tag{3.27}$$

Conditional on (Q, XQ), the expectation of the second term over the independent Gaussian vector $X\theta_\perp$ is

$$\begin{aligned}\mathbb{E}\Big[(X\theta_\perp)^\top(XQ)\left((XQ)^\top XQ\right)^{-2}(XQ)^\top X\theta_\perp \,\Big|\, Q, XQ\Big] \\ = \|\theta_\perp\|^2 \operatorname{Tr}\Big(\left((XQ)^\top XQ\right)^{-1}\Big).\end{aligned} \tag{3.28}$$

Gaussian quadratic-form concentration therefore yields

$$R^{(\mathrm{bias})}(\hat{\theta}) = \|\theta_\perp\|^2\left[1 + \operatorname{Tr}\Big(((XQ)^\top XQ)^{-1}\Big)\right] + o_{\mathbb{P}}(1). \tag{3.29}$$

The exact fully averaged identities provide a finite-dimensional check:

$$\begin{aligned}\mathbb{E}_{X,S}\Big[R^{(\mathrm{var})}(\hat{\theta})\Big] &= \sigma^2\frac{n}{m-n-1}, \\ \mathbb{E}_{X,S}\Big[R^{(\mathrm{bias})}(\hat{\theta})\Big] &= \left(1-\frac{n}{n_0}\right)\frac{m-1}{m-n-1}\|\theta_\star\|^2,\end{aligned} \tag{3.30}$$

whenever $n \le n_0$ and $m > n+1$. Indeed, (3.19) gives the omitted-signal factor $1 - n/n_0$, while (3.15) gives the amplification factor $1 + n/(m-n-1) = (m-1)/(m-n-1)$.

Here $XQ \in \mathbb{R}^{m\times n}$ is standard Gaussian, while $I_{n_0} - P_S$ projects onto a Haar-distributed (n_0-n)-dimensional subspace. Since $n/m \to \delta$ and $(n_0-n)/n_0 \to 1-\delta/\gamma$, applying Lemmas 3.2 and 3.4 gives

$$\begin{aligned}\operatorname{Tr}\Big(((XQ)^\top XQ)^{-1}\Big) &\xrightarrow{\mathbb{P}} \frac{\delta}{1-\delta}, \\ \|\theta_\perp\|^2 &\xrightarrow{\mathbb{P}} \left(1-\frac{\delta}{\gamma}\right)\|\theta_\star\|^2.\end{aligned} \tag{3.31}$$

The first limit in (3.31), combined with (3.26), gives the variance limit in (3.20). Since $1+\delta/(1-\delta) = (1-\delta)^{-1}$, substitution in (3.29) gives the bias limit. The same inverse-Wishart factor therefore amplifies both label noise and the omitted signal. □

Proposition 3.6 (Ambient-Bottleneck Calculation)**.** *Under the assumptions of Theorem 3.1, if* $\gamma < 1$ *and* $\delta > \gamma$*, then, in probability,*

$$\begin{aligned} R^{(\mathrm{var})}(\hat{\theta}) &\longrightarrow \sigma^2 \frac{\gamma}{1-\gamma}, \\ R^{(\mathrm{bias})}(\hat{\theta}) &\longrightarrow 0. \end{aligned} \tag{3.32}$$

Proof. Eventually $n_0 < m$ and $n \geq n_0$. The matrices X and S then have full column and full row rank, respectively, almost surely. The least-squares coefficient $\hat{\eta} = Z^{\dagger} y$ satisfies the normal equations

$$S^{\top} X^{\top} \left(X S \hat{\eta} - y \right) = 0. \tag{3.33}$$

Because $S^{\top}$ is injective and $\hat{\theta} = S\hat{\eta}$, these equations imply

$$X^{\top} \Big(X\hat{\theta} - y \Big) = 0, \qquad \hat{\theta} = (X^{\top} X)^{-1} X^{\top} y. \tag{3.34}$$

Thus the particular minimum-feature-norm solution maps to the unique ambient ordinary least-squares solution. Substituting the model gives

$$\hat{\theta} - \theta_{\star} = (X^{\top} X)^{-1} X^{\top} \varepsilon. \tag{3.35}$$

Hence the conditional bias vanishes and

$$R^{(\mathrm{var})}(\hat{\theta}) = \sigma^2 \operatorname{Tr}\Big((X^{\top} X)^{-1} X^{\top} X (X^{\top} X)^{-1} \Big) = \sigma^2 \operatorname{Tr}\Big((X^{\top} X)^{-1} \Big). \tag{3.36}$$

In particular,

$$\mathbb{E}_X \Big[R^{(\mathrm{var})}(\hat{\theta}) \Big] = \sigma^2 \frac{n_0}{m - n_0 - 1}, \tag{3.37}$$

whenever $m > n_0 + 1$. The tall-matrix case of Lemma 3.2 now gives $R^{(\mathrm{var})}(\hat{\theta}) \xrightarrow{\mathbb{P}} \sigma^2 \gamma/(1-\gamma)$, proving (3.32). □

Lemma 3.7 (Gaussian Row–Null Coordinates)**.** *Suppose* X *has full row rank, with* S *independent of* X *and distributed as in* (3.2)*. Set* $U_X \coloneqq X^{\top}(XX^{\top})^{-1/2}$ *and choose* $U_{\perp} \in \mathbb{R}^{n_0 \times (n_0 - m)}$ *with orthonormal columns spanning* $\ker(X)$*. Conditional on* X*, the matrices* $H \coloneqq U_X^{\top} S$ *and* $G_{\perp} \coloneqq U_{\perp}^{\top} S$ *are independent standard Gaussian matrices.*

Proof. The matrix U_X has orthonormal columns and is orthogonal to $U_{\perp}$. Indeed,

$$\begin{bmatrix} U_X & U_{\perp} \end{bmatrix}^{\top} \begin{bmatrix} U_X & U_{\perp} \end{bmatrix} = I_{n_0} \tag{3.38}$$

so $[U_X \;\; U_{\perp}]$ is orthogonal. Conditional on X, applying its transpose to each standard Gaussian column of S gives independent standard Gaussian row and null coordinates. □

Proposition 3.8 (Sample-Bottleneck Calculation)**.** *Under the assumptions of Theorem 3.1, if* $\gamma > 1$ *and* $\delta > 1$*, then, in probability,*

$$\begin{aligned} R^{(\mathrm{var})}(\hat{\theta}) &\longrightarrow \sigma^2 \left(\frac{1}{\gamma - 1} + \frac{1}{\delta - 1} \right), \\ R^{(\mathrm{bias})}(\hat{\theta}) &\longrightarrow \left(1 - \frac{1}{\gamma} \right) \frac{\delta}{\delta - 1} \left\| \theta_{\star} \right\|^2. \end{aligned} \tag{3.39}$$

Proof. Eventually $m < \min\{n_0, n\}$, so Z has full row rank and

$$Z^\dagger = Z^\top (ZZ^\top)^{-1}. \tag{3.40}$$

Use the matrices $U_X, U_\perp, H, G_\perp$ from Lemma 3.7. Since $[U_X \;\; U_\perp]$ is orthogonal and $H = (XX^\top)^{-1/2}Z$, projecting S onto these two coordinate blocks gives

$$\begin{aligned} S &= X^\top (XX^\top)^{-1} Z + U_\perp G_\perp, \\ Z &= (XX^\top)^{1/2} H. \end{aligned} \tag{3.41}$$

Substituting (3.41) into $M = SZ^\dagger$ and using $ZZ^\dagger = I_m$ gives

$$M = X^\top (XX^\top)^{-1} + U_\perp G_\perp Z^\dagger. \tag{3.42}$$

The two summands take values in the row space and null space of X, respectively, and are therefore orthogonal as linear maps into $\mathbb{R}^{n_0}$. Consequently, $R^{(\mathrm{var})}(\hat{\theta}) = \sigma^2 \left\|M\right\|_{\mathrm{F}}^2$, where

$$\left\|M\right\|_{\mathrm{F}}^2 = \mathrm{Tr}\Big((XX^\top)^{-1}\Big) + \left\|G_\perp Z^\dagger\right\|_{\mathrm{F}}^2. \tag{3.43}$$

Because Z is an invertible transform of H given X, conditioning on (X, Z) fixes H and leaves $G_\perp$ standard Gaussian. Thus its $n_0 - m$ rows are independent standard Gaussian vectors and $(Z^\dagger)^\top Z^\dagger = (ZZ^\top)^{-1}$. The conditional mean and variance of the second term in (3.43) are therefore

$$\begin{aligned} \mathbb{E}\left[\left\|G_\perp Z^\dagger\right\|_{\mathrm{F}}^2 \,\middle|\, X, Z\right] &= (n_0 - m)\,\mathrm{Tr}\Big((ZZ^\top)^{-1}\Big), \\ \mathrm{Var}\left(\left\|G_\perp Z^\dagger\right\|_{\mathrm{F}}^2 \,\middle|\, X, Z\right) &= 2(n_0 - m)\,\mathrm{Tr}\Big((ZZ^\top)^{-2}\Big). \end{aligned} \tag{3.44}$$

The displayed conditional variance is $o_{\mathbb{P}}(1)$ in the strict proportional regime, so the squared norm concentrates around the conditional mean.

The whitening in (3.41) also gives

$$\mathrm{Tr}\Big((ZZ^\top)^{-1}\Big) = \mathrm{Tr}\Big((XX^\top)^{-1}(HH^\top)^{-1}\Big). \tag{3.45}$$

Conditional on X, (3.15) yields

$$\mathbb{E}\Big[\mathrm{Tr}\Big((ZZ^\top)^{-1}\Big) \,\Big|\, X\Big] = \frac{1}{n - m - 1}\,\mathrm{Tr}\Big((XX^\top)^{-1}\Big). \tag{3.46}$$

Standard weighted inverse-Wishart concentration upgrades this conditional mean to

$$(n - m)\,\mathrm{Tr}\Big((ZZ^\top)^{-1}\Big) = \mathrm{Tr}\Big((XX^\top)^{-1}\Big) + o_{\mathbb{P}}(1). \tag{3.47}$$

Combining (3.43), (3.44), and (3.47) gives

$$R^{(\mathrm{var})}(\hat{\theta}) = \sigma^2 \left(1 + \frac{n_0 - m}{n - m}\right) \mathrm{Tr}\Big((XX^\top)^{-1}\Big) + o_{\mathbb{P}}(1). \tag{3.48}$$

Since

$$\mathrm{Tr}\Big((XX^\top)^{-1}\Big) \xrightarrow{\mathbb{P}} \frac{1}{\gamma - 1}, \qquad \frac{n_0 - m}{n - m} \longrightarrow \frac{\gamma - 1}{\delta - 1}, \tag{3.49}$$

the two terms in (3.48) give $\sigma^2/(\gamma-1)$ and $\sigma^2/(\delta-1)$, proving the variance limit in (3.39).

For the bias, define the row-space projector

$$P_X \coloneqq X^\top (XX^\top)^{-1} X. \tag{3.50}$$

Set $y = X\theta_\star$ and define

$$b \coloneqq Z^\top (ZZ^\top)^{-1} X\theta_\star, \qquad q \coloneqq \|b\|^2 = (X\theta_\star)^\top (ZZ^\top)^{-1} X\theta_\star. \tag{3.51}$$

Applying (3.42) to $X\theta_\star$ gives the exact error decomposition

$$\hat{\theta} - \theta_\star = -(I_{n_0} - P_X)\theta_\star + U_\perp G_\perp b. \tag{3.52}$$

Unlike the two estimator terms in (3.42), the two error terms in (3.52) both lie in $\ker(X)$ and need not be orthogonal. Conditional on (X, Z),

$$U_\perp G_\perp b \sim \mathcal{N}(0, q(I_{n_0} - P_X)). \tag{3.53}$$

Writing $v \coloneqq (I_{n_0} - P_X)\theta_\star$, the exact conditional moments are

$$\begin{aligned}
\mathbb{E}[\langle v, U_\perp G_\perp b\rangle \mid X, Z] &= 0, \\
\mathrm{Var}(\langle v, U_\perp G_\perp b\rangle \mid X, Z) &= q\,\|v\|^2, \\
\mathbb{E}\Big[\|U_\perp G_\perp b\|^2 \,\Big|\, X, Z\Big] &= (n_0 - m)q, \\
\mathrm{Var}\Big(\|U_\perp G_\perp b\|^2 \,\Big|\, X, Z\Big) &= 2(n_0 - m)q^2.
\end{aligned} \tag{3.54}$$

To calculate q, set

$$a \coloneqq (XX^\top)^{-1/2} X\theta_\star. \tag{3.55}$$

Then (3.41) gives

$$q = a^\top (HH^\top)^{-1} a, \qquad \|a\|^2 = \|P_X\theta_\star\|^2. \tag{3.56}$$

By Lemma 3.3, conditionally on X,

$$q \overset{d}{=} \frac{\|P_X\theta_\star\|^2}{\chi^2_{n-m+1}}, \tag{3.57}$$

with the zero-signal case interpreted directly. Thus $q = O_{\mathbb{P}}(m^{-1})$ and

$$(n-m)q = \|P_X\theta_\star\|^2 + o_{\mathbb{P}}(1). \tag{3.58}$$

The conditional variances in (3.54) now vanish. Expanding the squared norm in (3.52), the centered cross term is therefore $o_{\mathbb{P}}(1)$ and the correction norm concentrates around $(n_0 - m)q$. Using (3.58) gives

$$R^{(\mathrm{bias})}(\hat{\theta}) = \|(I_{n_0} - P_X)\theta_\star\|^2 + \frac{n_0 - m}{n - m}\|P_X\theta_\star\|^2 + o_{\mathbb{P}}(1). \tag{3.59}$$

The exact fully averaged risks in this regime are

$$\begin{aligned}
\mathbb{E}_{X,S}\Big[R^{(\mathrm{var})}(\hat{\theta})\Big] &= \sigma^2 \frac{m}{n_0 - m - 1}\left(1 + \frac{n_0 - m}{n - m - 1}\right), \\
\mathbb{E}_{X,S}\Big[R^{(\mathrm{bias})}(\hat{\theta})\Big] &= \frac{n_0 - m}{n_0}\left(1 + \frac{m}{n - m - 1}\right)\|\theta_\star\|^2,
\end{aligned} \tag{3.60}$$

whenever $n_0 > m+1$ and $n > m+1$.

Finally, the row space of X is Haar distributed, so Lemma 3.4 gives

$$\|P_X\theta_\star\|^2 \xrightarrow{\mathbb{P}} \frac{1}{\gamma}\|\theta_\star\|^2, \qquad \|(I_{n_0}-P_X)\theta_\star\|^2 \xrightarrow{\mathbb{P}} \left(1-\frac{1}{\gamma}\right)\|\theta_\star\|^2. \tag{3.61}$$

Substituting these limits and $(n_0-m)/(n-m) \to (\gamma-1)/(\delta-1)$ into (3.59) yields

$$R^{(\mathrm{bias})}(\hat{\theta}) \xrightarrow{\mathbb{P}} \left(1-\frac{1}{\gamma}+\frac{\gamma-1}{\gamma(\delta-1)}\right)\|\theta_\star\|^2 = \left(1-\frac{1}{\gamma}\right)\frac{\delta}{\delta-1}\|\theta_\star\|^2, \tag{3.62}$$

which proves the bias limit in (3.39). □

Proof of Theorem 3.1. Propositions 3.5, 3.6 and 3.8 prove the three strict regimes. It remains only to consider $\delta = \gamma < 1$. Along indices with $n \geq n_0$, the ambient calculation applies. Along indices with $n < n_0$, the finite-dimensional feature calculation remains valid. Since $n/n_0 \to 1$ and $n/m \to \gamma < 1$, the omitted-subspace bias vanishes, while the inverse-Wishart trace converges to $\gamma/(1-\gamma)$. Every subsequence therefore has the same ambient-bottleneck limit. The two remaining boundaries excluded in the theorem are precisely the square inverse-Wishart hard edges. □

The calculations have three main consequences:

- The smallest of m, n_0, and n determines whether features, ambient dimension, or samples are the bottleneck.
- The singular factors are rectangular Gaussian dimension gaps: when $\gamma > 1$ the feature gap closes at $\delta = 1$, whereas for $\gamma < 1$ the risk saturates once the features span $\mathbb{R}^{n_0}$.
- The isotropic model isolates the interpolation spike and second descent. Non-isotropic covariance can also produce the earlier descent shown in Figure 3.1.

The random projection model is linear, but it isolates mechanisms that recur in nonlinear networks: fixed-feature behavior, rank geometry, and sharp transitions near interpolation. The fixed-feature perspective connects back to the NNGP/NTK kernel limits of Chapter 2; Chapter 4 instead studies feature learning, where feature distributions evolve during training.

3.4 Exercises

1. Phase Diagram and Double Descent. Fix $\gamma > 1$ and assume that $\sigma^2 > 0$ or the limiting value of $\|\theta_\star\|^2$ is positive. Using parts (a) and (c) of Theorem 3.1, show that the limiting total risk is strictly increasing for $0 < \delta < 1$, diverges as $\delta \to 1$ from either side, and is strictly decreasing for $\delta > 1$. Compute its limit as $\delta \to \infty$. When $\gamma < 1$, verify that the feature-bottleneck formulas meet the ambient-bottleneck formulas continuously at $\delta = \gamma$, and explain why there is no interpolation singularity at $\delta = 1$.
2. Scale Invariance of Ridgeless Features. For a fixed scalar $\alpha \neq 0$, replace S by αS and hence Z by αZ. Use the Moore–Penrose pseudoinverse to show that $\hat{\eta}$ is rescaled by α^{-1} while $\hat{\theta}$ and its risk are unchanged. Explain why the global scalar normalization of S is therefore immaterial for the ridgeless estimator.
3. Interpolation as a Rank Condition. Show that $\mathrm{rank}(XS) = \min\{m, n_0, n\}$ almost surely. Deduce that the random-feature model can interpolate every response vector $y \in \mathbb{R}^m$

exactly if and only if $n_0 \geq m$ and $n \geq m$. Use this rank condition to explain once more why increasing n through m creates an interpolation threshold when $\gamma > 1$, but not when $\gamma < 1$.

Chapter 4

Feature Learning and Mean-Field Limits

Chapters 2 and 3 studied the kernel regime from its optimization and statistical perspectives. In this chapter we turn to *feature learning*: parameterizations and scaling limits in which internal representations evolve by a $\Theta(1)$ amount as the width $n \to \infty$.

Throughout the chapter, we consider a supervised dataset $\{(x^\alpha, y^\alpha)\}_{\alpha=1}^m$ with inputs $x^\alpha \in \mathbb{R}^{n_0}$ and targets $y^\alpha \in \mathbb{R}$ (for simplicity we take a scalar output). Given a width-n network $f(\cdot\,;\theta)$ (suppressing the subscript n), we abbreviate $f^\alpha := f(x^\alpha;\theta)$ and define the squared loss

$$L(\theta) = \frac{1}{2m}\sum_{\alpha=1}^{m}\big(f^\alpha - y^\alpha\big)^2. \tag{4.1}$$

We write $r^\alpha(t) := f^\alpha(t) - y^\alpha$ when a residual variable is useful and otherwise keep the difference explicit. Throughout this chapter, m and n_0 are fixed as $n \to \infty$, and

$$\Phi_0^{\alpha\beta} := \frac{1}{n_0}\langle x^\alpha, x^\beta\rangle \tag{4.2}$$

denotes the input kernel introduced in Chapter 2.

The learning-rate multiplier $\eta = \eta_n$ is allowed to depend on width. For continuous-time arguments we use the convention

$$\partial_t\theta(t) = -\eta_n \nabla_\theta L(\theta(t)). \tag{4.3}$$

Thus $\eta_n = \eta_0 n$ denotes an accelerated gradient flow; equivalently, it is unit-rate gradient flow observed on a time scale accelerated by n. We do not absorb the factor $1/m$ in the loss into time.

4.1 One Hidden Layer: How Features Move

Throughout this section we use a one-hidden-layer network of width n:

$$f(x;\theta) = \frac{1}{\gamma_n\sqrt{n}}\sum_{i=1}^{n} a_i\,\varphi(h_i(x)), \qquad h_i(x) := \frac{1}{\sqrt{n_0}}\langle w_i, x\rangle, \qquad \theta = \{(a_i, w_i)\}_{i=1}^n, \tag{4.4}$$

with activation $\varphi : \mathbb{R} \to \mathbb{R}$. Here γ_n is an extra scaling knob beyond the usual NTK prefactor $1/\sqrt{n}$: setting $\gamma_n = 1$ gives the standard (NTK) scaling, while $\gamma_n = \sqrt{n}$ gives an overall prefactor $1/n$.

We first make the notion of feature learning quantitative. For a fixed data point x^α, define the (scalar) hidden preactivations

$$h_i^\alpha := h_i(x^\alpha) = \frac{1}{\sqrt{n_0}} \langle w_i, x^\alpha \rangle, \qquad i = 1, \dots, n, \tag{4.5}$$

and collect them into the width-n vector $h^\alpha \in \mathbb{R}^n$.

Definition 4.1 (Feature Learning Criterion)**.** Fix a training step k and a sample α, and write $h^\alpha(k) \in \mathbb{R}^n$ for the vector of hidden preactivations on input x^α. Let $\Delta h^\alpha(k) := h^\alpha(k+1) - h^\alpha(k)$. We say that the hidden layer exhibits feature learning at step k on sample α if

$$\frac{1}{n} \left\| \Delta h^\alpha(k) \right\|^2 = \Theta(1) \qquad (n \to \infty). \tag{4.6}$$

When (4.6) holds we will sometimes write simply $\Delta h^\alpha(k) = \Theta(1)$, meaning the normalized squared norm above is order one. Precisely, this says that the root-mean-square coordinate change is order one; under exchangeability and a nonconcentration condition, a typical coordinate is then also order one.

To see what the feature-learning criterion requires, consider one step of full-batch gradient descent with learning rate η:

$$w_i^+ = w_i - \eta \, \partial_{w_i} L, \qquad a_i^+ = a_i - \eta \, \partial_{a_i} L. \tag{4.7}$$

A direct derivative gives

$$\partial_{w_i} L = \frac{1}{m} \sum_{\beta=1}^{m} (f^\beta - y^\beta) \, \partial_{w_i} f^\beta = \frac{1}{m} \sum_{\beta=1}^{m} (f^\beta - y^\beta) \, \frac{1}{\gamma_n \sqrt{n n_0}} \, a_i \, \varphi'(h_i^\beta) \, x^\beta. \tag{4.8}$$

Therefore the induced one-step change in the feature $h_i^\alpha = n_0^{-1/2} \langle w_i, x^\alpha \rangle$ is

$$\begin{aligned} \Delta h_i^\alpha := h_i^{\alpha,+} - h_i^\alpha &= \frac{1}{\sqrt{n_0}} \langle w_i^+ - w_i, x^\alpha \rangle \\ &= -\frac{\eta}{m \, \gamma_n \sqrt{n}} \sum_{\beta=1}^{m} (f^\beta - y^\beta) \, a_i \, \varphi'(h_i^\beta) \, \Phi_0^{\beta\alpha}. \end{aligned} \tag{4.9}$$

Heuristically, if $(f^\beta - y^\beta) = \Theta(1)$ and the inner products and activation factors are $\Theta(1)$, then

$$\Delta h_i^\alpha = \Theta\Big(\frac{\eta}{\gamma_n \sqrt{n}} \Big). \tag{4.10}$$

Since $h^\alpha = (h_i^\alpha)_{i=1}^n$, this suggests

$$\frac{1}{n} \left\| \Delta h^\alpha \right\|^2 \approx \Big(\frac{\eta}{\gamma_n \sqrt{n}} \Big)^2, \tag{4.11}$$

so achieving the feature-learning condition (4.6) requires (at least) the scaling $\eta \asymp \gamma_n \sqrt{n}$.

The same update must also move the predictor on an order-one scale. In this chapter it is convenient to factor the output multiplier out of the tangent kernel and define the normalized empirical NTK by

$$K_t^{(n),\alpha\beta} := \frac{1}{n}\sum_{i=1}^{n}\varphi(h_i^\alpha)\varphi(h_i^\beta) + \Phi_0^{\alpha\beta}\frac{1}{n}\sum_{i=1}^{n}a_i^2\varphi'(h_i^\alpha)\varphi'(h_i^\beta). \tag{4.12}$$

Throughout this chapter, "NTK" refers to this normalized kernel unless the raw gradient Gram is named explicitly. Thus the raw gradient Gram used in Chapter 2 is $\langle\nabla_\theta f^\alpha, \nabla_\theta f^\beta\rangle = \gamma_n^{-2}K_t^{(n),\alpha\beta}$. When the empirical averages are order one, $K_t^{(n)} = \Theta(1)$. Under the gradient-flow convention (4.3), the predictor dynamics satisfy the exact identity

$$\partial_t f^\alpha(t) = -\frac{\eta_n}{m\gamma_n^2}\sum_{\beta=1}^{m}K_t^{(n),\alpha\beta}r^\beta(t), \tag{4.13}$$

so the predictor speed is of order η_n/γ_n^2. A first-order gradient-descent update has the same width scaling. Balancing order-one feature motion with order-one predictor motion therefore requires $\eta \asymp \gamma_n\sqrt{n}$ and $\eta \asymp \gamma_n^2$, respectively. Solving these two constraints gives $\gamma_n \asymp \sqrt{n}$ and $\eta \asymp n$.

These two balance conditions distinguish three regimes:

1. **NTK (lazy) scaling.** Set $\gamma_n = 1$ and take $\eta = \Theta(1)$. Then (4.10) gives $\Delta h_i^\alpha = \Theta(n^{-1/2})$, and thus
$$\frac{1}{n}\left\|\Delta h^\alpha\right\|^2 = \Theta(n^{-1}), \tag{4.14}$$
so the features freeze as $n \to \infty$. This is the standard lazy/NTK regime.
2. **Naively increasing the learning rate.** Keeping $\gamma_n = 1$ but taking $\eta \asymp \sqrt{n}$ makes $\Delta h_i^\alpha = \Theta(1)$. For squared loss this often leads to exploding outputs because the function-space speed in (4.13) scales like $\partial_t f = \Theta(\sqrt{n})$. However, for cross-entropy loss, it can be acceptable for the logits to diverge while training remains stable, because gradients can stay controlled even when outputs grow; see [17] for a careful analysis of this "controlled divergence" phenomenon.
3. **Mean-field scaling.** Instead choose $\gamma_n = \sqrt{n}$, i.e.
$$f(x;\theta) = \frac{1}{n}\sum_{i=1}^{n}a_i\,\varphi\left(\frac{1}{\sqrt{n_0}}\langle w_i, x\rangle\right), \tag{4.15}$$
and scale the learning rate as $\eta = \eta_0 n$ with $\eta_0 = \Theta(1)$. Then (4.10) gives $\Delta h_i^\alpha = \Theta(1)$, hence the feature-learning condition remains $\frac{1}{n}\|\Delta h^\alpha\|^2 = \Theta(1)$ at infinite width. At the same time, (4.13) gives $\partial_t f^\alpha(t) = \Theta(1)$ at the scaling level of this calculation, so the predictor evolves on an order-one time scale.

At a mean-zero random initialization, $f(x)$ is typically $\Theta(n^{-1/2})$ for fixed x. This smaller initialization scale is secondary to the balance above: because $y^\alpha = \Theta(1)$, the initial residual $f^\alpha - y^\alpha$ remains $\Theta(1)$.

4.2 Mean-Field ODE: An Interacting Particle System

With mean-field scaling, the parameters (a_i, w_i) behave like an interacting particle system. To emphasize this, it is convenient to encode all neurons via an empirical measure.

Regime	Output prefactor	η scaling	Feature update Δh_i^α
NTK (lazy)	$1/\sqrt{n}$ $(\gamma_n = 1)$	$\Theta(1)$	$\Theta(n^{-1/2})$
Naive large-η	$1/\sqrt{n}$ $(\gamma_n = 1)$	$\Theta(\sqrt{n})$	$\Theta(1)$
Mean-field	$1/n$ $(\gamma_n = \sqrt{n})$	$\Theta(n)$	$\Theta(1)$

Table 4.1: A quick scaling comparison for the one-hidden-layer network.

Definition 4.2 (Empirical Measure of Neurons)**.** For the one-hidden-layer network (4.4), define the empirical measure on parameter space $\Theta = \mathbb{R} \times \mathbb{R}^{n_0}$ by

$$\rho^{(n)} := \frac{1}{n} \sum_{i=1}^{n} \delta_{(a_i, w_i)}. \tag{4.16}$$

With mean-field scaling ($\gamma_n = \sqrt{n}$), the network can be written as

$$f(x; \theta) = \int a \, \varphi\left(\frac{1}{\sqrt{n_0}} \langle w, x \rangle \right) \rho^{(n)}(\mathrm{d}a \, \mathrm{d}w). \tag{4.17}$$

With $\eta_n = \eta_0 n$, the convention (4.3) gives the accelerated continuous-time gradient flow

$$\frac{\mathrm{d}}{\mathrm{d}t}(a_i, w_i) = -\eta_0 n \, \nabla_{(a_i, w_i)} L. \tag{4.18}$$

Using the dynamics (4.18) (and differentiating the network output (4.4)), one finds an ODE of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}(a_i, w_i) = b((a_i, w_i), \rho_t^{(n)}), \qquad \rho_t^{(n)} := \frac{1}{n} \sum_{j=1}^{n} \delta_{(a_j(t), w_j(t))}, \tag{4.19}$$

where the velocity field $b(\theta, \rho)$ depends on the population only through the prediction function $f_\rho(x) = \int a \, \varphi(n_0^{-1/2} \langle w, x \rangle) \, \rho(\mathrm{d}a \, \mathrm{d}w)$. Thus each "particle" is driven by the empirical distribution of all particles.

Passing from this interacting particle system to a deterministic infinite-width limit rests on *propagation of chaos*: as $n \to \infty$, any *fixed* finite collection of neurons becomes asymptotically independent (and identically distributed), with a common law ρ_t that evolves deterministically.

Definition 4.3 (Chaos)**.** Fix a time $t \geq 0$, and let $\theta_i^{(n)}(t) = (a_i(t), w_i(t)) \in \Theta$ denote the particle parameters under the mean-field gradient flow introduced above. Let $P_t^{(n)}$ be the joint law of $(\theta_1^{(n)}(t), \ldots, \theta_n^{(n)}(t)) \in \Theta^n$. We say that $P_t^{(n)}$ is ρ_t*–chaotic* if for every fixed $k \in \mathbb{N}$ and every bounded continuous test function $\psi : \Theta^k \to \mathbb{R}$,

$$\lim_{n \to \infty} \mathbb{E}[\psi(\theta_1^{(n)}(t), \ldots, \theta_k^{(n)}(t))] = \int_{\Theta^k} \psi(\theta_1, \ldots, \theta_k) \, \rho_t^{\otimes k}(\mathrm{d}\theta_1 \cdots \mathrm{d}\theta_k). \tag{4.20}$$

We say that the particle system *propagates chaos* on a time interval if there exists a deterministic family (ρ_t) such that $P_t^{(n)}$ is ρ_t–chaotic for every t in that interval.

Under standard well-posedness assumptions on the McKean–Vlasov interaction field (e.g. φ is C^1 with bounded Lipschitz derivative, the data are bounded, and ρ_0 has suitable moments), one can show that for every finite time horizon T there exists a unique deterministic solution

$(\rho_t)_{t\in[0,T]}$ of the mean-field PDE in Definition 4.4, and the particle system (4.19) propagates chaos on $[0,T]$: for each fixed k, the k-particle marginal law converges to $\rho_t^{\otimes k}$ uniformly for $t \in [0,T]$. Quantitative rates (often $O(n^{-1/2})$ in Wasserstein distance) are available under Lipschitz interactions.

Propagation of chaos is helpful because it upgrades the heuristic "replace the empirical measure by its limit" into a principled approximation scheme: at large width, individual neurons behave as i.i.d. samples from ρ_t, and macroscopic objects (the predictor f_{ρ_t}, risks, kernels, etc.) become deterministic functionals of ρ_t. This is exactly the point of the mean-field framework: it replaces an $O(n)$ interacting system of ODEs by a single deterministic measure-valued evolution that can be analyzed using PDE and optimal-transport techniques. See, e.g., [48, 25, 8] and, for a treatment tailored to mean-field neural networks beyond logarithmic time, [16].

4.3 Mean-Field PDE and Wasserstein Gradient Flow

The empirical measure $\rho_t^{(n)}$ from Definition 4.2 satisfies a transport equation. In the $n \to \infty$ limit this becomes a deterministic PDE for ρ_t.

Definition 4.4 (Mean-Field Transport PDE)**.** Let ρ_t be a time-dependent probability measure on $\Theta = \mathbb{R} \times \mathbb{R}^{n_0}$. Define the mean-field predictor

$$f_{\rho_t}(x) := \int a\,\varphi\left(\frac{1}{\sqrt{n_0}}\langle w, x\rangle\right)\,\rho_t(\mathrm{d}a\,\mathrm{d}w), \tag{4.21}$$

and the velocity field $b(\theta, \rho_t)$ for $\theta = (a, w)$ by

$$b(\theta,\rho_t) := -\frac{\eta_0}{m}\sum_{\alpha=1}^{m}(f_{\rho_t}(x^\alpha) - y^\alpha)\begin{bmatrix}\varphi\left(n_0^{-1/2}\langle w, x^\alpha\rangle\right)\\ n_0^{-1/2}a\,\varphi'\left(n_0^{-1/2}\langle w, x^\alpha\rangle\right)\,x^\alpha\end{bmatrix}. \tag{4.22}$$

The mean-field PDE is the continuity equation

$$\partial_t\rho_t = -\nabla_{(a,w)}\cdot\big(b(\cdot,\rho_t)\rho_t\big), \tag{4.23}$$

interpreted in the weak sense.

To see the weak formulation directly, let $q : \Theta \to \mathbb{R}$ be a smooth compactly supported test function. For the empirical measure $\rho_t^{(n)} = \frac{1}{n}\sum_i \delta_{\theta_i(t)}$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int q(\theta)\,\rho_t^{(n)}(\mathrm{d}\theta) = \frac{1}{n}\sum_{i=1}^{n}\langle\nabla q(\theta_i(t)), \partial_t\theta_i(t)\rangle. \tag{4.24}$$

Using the McKean–Vlasov form (4.19), this becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\int q\,\mathrm{d}\rho_t^{(n)} = \int\langle\nabla q(\theta), b(\theta,\rho_t^{(n)})\rangle\,\rho_t^{(n)}(\mathrm{d}\theta). \tag{4.25}$$

Formally passing to $n \to \infty$ (where $\rho_t^{(n)} \xrightarrow{\mathrm{w}} \rho_t$ and $b(\theta,\rho_t^{(n)}) \to b(\theta,\rho_t)$) yields the weak formulation

$$\frac{\mathrm{d}}{\mathrm{d}t}\int q\,\mathrm{d}\rho_t = \int\langle\nabla q(\theta), b(\theta,\rho_t)\rangle\,\rho_t(\mathrm{d}\theta). \tag{4.26}$$

An integration by parts in (4.26) recovers the divergence form (4.23).

The same transport equation also has a Wasserstein gradient-flow interpretation. Define the functional on measures

$$F(\rho) := \frac{1}{2m} \sum_{\alpha=1}^{m} \big(f_\rho(x^\alpha) - y^\alpha\big)^2. \tag{4.27}$$

The mean-field dynamics can be viewed as the W_2–gradient flow of F in the space of probability measures, in the sense of Otto's calculus: the continuity equation (4.23) has the form $\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$ with velocity field $v_t(\theta) = -\eta_0 \nabla_\theta (\delta F/\delta \rho)(\theta)$. This is the same structure that underlies the Jordan–Kinderlehrer–Otto scheme [27] and the general theory of metric-space gradient flows [2].

The standard mean-field picture for two-layer networks emerged from several essentially concurrent papers between 2017 and 2020. Under an overall $1/n$ output scaling, these works viewed the network as an interacting particle system in parameter space whose empirical measure converges as $n \to \infty$ to a deterministic transport PDE with a Wasserstein gradient-flow interpretation. In particular, the perspectives in Nitanda and Suzuki [40], Rotskoff and Vanden-Eijnden [45], Mei–Montanari–Nguyen [37], Chizat and Bach [10], and Sirignano and Spiliopoulos [46] (first posted on arXiv in 2018) are strikingly aligned despite emphasizing different tools and goals. This rapid convergence is one reason the "neurons as particles" framework became a natural organizing principle for shallow feature learning.

Remark 4.5 (Permutation Invariance in Deep Networks). A one-hidden-layer network is invariant under permutations of the neurons: relabeling (a_i, w_i) does not change f. The empirical-measure description in Definition 4.2 is essentially the quotient by this symmetry.

For deep networks, there are independent permutations in each hidden layer, and the natural objects are distributions over layerwise paths through the network, or equivalently increasingly structured exchangeable arrays. This rapidly pushes us beyond the classical McKean–Vlasov setting and explains why the shallow mean-field PDE does not straightforwardly generalize to deep feature learning.

Tracking deep training dynamics at infinite width therefore requires additional machinery. One such tool is dynamical mean field theory (DMFT), which we sketch in Section 4.5 after introducing the maximal-update scaling.

4.4 Maximal Update and the μP Parameterization

Mean-field scaling in deep networks admits multiple inequivalent ways to distribute width-dependent factors across layers and to scale learning rates. The μP framework (also called maximal update parameterization) provides a systematic way to choose these scalings so that representation learning persists as $n \to \infty$ [50, 51]. (A personal note: I am not particularly attached to the naming "μP"—the term maximal update captures the core idea perfectly well.)

Consider a depth-d fully-connected network with d hidden layers of width n, forward preacti-

vations $h_\ell(x) \in \mathbb{R}^n$ and postactivations $\varphi_\ell(x) := \varphi(h_\ell(x))$:

$$f(x;\theta) = \frac{1}{\gamma_n \sqrt{n}} W_d \, \varphi_d(x), \tag{4.28}$$

$$h_{\ell+1}(x) = \frac{1}{\sqrt{n}} W_\ell \, \varphi_\ell(x), \qquad \ell = 1, \dots, d-1, \tag{4.29}$$

$$h_1(x) = \frac{1}{\sqrt{n_0}} W_0 \, x. \tag{4.30}$$

We take $W_0 \in \mathbb{R}^{n \times n_0}$, $W_\ell \in \mathbb{R}^{n \times n}$ for $1 \le \ell \le d-1$, and $W_d \in \mathbb{R}^{1 \times n}$. These are raw weight coordinates: the width-dependent multipliers are displayed explicitly in (4.30), and the same learning-rate multiplier η is applied to every raw weight matrix. A standard choice is i.i.d. Gaussian initialization $[W_\ell(0)]_{ij} \sim \mathcal{N}(0,1)$ (independently), together with a global learning-rate scaling that we derive heuristically below.

The activation function does not change the width dependence of the scaling analysis below, as long as φ and φ' remain $\Theta(1)$ on the typical scale of preactivations; it mainly contributes bounded multiplicative factors and modifies the limiting kernels.

To formulate maximal update at the representation level, we isolate the contribution to the next-layer preactivations coming only from the previous layer's weight update.

Definition 4.6 (Maximal Feature Update Criterion)**.** Fix $\ell \in \{1, \dots, d\}$ and consider one gradient-based update step producing a weight increment $\Delta W_{\ell-1}$. Define the induced change in the layer-ℓ preactivations coming only from $W_{\ell-1}$ by

$$\Delta_{W_{\ell-1}} h_\ell(x) := \begin{cases} \dfrac{1}{\sqrt{n_0}} \Delta W_0 x, & \ell = 1, \\ \dfrac{1}{\sqrt{n}} \Delta W_{\ell-1} \varphi_{\ell-1}(x), & 2 \le \ell \le d. \end{cases} \tag{4.31}$$

We say the parameterization achieves a maximal update at layer ℓ if, for typical training inputs and steps,

$$\frac{1}{n} \left\| \Delta_{W_{\ell-1}} h_\ell(x^\alpha) \right\|^2 = \Theta(1) \qquad (n \to \infty). \tag{4.32}$$

If (4.32) holds simultaneously for all layers $\ell = 1, \dots, d$ (with a single consistent scaling of γ_n and learning-rate hyperparameters), we say that the network satisfies the maximal feature update criterion. This is the operational layerwise condition used below; output stability will select the corresponding maximal-update/μP scaling.

Besides quantifying representation motion, the maximal-update criterion also explains when the NTK itself can evolve. In the lazy/NTK regime, the empirical forward feature kernels $\Phi_\ell^{(n),\alpha\beta}(t)$ are essentially frozen, so the NTK stays close to its initialization. In contrast, let $G_\ell^{(n),\alpha\beta}(t)$ denote the empirical backward kernels formed from the scaled sensitivities introduced below, with $G_{d+1}^{(n),\alpha\beta}(t) := 1$. Consistently with the shallow convention above, define the normalized empirical NTK by the exact finite-width decomposition

$$K_t^{(n),\alpha\beta} := \sum_{\ell=0}^{d} G_{\ell+1}^{(n),\alpha\beta}(t) \, \Phi_\ell^{(n),\alpha\beta}(t). \tag{4.33}$$

Because the sensitivities below contain an additional factor γ_n relative to those in Chapter 2, the raw gradient Gram is $\gamma_n^{-2} K_t^{(n)}$. Consequently, under (4.3), the exact predictor evolution is

$$\partial_t f^\alpha(t) = -\frac{\eta_n}{m\gamma_n^2} \sum_{\beta=1}^{m} K_t^{(n),\alpha\beta} r^\beta(t). \tag{4.34}$$

Thus $K_t^{(n)}$ can have a nonvanishing limit even when the raw gradient Gram vanishes, while the ratio η_n/γ_n^2 controls the output time scale. The maximal-update condition (4.32) keeps the feature changes induced by each layer's weight update at $\Theta(1)$ scale, allowing the corresponding Φ_ℓ factors to evolve at that scale. When these Φ_ℓ evolve, each layer's contribution to the normalized NTK can evolve as well, enabling a genuinely non-kernel limit. The reduction to a kernel method relies on a kernel function fixed independently of the training data and unchanged throughout optimization. Once feature learning produces a nontrivial evolution of the limiting tangent kernel, this fixed-kernel reduction no longer applies.

We now give a scaling calculation showing how γ_n and the learning-rate scaling should be chosen so that (i) the network function evolves at $\Theta(1)$ speed and (ii) each hidden layer satisfies maximal update in the sense of Definition 4.6. We perform the calculation for GD; for SGD the width dependence is the same (batch-size factors affect constants but not powers of n). We organize the calculation into six steps.

Step 1: Setup. For notational clarity, consider a single sample (x, y) and the per-sample squared loss

$$L(\theta) = \frac{1}{2}\big(f(x;\theta) - y\big)^2. \tag{4.35}$$

Let the residual be $r := f(x;\theta) - y$, which we treat as $\Theta(1)$ (width-independent) since $y = \Theta(1)$. Write the forward postactivations as $\varphi_\ell = \varphi(h_\ell)$.

For the backward pass, it is convenient in the μP setting to absorb the output scale factor γ_n into the backpropagated signal. Define the *μP-scaled backward sensitivity* by

$$g_\ell := \gamma_n \sqrt{n}\, \frac{\partial f}{\partial h_\ell} \in \mathbb{R}^n. \tag{4.36}$$

(Compared to the Chapter 2 normalization $\sqrt{n}\,\partial f/\partial h_\ell$, this inserts an extra factor of γ_n; the benefit is that g_ℓ is width-stable under μP scaling.)

Under the scaling (4.30), the exact backpropagation recursion is

$$g_\ell = \frac{1}{\sqrt{n}}\, \mathrm{diag}(\varphi'(h_\ell)) W_\ell^\top g_{\ell+1}, \qquad \ell = d-1, \dots, 1, \tag{4.37}$$

with top boundary

$$g_d = \gamma_n \sqrt{n}\, \frac{\partial}{\partial h_d}\Big(\frac{1}{\gamma_n \sqrt{n}}\, W_d\, \varphi(h_d)\Big) = W_d^\top \odot \varphi'(h_d), \tag{4.38}$$

so the typical-coordinate scale of g_ℓ is the same at every layer under stable wide-limit forward and backward signal propagation. We record this heuristic in normalized root-mean-square form:

$$\frac{1}{n}\, \|g_\ell\|^2 = \Theta(1), \qquad \ell = 1, \dots, d. \tag{4.39}$$

Step 2: Gradient Magnitudes. Since $L = \frac{1}{2}r^2$, we have $\nabla_\theta L = r\,\nabla_\theta f$. For $1 \le \ell \le d-1$, using $h_{\ell+1} = \sqrt{\frac{1}{n}}\,W_\ell \varphi_\ell$,

$$\nabla_{W_\ell} f = \frac{1}{\sqrt{n}}\Big(\frac{\partial f}{\partial h_{\ell+1}}\Big)\varphi_\ell^\top = \frac{1}{\gamma_n n}\, g_{\ell+1}\, \varphi_\ell^\top, \tag{4.40}$$

so a typical entry of $\nabla_{W_\ell} L$ scales as

$$[\nabla_{W_\ell} L]_{ij} = \Theta\Big(\frac{r}{\gamma_n n}\, g_{\ell+1,i}\, \varphi_{\ell,j}\Big) = \Theta\Big(\frac{1}{\gamma_n n}\Big), \tag{4.41}$$

since $r = \Theta(1)$, $\varphi_{\ell,j} = \Theta(1)$, and a typical coordinate $g_{\ell+1,i} = \Theta(1)$ by (4.39). Thus an SGD/GD update with learning rate η satisfies

$$[\Delta W_\ell]_{ij} = \Theta\Big(\frac{\eta}{\gamma_n n}\Big). \tag{4.42}$$

Step 3: Maximal Update Condition. Now consider the layerwise preactivation change from (4.31):

$$\Delta_{W_\ell} h_{\ell+1} = \frac{1}{\sqrt{n}}\,\Delta W_\ell\, \varphi_\ell. \tag{4.43}$$

Using that (to leading order in a single-sample calculation) ΔW_ℓ is proportional to the outer product $g_{\ell+1}\varphi_\ell^\top$, we obtain componentwise

$$[\Delta_{W_\ell} h_{\ell+1}]_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [\Delta W_\ell]_{ij}\, \varphi_{\ell,j} \approx -\frac{\eta r}{\sqrt{n}}\, \frac{1}{\gamma_n n}\, g_{\ell+1,i} \sum_{j=1}^{n} \varphi_{\ell,j}^2. \tag{4.44}$$

Since $\frac{1}{n}\sum_j \varphi_{\ell,j}^2 = \Theta(1)$, the sum $\sum_j \varphi_{\ell,j}^2 = \Theta(n)$ and thus

$$[\Delta_{W_\ell} h_{\ell+1}]_i = \Theta\Big(\frac{\eta}{\sqrt{n}}\, \frac{1}{\gamma_n}\, g_{\ell+1,i}\Big) = \Theta\Big(\frac{\eta}{\gamma_n \sqrt{n}}\Big). \tag{4.45}$$

Therefore

$$\frac{1}{n}\,\|\Delta_{W_\ell} h_{\ell+1}\|^2 = \Theta\Big(\frac{\eta^2}{\gamma_n^2 n}\Big), \tag{4.46}$$

and enforcing the maximal-update criterion (4.32) requires

$$\eta \asymp \gamma_n \sqrt{n}. \tag{4.47}$$

Step 4: Input Boundary. The first weight matrix has the exact derivative

$$\nabla_{W_0} f = \frac{1}{\gamma_n \sqrt{n n_0}}\, g_1 x^\top. \tag{4.48}$$

Consequently, for the single-sample update used here,

$$\Delta_{W_0} h_1 = -\frac{\eta r}{\gamma_n \sqrt{n}}\, g_1 \frac{\|x\|^2}{n_0} = \Theta\left(\frac{\eta}{\gamma_n \sqrt{n}}\right) \tag{4.49}$$

when $n_0^{-1}\|x\|^2 = \Theta(1)$. Thus the same maximal-update constraint applies at the input boundary as at every internal hidden layer.

Step 5: Stable Function-Space Dynamics. We also require that the output itself changes on a $\Theta(1)$ scale. Consider the contribution coming from the output weights W_d. Since $f(x) = \frac{1}{\gamma_n\sqrt{n}} W_d \varphi_d$, we have

$$\nabla_{W_d} f = \frac{1}{\gamma_n\sqrt{n}}\, \varphi_d^\top, \quad \Rightarrow \quad [\nabla_{W_d} L]_j = \Theta\Big(\frac{r}{\gamma_n\sqrt{n}}\Big) = \Theta\Big(\frac{1}{\gamma_n\sqrt{n}}\Big). \tag{4.50}$$

Thus $[\Delta W_d]_j = \Theta(\eta/(\gamma_n\sqrt{n}))$ and the resulting change in the output is

$$\Delta f = \frac{1}{\gamma_n\sqrt{n}}\, \Delta W_d\, \varphi_d \approx -\frac{\eta r}{\gamma_n^2 n} \sum_{j=1}^{n} \varphi_{d,j}^2 = \Theta\Big(\frac{\eta}{\gamma_n^2}\Big), \tag{4.51}$$

using again that $\frac{1}{n}\sum_j \varphi_{d,j}^2 = \Theta(1)$. To keep $\Delta f = \Theta(1)$ we therefore need

$$\eta \asymp \gamma_n^2. \tag{4.52}$$

Step 6: Solving the Scaling Constraints. Combining (4.52) and (4.47) yields

$$\gamma_n \asymp \sqrt{n}, \qquad \eta \asymp n. \tag{4.53}$$

Equivalently, the overall output prefactor becomes $1/(\gamma_n\sqrt{n}) = 1/n$, and the learning rate is scaled up by n. This is the basic μP/maximal-update scaling for MLPs.

Remark 4.7 (Shallow Mean Field as the $d = 1$ Case). For $d = 1$, identify $a_i = (W_d)_{1i}$ and let $w_i^\top$ be the ith row of W_0. Then (4.30) is exactly the shallow model (4.4):

$$f(x) = \frac{1}{\gamma_n\sqrt{n}} \sum_{i=1}^{n} a_i \varphi\left(\frac{1}{\sqrt{n_0}}\langle w_i, x\rangle\right). \tag{4.54}$$

Hence $\gamma_n = \sqrt{n}$ and the common raw learning-rate multiplier $\eta = \eta_0 n$ recover the shallow mean-field parameterization and particle dynamics exactly [50].

Remark 4.8 (Generality). The maximal-update/μP philosophy extends far beyond fully-connected MLPs: one can, in principle, derive analogous width-dependent scalings for other architectures (CNNs, attention, residual connections, etc.) and for other training algorithms (Adam, momentum, weight decay, etc.). But each new setting requires its own calculation, because the relevant forward/backward signal propagation and parameter grouping changes.

4.5 A Sketch of DMFT for Deep Feature Learning

The deep μP scaling does not collapse to a deterministic PDE like the shallow mean-field limit. Instead, a representative forward/backward coordinate remains stochastic at infinite width, while its statistics close through deterministic time-dependent kernel order parameters. *Dynamical mean field theory* (DMFT) provides this reduced description. We follow the infinite-width self-consistent DMFT of Bordelon and Pehlevan [6, Secs. 2–3 and App. D.7]; their later work develops finite-width fluctuations around this limit [7].

We begin by introducing the state variables and kernels that close the limiting dynamics. At width n, let $h_\ell^\alpha(t) \in \mathbb{R}^n$ denote the forward preactivations at layer ℓ on input x^α. For the backward pass, use the μP-scaled sensitivity

$$g_\ell^\alpha(t) := \gamma_n\sqrt{n}\, \frac{\partial f^\alpha(t)}{\partial h_\ell^\alpha(t)} \in \mathbb{R}^n, \tag{4.55}$$

whose typical coordinates are order one under the stable scaling assumptions of Section 4.4. We suppress width dependence on the state variables h_ℓ^α and g_ℓ^α, reserving the superscript (n) for empirical kernels.

The finite-width two-time feature and gradient kernels are

$$\Phi_\ell^{(n),\alpha\beta}(t,s) := \frac{1}{n}\left\langle \varphi(h_\ell^\alpha(t)), \varphi\Big(h_\ell^\beta(s)\Big)\right\rangle, \qquad G_\ell^{(n),\alpha\beta}(t,s) := \frac{1}{n}\left\langle g_\ell^\alpha(t), g_\ell^\beta(s)\right\rangle. \tag{4.56}$$

Their diagonal $t = s$ agrees with the empirical kernels used above. DMFT also introduces response (or susceptibility) kernels A_ℓ and B_ℓ, defined below.

With this notation in place, we can state a representative form of the DMFT equations. Fix $\gamma_n = \gamma_0\sqrt{n}$ with $\gamma_0 = \Theta(1)$ and the stable learning-rate scale $\eta_n \asymp \gamma_n^2$ (Section 4.4). If u denotes physical time under (4.3), define macroscopic time by $t := \eta_n u/(m\gamma_n^2)$. At fixed depth, fixed m, and over a finite time horizon, the DMFT limit treats the coordinates as asymptotically i.i.d. across neurons. From this point onward, $h_\ell^\alpha(t)$, $z_\ell^\alpha(t)$, and $g_\ell^\alpha(t)$ denote one representative scalar coordinate of the limiting single-site process [6, Eqs. (9)–(10)]. Here *single site* is statistical-physics language for the effective process associated with one typical neuron index after the influence of all other neurons has been summarized by the mean fields. In probabilistic language, this is a representative mean-field particle—more precisely here, its coupled forward/backward coordinate trajectory across data examples and training times.

For $1 \le \ell \le d$, the representative process satisfies

$$h_\ell^\alpha(t) = u_\ell^\alpha(t) + \gamma_0 \int_0^t \sum_{\beta=1}^m \Big[A_{\ell-1}^{\alpha\beta}(t,s) - \Phi_{\ell-1}^{\alpha\beta}(t,s)\, r^\beta(s)\Big]\, g_\ell^\beta(s)\,\mathrm{d}s, \tag{4.57}$$

$$z_\ell^\alpha(t) = v_\ell^\alpha(t) + \gamma_0 \int_0^t \sum_{\beta=1}^m \Big[B_\ell^{\alpha\beta}(t,s) - G_{\ell+1}^{\alpha\beta}(t,s)\, r^\beta(s)\Big]\, \varphi\big(h_\ell^\beta(s)\big)\,\mathrm{d}s, \tag{4.58}$$

$$g_\ell^\alpha(t) = \varphi'\big(h_\ell^\alpha(t)\big)\, z_\ell^\alpha(t), \tag{4.59}$$

Here u_ℓ and v_ℓ are independent centered Gaussian drives with

$$\mathbb{E}[u_\ell^\alpha(t)\, u_\ell^\beta(s)] = \Phi_{\ell-1}^{\alpha\beta}(t,s), \qquad \mathbb{E}[v_\ell^\alpha(t)\, v_\ell^\beta(s)] = G_{\ell+1}^{\alpha\beta}(t,s), \qquad u_\ell \perp v_\ell, \tag{4.60}$$

Self-consistency means that the empirical kernels in (4.56) converge, at each fixed pair (t,s), to the deterministic expectations

$$\begin{aligned} \Phi_\ell^{\alpha\beta}(t,s) &:= \mathbb{E}\Big[\varphi\big(h_\ell^\alpha(t)\big)\varphi\big(h_\ell^\beta(s)\big)\Big], & G_\ell^{\alpha\beta}(t,s) &:= \mathbb{E}\Big[g_\ell^\alpha(t) g_\ell^\beta(s)\Big], \\ \Phi_0^{\alpha\beta}(t,s) &:= \Phi_0^{\alpha\beta} = \frac{1}{n_0}\langle x^\alpha, x^\beta\rangle, & G_{d+1}^{\alpha\beta}(t,s) &:= 1. \end{aligned} \tag{4.61}$$

For $1 \le \ell \le d-1$, the response kernels are the causal functional derivatives

$$\begin{aligned} A_\ell^{\alpha\beta}(t,s) &:= \frac{1}{\gamma_0}\mathbb{E}\left[\frac{\delta\varphi\big(h_\ell^\alpha(t)\big)}{\delta v_\ell^\beta(s)}\right], & B_\ell^{\alpha\beta}(t,s) &:= \frac{1}{\gamma_0}\mathbb{E}\left[\frac{\delta g_{\ell+1}^\alpha(t)}{\delta u_{\ell+1}^\beta(s)}\right], \\ A_0 &:= 0, & B_d &:= 0. \end{aligned} \tag{4.62}$$

They vanish for $s > t$; the expectation is over the same single-site process [6, Sec. 3.2]. Taken together, (4.57)–(4.62) form a fixed-point equation for the law of the representative process and

its kernel order parameters: a candidate collection (Φ, G, A, B) determines the Gaussian drives and memory terms, hence the process, and the correlations and response derivatives of that process must reproduce the same (Φ, G, A, B).

Finally, the limiting kernels couple back to the training residual. On the equal-time diagonal, the limiting normalized NTK is

$$K_t^{\alpha\beta} := \sum_{\ell=0}^{d} G_{\ell+1}^{\alpha\beta}(t,t)\, \Phi_\ell^{\alpha\beta}(t,t). \tag{4.63}$$

Together with $r^\alpha(t) = f^\alpha(t) - y^\alpha$, it closes the macroscopic-time dynamics through

$$r^\alpha(t) = r^\alpha(0) - \int_0^t \sum_{\beta=1}^{m} K_s^{\alpha\beta} r^\beta(s)\, \mathrm{d}s, \tag{4.64}$$

equivalently $\partial_t r(t) = -K_t r(t)$ for $r(t) \in \mathbb{R}^m$ [6, Eq. (3)]. Under the centered initialization used here, $f^\alpha(0) \xrightarrow{\mathbb{P}} 0$, so $r^\alpha(0) \xrightarrow{\mathbb{P}} -y^\alpha$. Writing the same quantities as functions of physical time u,

$$\frac{\mathrm{d}}{\mathrm{d}u} f^\alpha(u) \;=\; \frac{\mathrm{d}}{\mathrm{d}u} r^\alpha(u) = -\frac{\eta_n}{m\gamma_n^2} \sum_{\beta=1}^{m} K_u^{\alpha\beta} r^\beta(u). \tag{4.65}$$

The practical takeaway is the contrast with the shallow mean-field limit: deep feature learning at infinite width is not captured by a single deterministic transport PDE for a parameter measure. The representative coordinate retains non-Markovian memory and Gaussian randomness, while the self-consistent kernels and predictor trajectory are deterministic.

4.6 Summary and Outlook

The general layerwise feature-learning diagnostic is $n^{-1}\|\Delta h_\ell^\alpha\|^2$, which vanishes in the lazy regime. Maximal update sharpens this diagnostic by requiring the component induced by each layer's own weights to satisfy $n^{-1}\|\Delta_{W_{\ell-1}} h_\ell^\alpha\|^2 = \Theta(1)$. The key distinction is between a scaling and a description of its infinite-width limit. For $d = 1$, maximal update is exactly shallow mean-field scaling, whose limit is a deterministic neuron law; for deep networks, DMFT describes the corresponding feature-learning limit. Thus DMFT is not an additional parameterization.

At output level, the normalized NTK is the common bridge: it freezes to K in the lazy limit and evolves with the state in feature-learning limits. The maximal-update (μP) limit in this chapter takes width to infinity at fixed depth, whereas Chapters 5 and 6 study the proportional depth–width limit. These regimes are not compatible: finite-width effects that typically vanish at fixed depth can accumulate to a nontrivial effect as depth tends to infinity. As we will see, the proportional limit is precisely the regime that captures this phenomenon.

Framework and role	Width/time balance	Infinite-width description
Lazy NTK (scaling regime)	$\gamma_n = 1$, $\eta_n = \Theta(1)$; $n^{-1}\|\Delta h_\ell^\alpha\|^2 = o(1)$.	The normalized NTK converges to a time-independent deterministic kernel K; the residual follows linear kernel gradient flow.
Shallow mean field, $d = 1$ (scaling and limit)	$\gamma_n = \sqrt{n}$, $\eta_n = \eta_0 n$; $n^{-1}\|\Delta h_1^\alpha\|^2 = \Theta(1)$.	The empirical neuron measure converges to a deterministic law ρ_t governed by a transport PDE/Wasserstein gradient flow; its predictor and K_t can evolve.
Deep maximal update/μP (scaling principle)	$\gamma_n = \gamma_0\sqrt{n}$, $\eta_n \asymp \gamma_n^2 \asymp n$; every hidden layer satisfies $n^{-1}\|\Delta_{W_{\ell-1}} h_\ell^\alpha\|^2 = \Theta(1)$.	The scaling selects a stable feature-learning limit but does not by itself specify the limiting dynamics; its $d = 1$ case is the preceding row.
DMFT under deep maximal update (limit description)	With physical time u, use $t = \eta_n u/(m\gamma_n^2)$.	A stochastic, non-Markovian representative coordinate closes through deterministic $\Phi_\ell, G_\ell, A_\ell, B_\ell$; the limiting K_t and $r(t)$ are deterministic, with $\partial_t r = -K_t r$ in macroscopic time.

Table 4.2: Relationship among the scalings and infinite-width descriptions in this chapter.

4.7 Exercises

1. Shallow Scaling Balance. Assume that the residuals, activation factors, input kernels, and normalized empirical NTK remain order one, and write $\gamma_n \asymp n^p$ and $\eta_n \asymp n^q$. Use (4.10) and (4.13) to show that a typical feature coordinate and the predictor speed scale as

$$\Delta h_i^\alpha = \Theta\big(n^{q-p-1/2}\big), \qquad \partial_t f^\alpha = \Theta\big(n^{q-2p}\big). \tag{4.66}$$

   Require both quantities to remain order one and solve for p and q. Conclude that $\gamma_n \asymp \sqrt{n}$, $\eta_n \asymp n$, and the output prefactor is of order $1/n$. Finally, recover from the same formulas the feature and output scalings in the NTK regime $(p, q) = (0, 0)$ and the naive large-learning-rate regime $(p, q) = (0, 1/2)$.
2. Shallow Mean Field as the $d = 1$ Maximal-Update Model. Set $d = 1$, identify $a_i = (W_1)_{1i}$, and let $w_i^\top$ be the ith row of W_0. First verify that (4.30) becomes exactly (4.4). Next show that $g_{1,i}^\alpha = a_i\varphi'(h_i^\alpha)$. Using $G_2^{(n),\alpha\beta} = 1$, show that the deep normalized-NTK decomposition (4.33) reduces exactly to the shallow formula (4.12). Finally, with $\gamma_n = \sqrt{n}$ and $\eta_n = \eta_0 n$, verify coordinatewise that the gradient updates of W_0 and W_1 agree with the shallow particle updates of w_i and a_i. Conclude that the shallow mean-field model is the $d = 1$ maximal-update model without any additional change of coordinates or layerwise learning rates.
3. DMFT Residual, Sign, and Time Conventions. Starting from the physical-time equation (4.65) and the change of time $t = \eta_n u/(m\gamma_n^2)$, use $r^\alpha = f^\alpha - y^\alpha$ to derive

$$\frac{\mathrm{d}}{\mathrm{d}t} r^\alpha(t) = -\sum_{\beta=1}^{m} K_t^{\alpha\beta} r^\beta(t) \tag{4.67}$$

   and then recover the integral equation (4.64). Check that centered initialization gives $r^\alpha(0) \xrightarrow{\mathbb{P}} -y^\alpha$.

If the data-dependent terms in the representative DMFT equations are temporarily written as $A_{\ell-1}+\Phi_{\ell-1}(y-f)$ and $B_\ell+G_{\ell+1}(y-f)$, use $y^\beta-f^\beta=-r^\beta$ to verify the signs $A_{\ell-1}-\Phi_{\ell-1}r$ in (4.57) and $B_\ell - G_{\ell+1}r$ in (4.58). Finally, substitute $A_0=0$ and $\Phi_0^{\alpha\beta}=n_0^{-1}\langle x^\alpha, x^\beta\rangle$ at the input boundary, and $B_d=0$ and $G_{d+1}^{\alpha\beta}=1$ at the output boundary, and write the resulting $\ell=1$ and $\ell=d$ equations explicitly.

Chapter 5

Depth Scaling and Covariance SDE Limits

These notes study what happens when network *depth* becomes large. The recurring template is: (i) identify a scaling (step size) per layer so that each layer contributes a small perturbation, and (ii) pass to a continuous-time limit.

Throughout this chapter, unless stated otherwise, weight entries are i.i.d. standard Gaussian and activations are applied coordinatewise. For each model with independent layer weights, let $\mathcal{F}_\ell$ denote the sigma-field generated by all hidden states through layer ℓ (including every propagated input when there are several). Conditioning on $\mathcal{F}_\ell$ makes the next-layer preactivations exactly Gaussian at every finite width, as in (2.26). Large-width approximations enter when the resulting random empirical covariances and their one-step fluctuations are passed to a limiting recursion or diffusion. Although this chapter focuses on networks at initialization with i.i.d. Gaussian weight entries, Section 6.3 in Chapter 6 develops a one-step training-time calculation in the same proportional depth–width framework. We use φ for the activation and write data points as $\{x^\alpha\}_{\alpha=1}^m$.

5.1 ResNet Scaling Limits

We begin with the Euler approximation underlying the deterministic-depth regime. An ordinary differential equation (ODE) is determined by a vector field b and an initial condition x_0.

Definition 5.1 (ODE Flow)**.** Let $b : \mathbb{R}^n \to \mathbb{R}^n$ and $x_0 \in \mathbb{R}^n$. An ODE solution is a differentiable path $(X_t)_{t\in[0,T]}$ satisfying

$$\partial_t X_t = b(X_t), \qquad X_0 = x_0. \tag{5.1}$$

The simplest numerical scheme is Euler's method with step size $\eta > 0$:

$$Y_{k+1} = Y_k + \eta\, b(Y_k), \qquad Y_0 = x_0. \tag{5.2}$$

It is helpful to compare the discrete sequence to a piecewise-constant interpolation:

$$Y_t^{(\eta)} := Y_{\lfloor t/\eta \rfloor}, \qquad t \in [0, T]. \tag{5.3}$$

Theorem 5.2 (Euler Convergence)**.** *Assume b is globally Lipschitz on $\mathbb{R}^n$:*

$$\|b(x) - b(y)\| \le C_1 \|x - y\| \quad \forall x, y \in \mathbb{R}^n. \tag{5.4}$$

Then the ODE (5.1) *has a unique solution on $[0, T]$, and the Euler interpolation* (5.3) *converges uniformly to it as $\eta \to 0$:*

$$\sup_{t\in[0,T]} \left\| X_t - Y_t^{(\eta)} \right\| \to 0. \tag{5.5}$$

The conclusion in (5.5) compares the whole depth trajectory with the ODE path, rather than only the network state at one chosen layer. With layer k identified with time $k\eta$, the approximation error is small simultaneously for every layer k with $k\eta \le T$, so a single ODE path describes the full deep network.

The Euler viewpoint becomes structural in deep residual networks (ResNets) [22].

For comparison, consider an MLP layer of width n:

$$h_{\ell+1} = \frac{1}{\sqrt{n}} W_\ell \, \varphi(h_\ell), \qquad W_\ell \in \mathbb{R}^{n\times n}. \tag{5.6}$$

In contrast, a (basic) ResNet layer adds a skip connection:

$$h_{\ell+1} = h_\ell \; + \; \frac{1}{\sqrt{n}} W_\ell \, \varphi(h_\ell). \tag{5.7}$$

The skip term makes it natural to rescale the residual block so that each layer is a *small* update.

A classical deterministic-depth scaling is

$$h_{\ell+1} = h_\ell + \varepsilon \, \frac{1}{\sqrt{n}} W \, \varphi(h_\ell), \tag{5.8}$$

with a *shared* weight matrix W across layers (a recurrent-in-depth model). Fix a "layer time" horizon $\bar\tau > 0$ and, for a depth-d network, set $\varepsilon := \bar\tau/d$ and $\tau := \varepsilon\ell$. Then the update (5.8) is exactly Euler discretization of

$$\partial_\tau h_\tau = \frac{1}{\sqrt{n}} W \, \varphi(h_\tau), \qquad \tau \in [0, \bar\tau]. \tag{5.9}$$

For fixed n and W, a Lipschitz activation (including ReLU) makes the vector field in (5.9) globally Lipschitz, so Theorem 5.2 applies as $d \to \infty$. This connects deep ResNets to *Neural ODEs* [9].

The same viewpoint accommodates non-uniform layer times: varying step sizes ε_ℓ and/or weights W_ℓ can approximate a time-inhomogeneous ODE, with the Euler scheme now tracking a time-dependent vector field.

5.1.1 CLT Depth Scaling: From Random Walk to SDE

If each layer contributes random fluctuations of size $\sqrt{\eta}$ (rather than deterministic size η), then a diffusion limit emerges.

The basic template is Donsker's random-walk limit in (1.3)–(1.5): centered increments of size $\sqrt{\eta}$ accumulate to Brownian motion. Allowing state-dependent variance and adding an $O(\eta)$ drift leads to an Itô SDE

$$\mathrm{d}X_\tau = b(X_\tau)\,\mathrm{d}\tau + \sigma(X_\tau)\,\mathrm{d}B_\tau. \tag{5.10}$$

The same mechanism appears in the *CLT-scaled* ResNet

$$h_{\ell+1} = h_\ell + \frac{1}{\sqrt{d}}\,\frac{1}{\sqrt{n}} W_\ell\,\varphi(h_\ell), \tag{5.11}$$

where $(W_\ell)_{\ell\geq 0}$ are independent with i.i.d. $\mathcal{N}(0,1)$ entries. Write $\eta := 1/d$ and interpret the *layer time* $\tau := \eta\ell$ (so $\ell \approx \tau/\eta$). Since W_ℓ is independent of $\mathcal{F}_\ell$, the next-layer transition is exactly Gaussian:

$$h_{\ell+1} \mid \mathcal{F}_\ell \sim \mathcal{N}\Big(h_\ell, \eta\,\frac{1}{n}\,\|\varphi(h_\ell)\|^2\, I_n\Big). \tag{5.12}$$

Thus (5.11) is exactly, in transition law, the Euler–Maruyama scheme for

$$\mathrm{d}h_\tau = \frac{1}{\sqrt{n}}\,\|\varphi(h_\tau)\|\,\mathrm{d}B_\tau, \tag{5.13}$$

where B_τ is an n-dimensional Brownian motion. For fixed n and a globally Lipschitz activation (including ReLU), the diffusion coefficient is globally Lipschitz, so standard Euler–Maruyama convergence applies as $d \to \infty$.

For the no-bias Gaussian ReLU specialization of (5.11), Hayou and Yang prove that the limiting one-neuron preactivation law and empirical covariance flow are independent of the relative rates at which n and d tend to infinity, provided $\min\{n,d\} \to \infty$ [20, Theorems 1 and 2]. In particular, the two sequential orders and proportional paths such as $d/n \to \bar{\tau}$ agree for these observables. This is a model-specific commutation result: for plain MLPs, proportional and sequential limits can instead differ. The representative routes are summarized in Figure 5.1.

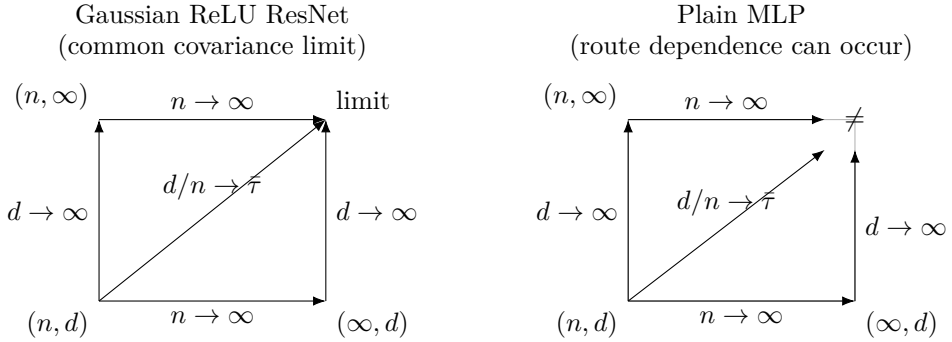

Figure 5.1: Representative depth–width routes for initialization observables. Left: for the no-bias Gaussian ReLU ResNet (5.11), the covariance limit is independent of the route. Right: for plain MLPs, proportional and sequential limits can differ.

5.2 Proportional Depth–Width Limits at Initialization

To see why joint limits appear for plain deep networks, it is instructive to start with the simplest case: a deep *linear* network at initialization.

Fix $n_0 \in \mathbb{N}$ and a nonzero input $x \in \mathbb{R}^{n_0}$. For each width n, let $W_0 \in \mathbb{R}^{n \times n_0}$ and $(W_\ell)_{\ell \geq 1}$, $W_\ell \in \mathbb{R}^{n \times n}$, be mutually independent matrices with i.i.d. $\mathcal{N}(0,1)$ entries. Set

$$h_0 := x, \qquad h_1 := \frac{1}{\sqrt{n_0}} W_0 x, \qquad h_{\ell+1} := \frac{1}{\sqrt{n}} W_\ell h_\ell, \quad \ell \geq 1. \tag{5.14}$$

A depth-d network is the prefix ending at h_d. Define

$$\Phi_0 := \frac{1}{n_0} \|x\|^2, \qquad \Phi_\ell := \frac{1}{n} \|h_\ell\|^2, \quad \ell \geq 1. \tag{5.15}$$

Lemma 5.3 (Multiplicative Variance Recursion)**.** *For every $\ell \geq 0$, conditioned on $\mathcal{F}_\ell$, we have*

$$h_{\ell+1} \,|\, \mathcal{F}_\ell \sim \mathcal{N}(0,\, \Phi_\ell I_n), \tag{5.16}$$

and consequently

$$\Phi_{\ell+1} = \Phi_\ell \frac{1}{n} \|\xi_\ell\|^2, \qquad (\xi_\ell)_{\ell \geq 0} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_n), \tag{5.17}$$

where the innovations are independent of the past.

Proof Sketch. For $\ell \geq 1$, write the i-th row of W_ℓ as $w_{\ell,i}^\top \in \mathbb{R}^n$. Then

$$h_{\ell+1,i} = \frac{1}{\sqrt{n}} w_{\ell,i}^\top h_\ell. \tag{5.18}$$

Each coordinate is centered Gaussian with variance

$$\mathrm{Var}(h_{\ell+1,i} \mid \mathcal{F}_\ell) = \frac{1}{n} \|h_\ell\|^2 = \Phi_\ell, \tag{5.19}$$

and row independence gives the displayed conditional law. The input step $\ell = 0$ is the same calculation in $\mathbb{R}^{n_0}$ with normalization $n_0^{-1/2}$ and variance $n_0^{-1} \|x\|^2 = \Phi_0$. Independence of the layer matrices lets us write $h_{\ell+1} = \sqrt{\Phi_\ell}\, \xi_\ell$ with independent standard Gaussian innovations. Taking squared norms gives

$$\Phi_{\ell+1} = \frac{1}{n} \|h_{\ell+1}\|^2 = \Phi_\ell \frac{1}{n} \|\xi_\ell\|^2, \tag{5.20}$$

which is (5.17). □

Taking logs turns the product into a sum:

$$\log\left(\frac{\Phi_d}{\Phi_0}\right) = \sum_{\ell=0}^{d-1} \log\left(\frac{1}{n} \|\xi_\ell\|^2\right). \tag{5.21}$$

Since $\frac{1}{n} \|\xi_\ell\|^2$ concentrates near 1 with fluctuations of order $n^{-1/2}$, the sum in (5.21) is nontrivial when d is of order n.

Theorem 5.4 (Scalar Linear Variance Limit)**.** *In the joint limit $d, n \to \infty$ with $d/n \to \bar{\tau} \in (0, \infty)$,*

$$\log\left(\frac{\Phi_d}{\Phi_0}\right) \xrightarrow{d} \mathcal{N}(-\bar{\tau},\, 2\bar{\tau}). \tag{5.22}$$

This endpoint law is the equal-hidden-width Gaussian specialization of [19, Theorem 1 and Eq. (9)].

More strongly, define $\Phi_\tau^{(n)} := \Phi_{\lfloor n\tau \rfloor}$. *For every fixed* $T > 0$, $\Phi^{(n)} \xrightarrow{d} \Phi$ *in* $D([0,T],\mathbb{R})$, *where* Φ *is the geometric Brownian motion*

$$\mathrm{d}\Phi_\tau = \sqrt{2}\,\Phi_\tau\,\mathrm{d}B_\tau, \qquad \Phi_0 = \frac{1}{n_0}\,\|x\|^2\,. \tag{5.23}$$

The endpoint convergence in (5.22) *is a consequence of this functional limit.*

Proof Sketch. By Lemma 5.3, the scalar variance satisfies the multiplicative recursion

$$\Phi_{\ell+1} = \Phi_\ell\, U_{\ell,n}, \qquad U_{\ell,n} := \frac{1}{n}\,\|\xi_\ell\|^2\,, \qquad \xi_\ell \sim \mathcal{N}(0, I_n) \text{ i.i.d.} \tag{5.24}$$

Equivalently,

$$\Phi_{\ell+1} - \Phi_\ell = \Phi_\ell\,(U_{\ell,n} - 1). \tag{5.25}$$

Since $\|\xi_\ell\|^2 \sim \chi^2_n$, we have $\mathbb{E}[U_{\ell,n} - 1] = 0$, $\mathrm{Var}(U_{\ell,n} - 1) = 2/n$, and $\mathbb{E}[|U_{\ell,n} - 1|^3] = O(n^{-3/2})$. Thus, conditioning on the forward history gives

$$\begin{aligned} n\,\mathbb{E}[\Phi_{\ell+1} - \Phi_\ell \mid \mathcal{F}_\ell] &= 0, \\ n\ \mathrm{Var}(\Phi_{\ell+1} - \Phi_\ell \mid \mathcal{F}_\ell) &= 2\Phi_\ell^2, \\ n\,\mathbb{E}[|\Phi_{\ell+1} - \Phi_\ell|^3 \mid \mathcal{F}_\ell] &= O(n^{-1/2})\Phi_\ell^3. \end{aligned} \tag{5.26}$$

These conditional moments depend on the history only through Φ_ℓ, so they also give the Markov-state coefficients uniformly on compact sets. They match the zero drift and diffusion covariance $2\Phi^2$ of (5.23), with a vanishing third-moment remainder. Thus the interpolated chain has the corresponding local diffusion approximation; the precise sufficient result is deferred to Theorem A.4. The prelimit chain and the solution of (5.23) are nonnegative martingales, so Doob's inequality gives, for each $T > 0$,

$$\mathbb{P}\left(\sup_{0\le\tau\le T} \Phi_\tau^{(n)} \ge R\right) \le \frac{\Phi_0}{R}, \qquad \mathbb{P}\left(\sup_{0\le\tau\le T} \Phi_\tau \ge R\right) \le \frac{\Phi_0}{R}. \tag{5.27}$$

Letting $R \to \infty$ removes the compact stopping and proves the stated functional convergence.

Now let $\tau_n := d/n \to \bar\tau$ and choose $T > \bar\tau$. Since the limiting process is continuous, evaluation at τ_n gives $\Phi_d = \Phi_{\tau_n}^{(n)} \xrightarrow{d} \Phi_{\bar\tau}$. The geometric Brownian motion is strictly positive, so the continuous mapping theorem gives $\log(\Phi_d/\Phi_0) \xrightarrow{d} \log(\Phi_{\bar\tau}/\Phi_0)$. Itô's lemma applied to $g(\Phi) = \log\Phi$ gives

$$\mathrm{d}\log\Phi_\tau = \frac{1}{\Phi_\tau}\mathrm{d}\Phi_\tau - \frac{1}{2}\frac{1}{\Phi_\tau^2}(\mathrm{d}\Phi_\tau)^2 = -\mathrm{d}\tau + \sqrt{2}\,\mathrm{d}B_\tau, \tag{5.28}$$

so $\log\Phi_{\bar\tau} - \log\Phi_0 \sim \mathcal{N}(-\bar\tau,\, 2\bar\tau)$, which proves (5.22). □

The term *geometric* in (5.23) refers to its multiplicative noise: randomness scales with Φ_τ itself. This is the probabilistic signature of deep products, in which each layer randomly stretches or shrinks the current state.

Even in this simplest linear setting, the proportional regime already produces *non-Gaussian* marginal laws. For any fixed coordinate i, $h_{d,i} \mid \mathcal{F}_{d-1} \sim \mathcal{N}(0, \Phi_{d-1})$. Since $\Phi_{d-1} \xrightarrow{d} \Phi_{\bar{\tau}}$ when $d/n \to \bar{\tau}$, the limiting marginal law is a Gaussian scale mixture with random mixing variance $\Phi_{\bar{\tau}}$. This is a key motivation for taking a *joint* depth–width limit.

Here "stable" means order one in a two-sided probabilistic sense. For a positive random sequence X_k, the condition $X_k = O_{\mathbb{P}}(1)$ only rules out escape to infinity and still permits collapse to zero. We write $X_k = \Theta_{\mathbb{P}}(1)$ when both X_k and X_k^{-1} are $O_{\mathbb{P}}(1)$, so typical realizations remain bounded away from both zero and infinity.

The three depth–width regimes behave differently in this sense. If $n \to \infty$ first at fixed d, the layerwise fluctuations are averaged away and Φ_d converges in probability to the deterministic order-one value Φ_0. If $d \to \infty$ first at fixed n, repeated multiplication instead sends $\Phi_d \to 0$ almost surely, even though $\mathbb{E}[\Phi_d] = \Phi_0$ for every d. Thus $\Phi_d = O_{\mathbb{P}}(1)$ but not $\Theta_{\mathbb{P}}(1)$: increasingly rare large realizations preserve the mean while a typical realization collapses. Under the proportional scaling $d/n \to \bar{\tau}$, the variance converges to the positive, finite, nondegenerate random variable $\Phi_{\bar{\tau}}$, so it remains $\Theta_{\mathbb{P}}(1)$ while retaining nontrivial randomness [19, §1.1, Theorem 1, and Eq. (9)].

Although the discussion so far concerns initialization, proportional depth–width scaling also changes the NTK picture for fully connected ReLU networks. For equal hidden width n, Hanin and Nica show that the on-diagonal NTK at initialization need not concentrate when d/n has a nonzero limit [18, Theorem 1]. Their training-time result is narrower: for square loss and a single-example SGD step, they compute a potentially non-negligible mean first update of the same kernel entry. This motivates, but does not prove, their conjectured "weak feature learning" regime [18, Theorem 2 and §3.2].

5.3 Multiple Data Points: A Covariance SDE

We write $\mathrm{Sym}(m)$ for the vector space of real symmetric $m \times m$ matrices. The empirical covariance matrices below are positive semidefinite; they need not be strictly positive definite.

Now let $m \geq 2$ and propagate inputs $x^1, \ldots, x^m \in \mathbb{R}^{n_0}$ through the same random network. For the linear model (5.14), define the input and hidden-layer Gram matrices by

$$\begin{aligned} \Phi_0^{\alpha\beta} &:= \frac{1}{n_0} \left\langle x^\alpha, x^\beta \right\rangle, \qquad \Phi_\ell^{\alpha\beta} := \frac{1}{n} \left\langle h_\ell^\alpha, h_\ell^\beta \right\rangle, \quad \ell \geq 1, \\ \Phi_\ell &:= [\Phi_\ell^{\alpha\beta}]_{\alpha,\beta=1}^m \in \mathrm{Sym}(m), \quad \ell \geq 0. \end{aligned} \tag{5.29}$$

Let $\mathcal{F}_0 := \{\varnothing, \Omega\}$ be the trivial sigma-field. For $\ell \geq 1$, define

$$\mathcal{F}_\ell := \sigma\Big(\{h_k^\alpha\}_{1 \leq k \leq \ell,\ \alpha \in [m]} \Big). \tag{5.30}$$

Thus $\mathcal{F}_\ell$ records the forward hidden-state history through layer ℓ. Conditioned on $\mathcal{F}_\ell$, $\{h_{\ell+1}^\alpha\}_{\alpha=1}^m$ is jointly Gaussian with covariance $\Phi_\ell \otimes I_n$. As a result, each entry of the centered increment $\Phi_{\ell+1} - \Phi_\ell$ is an empirical average of conditionally i.i.d. centered terms. For $1 \leq \alpha \leq \beta \leq m$, write

$$\Phi_{\ell+1}^{\alpha\beta} = \Phi_\ell^{\alpha\beta} + \frac{1}{\sqrt{n}} R_\ell^{\alpha\beta}, \qquad \mathbb{E}[R_\ell^{\alpha\beta} \mid \mathcal{F}_\ell] = 0, \tag{5.31}$$

where

$$R_\ell^{\alpha\beta} := \frac{1}{\sqrt{n}} \sum_{i=1}^n \Big(h_{\ell+1,i}^\alpha h_{\ell+1,i}^\beta - \Phi_\ell^{\alpha\beta} \Big). \tag{5.32}$$

The next step is to compute the conditional covariance between these fluctuation entries.

Proposition 5.5 (Isserlis Covariance for the Gram Increment)**.** *Let indices satisfy $\alpha \le \beta$ and $\gamma \le \delta$ (to avoid duplication from symmetry), and let $R_\ell^{\alpha\beta}$ be the centered Gram fluctuation defined in* (5.32)*. Then, conditional on $\mathcal{F}_\ell$, the neuron-wise vectors*

$$h_{\ell+1,i} := (h^1_{\ell+1,i}, \dots, h^m_{\ell+1,i}) \in \mathbb{R}^m, \qquad i = 1, \dots, n, \tag{5.33}$$

are i.i.d. Gaussian with law $\mathcal{N}(0, \Phi_\ell)$, and we have the exact *identities*

$$\mathbb{E}\Big[R_\ell^{\alpha\beta} \mid \mathcal{F}_\ell\Big] = 0, \qquad \mathrm{Cov}\Big(R_\ell^{\alpha\beta}, R_\ell^{\gamma\delta} \mid \mathcal{F}_\ell\Big) = \Phi_\ell^{\alpha\gamma}\,\Phi_\ell^{\beta\delta} + \Phi_\ell^{\alpha\delta}\,\Phi_\ell^{\beta\gamma}. \tag{5.34}$$

Equivalently, in the flattened coordinates on $\mathrm{Sym}(m)$*, the conditional covariance matrix is $\Sigma(\Phi_\ell)$ with entries*

$$\Sigma_{\alpha\beta,\gamma\delta}(\Phi) := \Phi^{\alpha\gamma}\Phi^{\beta\delta} + \Phi^{\alpha\delta}\Phi^{\beta\gamma}. \tag{5.35}$$

Proof. Fix ℓ and condition on $\mathcal{F}_\ell$. For each i, set

$$R^{\alpha\beta}_{\ell,i} := h^\alpha_{\ell+1,i}\, h^\beta_{\ell+1,i} - \Phi_\ell^{\alpha\beta}, \qquad \text{so that} \qquad R_\ell^{\alpha\beta} = \frac{1}{\sqrt{n}} \sum_{i=1}^n R^{\alpha\beta}_{\ell,i}. \tag{5.36}$$

Since $(h_{\ell+1,i})_{i=1}^n$ are conditionally i.i.d. $\mathcal{N}(0, \Phi_\ell)$, the centered variables $(R^{\alpha\beta}_{\ell,i})_{i=1}^n$ are conditionally i.i.d. as well, and therefore

$$\mathrm{Cov}\Big(R_\ell^{\alpha\beta}, R_\ell^{\gamma\delta} \mid \mathcal{F}_\ell\Big) = \mathbb{E}\Big[R^{\alpha\beta}_{\ell,1} R^{\gamma\delta}_{\ell,1} \mid \mathcal{F}_\ell\Big]. \tag{5.37}$$

Expanding and using $\mathbb{E}[h^\alpha_{\ell+1,1} h^\beta_{\ell+1,1} \mid \mathcal{F}_\ell] = \Phi_\ell^{\alpha\beta}$ gives

$$\mathbb{E}\Big[R^{\alpha\beta}_{\ell,1} R^{\gamma\delta}_{\ell,1} \mid \mathcal{F}_\ell\Big] = \mathbb{E}\Big[h^\alpha_{\ell+1,1} h^\beta_{\ell+1,1} h^\gamma_{\ell+1,1} h^\delta_{\ell+1,1} \mid \mathcal{F}_\ell\Big] - \Phi_\ell^{\alpha\beta}\Phi_\ell^{\gamma\delta}. \tag{5.38}$$

By Isserlis' theorem for the centered Gaussian vector $h_{\ell+1,1}$ (a.k.a. Wick's formula),

$$\mathbb{E}\Big[h^\alpha_{\ell+1,1} h^\beta_{\ell+1,1} h^\gamma_{\ell+1,1} h^\delta_{\ell+1,1} \mid \mathcal{F}_\ell\Big] = \Phi_\ell^{\alpha\beta}\Phi_\ell^{\gamma\delta} + \Phi_\ell^{\alpha\gamma}\Phi_\ell^{\beta\delta} + \Phi_\ell^{\alpha\delta}\Phi_\ell^{\beta\gamma}, \tag{5.39}$$

and subtracting the first term yields the claimed covariance. □

By Proposition 5.5, one Gram increment has conditional covariance $n^{-1}\Sigma(\Phi_\ell)$. Thus, on the layer-time scale $\tau = \ell/n$, the fluctuations from $O(n)$ layers accumulate at order one. The following theorem formalizes this process-level limit.

Theorem 5.6 (Linear Covariance SDE Limit [34, Theorem 3.2 and Appendix C.1])**.** *For each n, define the piecewise-constant interpolation $\Phi^{(n)}_\tau := \Phi_{\lfloor n\tau \rfloor}$, $\tau \ge 0$. Let $\bar m := m(m+1)/2$. After flattening Φ into its $\bar m$ distinct upper-triangular entries, for every $\bar\tau > 0$, as $n \to \infty$, $\Phi^{(n)}$ converges in distribution to Φ in $D([0, \bar\tau], \mathbb{R}^{\bar m})$, where Φ solves*

$$\mathrm{d}\Phi_\tau = \Sigma(\Phi_\tau)^{1/2}\, \mathrm{d}B_\tau, \qquad \Phi_0 = \left[\frac{1}{n_0}\left\langle x^\alpha, x^\beta \right\rangle\right]_{\alpha,\beta=1}^m. \tag{5.40}$$

Here B is standard Brownian motion in $\mathbb{R}^{\bar m}$, and Σ is the covariance matrix in (5.35)*.*

Proof Sketch. By (5.31) and Proposition 5.5, one layer has conditional mean zero and conditional covariance $n^{-1}\Sigma(\Phi_\ell)$. The same Gaussian moment bounds verify the third-moment hypothesis of Theorem A.4. These conditional moments depend on the history only through Φ_ℓ, so they are the Markov-state coefficients required by Theorem A.4. We have therefore matched the zero drift and diffusion covariance Σ, yielding the stated diffusion approximation; the precise technical result is deferred to Theorem A.4. □

When $m = 1$, (5.35) gives $\Sigma(\Phi) = 2\Phi^2$, so Theorem 5.6 reduces to the geometric Brownian motion in Theorem 5.4. If the inputs are linearly dependent, those linear relations are preserved by every linear layer. One may therefore apply the theorem to a maximal linearly independent subset and reconstruct the full Gram matrix from it.

The same covariance tensor has a geometric interpretation: on $\mathrm{SPD}(m)$, $\Sigma(\Phi)$ is the cometric, or inverse metric tensor, associated with the affine-invariant Riemannian metric. See Section 6.2.2 for this metric interpretation and Section 6.2.3 for the resulting eigenvalue dynamics.

5.4 Nonlinearities, Correlation Collapse, and Shaping

For nonlinear activations, the depth limit depends strongly on normalization and on how quickly correlations are driven toward a fixed point.

Consider first the ReLU activation $\varphi(u) = \max\{u, 0\}$. If $g \sim \mathcal{N}(0, \sigma^2)$ then $\mathbb{E}[\varphi(g)^2] = \sigma^2/2$. Thus, for the naive ReLU MLP

$$h_{\ell+1} = \frac{1}{\sqrt{n}} W_\ell \varphi(h_\ell), \tag{5.41}$$

the variance contracts by a factor $1/2$ per layer in expectation. The standard fix is Kaiming normalization [21]: choose $c > 0$ so that $c^{-1} = \mathbb{E}[\varphi(\omega)^2]$ for $\omega \sim \mathcal{N}(0, 1)$, and scale the layer by $\sqrt{c}$. For ReLU, $c = 2$.

Proposition 5.7 (ReLU SDE Limit for One Sample)**.** *Fix $n_0 \in \mathbb{N}$ and $h_0 := x \in \mathbb{R}^{n_0} \setminus \{0\}$. Consider the Kaiming-normalized ReLU network*

$$h_1 = \frac{1}{\sqrt{n_0}} W_0 x, \qquad h_{\ell+1} = \sqrt{\frac{c}{n}}\, W_\ell\, \varphi(h_\ell), \qquad \ell \geq 1, \qquad c = 2. \tag{5.42}$$

Here $W_0 \in \mathbb{R}^{n \times n_0}$ and $W_\ell \in \mathbb{R}^{n \times n}$ for $\ell \geq 1$ are mutually independent matrices with i.i.d. $\mathcal{N}(0, 1)$ entries. Define the (normalized) squared post-activation norm

$$\Phi_0 := \frac{1}{n_0}\|x\|_2^2, \qquad \Phi_\ell := \frac{c}{n}\,\|\varphi(h_\ell)\|_2^2, \quad \ell \geq 1. \tag{5.43}$$

Let $d, n \to \infty$ with $d/n \to \bar{\tau} \in (0, \infty)$, and for $\tau \in [0, \bar{\tau}]$ set $\Phi_\tau^{(n)} := \Phi_{\lfloor \tau n \rfloor}$. Then $\Phi^{(n)}$ converges in distribution to the geometric Brownian motion Φ in $D([0, \bar{\tau}], \mathbb{R})$ with the Skorohod topology (see Appendix A.3), where

$$\mathrm{d}\Phi_\tau = \sqrt{5}\,\Phi_\tau\,\mathrm{d}B_\tau, \tag{5.44}$$

with the same initial condition. At the terminal depth,

$$\log\left(\frac{\Phi_d}{\Phi_0}\right) \xrightarrow{d} \mathcal{N}\left(-\frac{5}{2}\bar{\tau},\ 5\bar{\tau}\right). \tag{5.45}$$

The proof is deferred to Exercise 5.14.

For one input, Kaiming normalization gives $\mathbb{E}[\Phi_{\ell+1} \mid \mathcal{F}_\ell] = \Phi_\ell$; the remaining order-$n^{-1/2}$ finite-width fluctuation accumulates into the SDE of Proposition 5.7. This controls only the diagonal of the multi-input Gram matrix; the off-diagonal correlation update remains the obstruction. For two inputs with nonzero postactivation norms, define the scaled postactivation Gram entry and its correlation at layer ℓ by

$$\Phi_\ell^{\alpha\beta} := \frac{c}{n}\left\langle \varphi(h_\ell^\alpha), \varphi(h_\ell^\beta)\right\rangle, \qquad \rho_\ell^{\alpha\beta} := \frac{\Phi_\ell^{\alpha\beta}}{\sqrt{\Phi_\ell^{\alpha\alpha}\Phi_\ell^{\beta\beta}}} \in [-1,1]. \tag{5.46}$$

In the infinite-width (fixed-depth) limit, the Cho–Saul formula in Proposition 2.5, applied with unit variances, gives the ReLU correlation update

$$\rho_{\ell+1}^{\alpha\beta} = \mathcal{T}(\rho_\ell^{\alpha\beta}), \qquad \mathcal{T}(\rho) := \frac{\mathbb{E}[\varphi(w_1)\varphi(w_2)]}{\mathbb{E}[\varphi(g)^2]} = \frac{1}{\pi}\left(\sqrt{1-\rho^2} + (\pi - \arccos\rho)\,\rho\right), \tag{5.47}$$

where $g \sim \mathcal{N}(0,1)$ and (w_1, w_2) is a jointly standard Gaussian pair with correlation ρ. For $-1 \le \rho < 1$, one has $\mathcal{T}(\rho) > \rho$; hence 1 is the unique fixed point in $[-1,1]$, and every correlation recursion starting below 1 converges to 1.

Remark 5.8 (Representation and Kernel Collapse). As an off-diagonal correlation approaches 1, the corresponding normalized postactivation representations align and their normalized kernel entry approaches its diagonal value. This is a forward representation/kernel-collapse statement; it does not by itself imply that backpropagated gradients vanish.

Away from 1, the map $\mathcal{T}$ in (5.47) differs from the identity by an order-one amount, so the unshaped off-diagonal update is not on the order-n^{-1} drift scale needed over $d \asymp n$ layers. Activation shaping repairs this scale mismatch: the full normalized kernel map becomes the identity plus an order-n^{-1} bias while the empirical Gram fluctuation remains of order $n^{-1/2}$. This viewpoint is closely related to *deep kernel shaping* and *tailored rectifiers* [35, 52], and it underlies the covariance SDE limits of [34, Sections 2.3–3.2].

For concreteness, fix inputs $x^1, \dots, x^m \in \mathbb{R}^{n_0}$ and consider the Kaiming-normalized network

$$h_1^\alpha = \frac{1}{\sqrt{n_0}} W_0 x^\alpha, \qquad h_{\ell+1}^\alpha = \sqrt{\frac{c}{n}}\, W_\ell\, \varphi_s(h_\ell^\alpha), \quad 1 \le \ell \le d-1, \tag{5.48}$$

where $W_0 \in \mathbb{R}^{n\times n_0}$ and $W_\ell \in \mathbb{R}^{n\times n}$ are independent with i.i.d. $\mathcal{N}(0,1)$ entries, and $c = c(n)$ is chosen so that $\mathbb{E}[\varphi_s(g)^2] = c^{-1}$ for $g \sim \mathcal{N}(0,1)$. Define the input boundary and the scaled postactivation Gram matrices by

$$\begin{aligned} \Phi_0^{\alpha\beta} &:= \frac{1}{n_0}\left\langle x^\alpha, x^\beta\right\rangle, \\ \Phi_\ell^{\alpha\beta} &:= \frac{c}{n}\left\langle \varphi_s(h_\ell^\alpha), \varphi_s(h_\ell^\beta)\right\rangle, \quad 1 \le \ell \le d, \qquad \Phi_\ell := \left(\Phi_\ell^{\alpha\beta}\right)_{\alpha,\beta=1}^m \in \mathrm{Sym}(m). \end{aligned} \tag{5.49}$$

Thus Φ_ℓ for $\ell \ge 1$ is the postactivation Gram matrix at layer ℓ. Every empirical Φ_ℓ is positive semidefinite. We assume that the data Gram matrix Φ_0 is strictly positive definite, although the one-step calculations below remain valid for positive-semidefinite covariances. The dependence of the finite-width chain Φ_ℓ on n is suppressed. With $\mathcal{F}_\ell$ as in (5.30), conditional on $\mathcal{F}_\ell$ the vectors $h_{\ell+1,i} = (h_{\ell+1,i}^1, \dots, h_{\ell+1,i}^m)$, $i = 1, \dots, n$, are i.i.d. $\mathcal{N}(0, \Phi_\ell)$ for $0 \le \ell \le d-1$.

ReLU-like shaping. For $p > 0$, let

$$\varphi_s(x) = s_+ \max\{x, 0\} + s_- \min\{x, 0\}, \qquad s_\pm = 1 + \frac{c_\pm}{n^p}. \tag{5.50}$$

At the critical exponent $p = \frac{1}{2}$, this is [34, Definition 3.1]; the general exponent family is analyzed in [34, Proposition 3.4]. The exponent $p = \frac{1}{2}$ is critical because the deterministic bias of the normalized kernel is then $O(n^{-1})$ and accumulates over $d \asymp n$ layers.

Smooth shaping. Assume $\varphi \in C^4(\mathbb{R})$, $\varphi(0) = 0$, $\varphi'(0) = 1$, and $\left|\varphi^{(4)}(x)\right| \le C\,(1 + |x|^q)$ for some $C, q > 0$, as in [34, Assumption 3.5]. For a constant $a > 0$, use the scaling

$$\varphi_s(x) := s\,\varphi(x/s), \qquad s = a\sqrt{n}, \qquad c^{-1} = \mathbb{E}[\varphi_s(g)^2], \quad g \sim \mathcal{N}(0, 1). \tag{5.51}$$

Then $\varphi_s(x) \to x$ pointwise as $n \to \infty$, with the fourth-derivative bound controlling the Taylor remainder in expectation, so the kernel map is again a small perturbation of the identity [34, Definition 3.6].

Conditional remainder notation. For a matrix-valued random variable X, possibly depending on ℓ and n, we write $X = \widetilde{o}_\ell(n^{-1})$ if

$$\|\mathbb{E}[X \mid \mathcal{F}_\ell]\|_{\mathrm{F}} = o(n^{-1}), \qquad \mathbb{E}\Big[\|X - \mathbb{E}[X \mid \mathcal{F}_\ell]\|_{\mathrm{F}}^2 \;\Big|\; \mathcal{F}_\ell\Big] = o(n^{-1}). \tag{5.52}$$

The subscript in $\widetilde{o}_\ell(n^{-1})$ records conditioning on $\mathcal{F}_\ell$. Thus its conditional mean is $o(n^{-1})$, while its centered part is $o(n^{-1/2})$ in conditional L^2; this does not assert $o_{\mathbb{P}}(n^{-1})$. The conditioning rules out predictable one-step errors that disappear only after averaging over the current state but can accumulate over layers.

Theorem 5.9 (Euler–Maruyama Increment Expansion under Critical Shaping)**.** *Assume either ReLU-like shaping* (5.50) *at the critical exponent $p = \frac{1}{2}$, or smooth shaping* (5.51) *with $s = a\sqrt{n}$. After flattening the upper-triangular entries as coordinates on* $\mathrm{Sym}(m)$*, one can write*

$$\Phi_{\ell+1} - \Phi_\ell = \frac{1}{n}\, b(\Phi_\ell) + \frac{1}{\sqrt{n}}\, \Sigma(\Phi_\ell)^{1/2} \xi_\ell + \widetilde{o}_\ell(n^{-1}), \tag{5.53}$$

where $\widetilde{o}_\ell(n^{-1})$ is understood in the sense of (5.52)*. The noise vector ξ_ℓ satisfies*

$$\mathbb{E}[\xi_\ell \mid \mathcal{F}_\ell] = 0, \qquad \mathrm{Cov}(\xi_\ell \mid \mathcal{F}_\ell) = I_{\bar{m}}, \qquad \bar{m} := \frac{m(m+1)}{2}, \tag{5.54}$$

and Σ is the covariance matrix in (5.35)*. The drift b is given as follows.*

- *In the ReLU-like shaping scheme, one has the entrywise formula*

$$\begin{aligned} b^{\alpha\beta}(\Phi) &= \sqrt{\Phi^{\alpha\alpha}\Phi^{\beta\beta}}\, \nu(\rho^{\alpha\beta}), \qquad \rho^{\alpha\beta} := \frac{\Phi^{\alpha\beta}}{\sqrt{\Phi^{\alpha\alpha}\Phi^{\beta\beta}}}, \\ \nu(\rho) &:= \frac{(c_+ - c_-)^2}{2\pi}\Big(\sqrt{1-\rho^2} - \rho \arccos(\rho)\Big). \end{aligned} \tag{5.55}$$

- *In the smooth-shaping scheme, the drift admits the entrywise form*

$$b^{\alpha\beta}(\Phi) = \frac{\varphi''(0)^2}{4a^2}\Big(\Phi^{\alpha\alpha}\Phi^{\beta\beta} + \Phi^{\alpha\beta}\big(2\Phi^{\alpha\beta} - 3\big)\Big) + \frac{\varphi'''(0)}{2a^2}\,\Phi^{\alpha\beta}\Big(\Phi^{\alpha\alpha} + \Phi^{\beta\beta} - 2\Big). \tag{5.56}$$

The ReLU-like formula is understood by continuity if a diagonal covariance entry vanishes.

Proof. Fix ℓ and condition on $\mathcal{F}_\ell$. Since W_ℓ has i.i.d. Gaussian rows, the neuron-wise preactivation vectors

$$h_{\ell+1,i} := (h^1_{\ell+1,i}, \dots, h^m_{\ell+1,i}) \in \mathbb{R}^m, \qquad i = 1, \dots, n, \tag{5.57}$$

are conditionally i.i.d. $\mathcal{N}(0, \Phi_\ell)$.

Drift calculation. Let $w \mid \mathcal{F}_\ell \sim \mathcal{N}(0, \Phi_\ell)$. In the ReLU-like case, set $\rho_\ell^{\alpha\beta} := \Phi_\ell^{\alpha\beta}/\sqrt{\Phi_\ell^{\alpha\alpha}\Phi_\ell^{\beta\beta}}$ and let (z_1, z_2) be a standard Gaussian pair with correlation $\rho_\ell^{\alpha\beta}$. Positive homogeneity and the corresponding arc-cosine calculation [34, Proposition 3.4] give, at $p = \frac{1}{2}$,

$$\begin{aligned} \mathbb{E}[c\,\varphi_s(z_1)\varphi_s(z_2)] &= \rho_\ell^{\alpha\beta} + \frac{1}{n}\nu(\rho_\ell^{\alpha\beta}) + O(n^{-3/2}), \\ \mathbb{E}\Big[c\,\varphi_s(w^\alpha)\varphi_s(w^\beta) \mid \mathcal{F}_\ell\Big] &= \sqrt{\Phi_\ell^{\alpha\alpha}\Phi_\ell^{\beta\beta}}\left(\rho_\ell^{\alpha\beta} + \frac{1}{n}\nu(\rho_\ell^{\alpha\beta}) + O(n^{-3/2})\right) \\ &= \Phi_\ell^{\alpha\beta} + \frac{1}{n}b^{\alpha\beta}(\Phi_\ell) + O(n^{-3/2}), \end{aligned} \tag{5.58}$$

which yields (5.55).

In the smooth case, Taylor expansion at the origin gives, in conditional L^4,

$$\varphi_s(x) = x + \frac{\varphi''(0)}{2a\sqrt{n}}x^2 + \frac{\varphi'''(0)}{6a^2 n}x^3 + O_{L^4}(n^{-3/2}). \tag{5.59}$$

Taking $g \sim \mathcal{N}(0,1)$ in (5.59) and using the Gaussian moments gives

$$c^{-1} = 1 + \frac{1}{a^2 n}\left(\frac{3}{4}\varphi''(0)^2 + \varphi'''(0)\right) + o(n^{-1}), \qquad c = 1 - \frac{1}{a^2 n}\left(\frac{3}{4}\varphi''(0)^2 + \varphi'''(0)\right) + o(n^{-1}). \tag{5.60}$$

The odd cubic terms have zero expectation, while Isserlis' formula gives $\mathbb{E}[(w^\alpha)^3 w^\beta \mid \mathcal{F}_\ell] = 3\Phi_\ell^{\alpha\alpha}\Phi_\ell^{\alpha\beta}$ and $\mathbb{E}[(w^\alpha)^2(w^\beta)^2 \mid \mathcal{F}_\ell] = \Phi_\ell^{\alpha\alpha}\Phi_\ell^{\beta\beta} + 2(\Phi_\ell^{\alpha\beta})^2$. Consequently,

$$\begin{aligned} &\mathbb{E}\Big[\varphi_s(w^\alpha)\varphi_s(w^\beta) \mid \mathcal{F}_\ell\Big] \\ &\quad = \Phi_\ell^{\alpha\beta} + \frac{\varphi''(0)^2}{4a^2 n}\left(\Phi_\ell^{\alpha\alpha}\Phi_\ell^{\beta\beta} + 2(\Phi_\ell^{\alpha\beta})^2\right) + \frac{\varphi'''(0)}{2a^2 n}\,\Phi_\ell^{\alpha\beta}\left(\Phi_\ell^{\alpha\alpha} + \Phi_\ell^{\beta\beta}\right) + o(n^{-1}). \end{aligned} \tag{5.61}$$

Multiplying (5.61) by (5.60) gives

$$\mathbb{E}\Big[c\,\varphi_s(w^\alpha)\varphi_s(w^\beta) \mid \mathcal{F}_\ell\Big] = \Phi_\ell^{\alpha\beta} + \frac{1}{n}b^{\alpha\beta}(\Phi_\ell) + o(n^{-1}), \tag{5.62}$$

with b exactly as in (5.56).

Covariance calculation. For each α, β, define

$$R_{\ell,i}^{\alpha\beta} := c\,\varphi_s(h^\alpha_{\ell+1,i})\varphi_s(h^\beta_{\ell+1,i}) - \mathbb{E}\Big[c\,\varphi_s(w^\alpha)\varphi_s(w^\beta) \mid \mathcal{F}_\ell\Big], \qquad R_\ell^{\alpha\beta} := \frac{1}{\sqrt{n}}\sum_{i=1}^n R_{\ell,i}^{\alpha\beta}. \tag{5.63}$$

Then

$$\Phi_{\ell+1}^{\alpha\beta} - \Phi_\ell^{\alpha\beta} = \frac{1}{n}b^{\alpha\beta}(\Phi_\ell) + \frac{1}{\sqrt{n}}R_\ell^{\alpha\beta} + o(n^{-1}), \tag{5.64}$$

where the final term is the conditional-mean remainder from the preceding drift calculation. In both shaping schemes, $c\,\varphi_s(h^{\alpha}_{\ell+1,1})\varphi_s(h^{\beta}_{\ell+1,1}) = h^{\alpha}_{\ell+1,1}h^{\beta}_{\ell+1,1} + O(n^{-1/2})$ in conditional L^2. After centering, this gives $R^{\alpha\beta}_{\ell,1} = h^{\alpha}_{\ell+1,1}h^{\beta}_{\ell+1,1} - \Phi^{\alpha\beta}_{\ell} + O(n^{-1/2})$ in conditional L^2. Conditional on $\mathcal{F}_\ell$, the leading term has the same law as the Gaussian-product fluctuation computed in Proposition 5.5 and (5.35). Conditional independence across i then gives

$$\begin{aligned}\operatorname{Cov}(R^{\alpha\beta}_{\ell}, R^{\gamma\delta}_{\ell} \mid \mathcal{F}_\ell) &= \operatorname{Cov}(R^{\alpha\beta}_{\ell,1}, R^{\gamma\delta}_{\ell,1} \mid \mathcal{F}_\ell)\\ &= \operatorname{Cov}(w^{\alpha}w^{\beta}, w^{\gamma}w^{\delta} \mid \mathcal{F}_\ell) + o(1)\\ &= \Sigma_{\alpha\beta,\gamma\delta}(\Phi_\ell) + o(1).\end{aligned} \tag{5.65}$$

Let R_ℓ now denote the vector of its distinct upper-triangular entries. Then, in flattened coordinates,

$$\Sigma_n(\Phi_\ell) := \operatorname{Cov}(R_\ell \mid \mathcal{F}_\ell) = \Sigma(\Phi_\ell) + o(1). \tag{5.66}$$

One may choose ξ_ℓ with the conditional moments in (5.54) so that

$$R_\ell = \Sigma_n(\Phi_\ell)^{1/2}\xi_\ell. \tag{5.67}$$

If $\Sigma_n(\Phi_\ell)$ is singular, its zero-variance directions do not contribute to R_ℓ; one can add independent unit-variance coordinates in those directions so that $\operatorname{Cov}(\xi_\ell \mid \mathcal{F}_\ell) = I_{\bar{m}}$ without changing the factorization. By this convergence and continuity of the positive-semidefinite square root,

$$\frac{1}{\sqrt{n}}R_\ell = \frac{1}{\sqrt{n}}\Sigma(\Phi_\ell)^{1/2}\xi_\ell + \widetilde{o}_\ell(n^{-1}). \tag{5.68}$$

Combining this identity with (5.58) or (5.62) proves (5.53). The same Gaussian moment bounds give the conditional Lindeberg estimate used in the diffusion approximation below; see [34, Appendices C–D] for the corresponding uniform estimates. □

The expansion (5.53) has the form of one Euler–Maruyama step for a diffusion with drift b, diffusion covariance Σ, and diffusion coefficient $\Sigma^{1/2}$. This leads to the following process-level limit (see [34]).

Theorem 5.10 (Neural Covariance SDE Limit [34])**.** *Fix $m \in \mathbb{N}$ and inputs $x^1, \dots, x^m$. Let $\bar{m} := m(m+1)/2$, and let $\{\Phi_\ell\}_{\ell=0}^{d}$ be the shaped Gram-matrix Markov chain defined by* (5.49)*. Assume a critical shaping scheme as in Theorem 5.9, and suppose the input Gram matrix Φ_0 is strictly positive definite. Define the piecewise-constant interpolation*

$$\Phi^{(n)}_{\tau} := \Phi_{\min\{\lfloor \tau n \rfloor, d\}}, \qquad \tau \geq 0. \tag{5.69}$$

For $r > 0$, set

$$\tau^{(n)}_r := \inf\{\tau \geq 0 : \|\Phi^{(n)}_\tau\| > r\}, \qquad \tau_r := \inf\{\tau \geq 0 : \|\Phi_\tau\| > r\}, \tag{5.70}$$

where $\|\cdot\|$ denotes the Euclidean norm of the upper triangular entries. If $n, d \to \infty$ with $d/n \to \bar{\tau} \in (0, \infty)$, then for each fixed $r > 0$,

$$\Phi^{(n)}_{\cdot\wedge\tau^{(n)}_r} \xrightarrow{d} \Phi_{\cdot\wedge\tau_r} \qquad \text{in } D\big([0,\bar{\tau}], \mathbb{R}^{\bar{m}}\big), \tag{5.71}$$

where the upper-triangular entries are used as coordinates, the convergence is in the Skorohod topology, and Φ *is the unique strong solution of*

$$\mathrm{d}\Phi_\tau = b(\Phi_\tau)\,\mathrm{d}\tau + \Sigma(\Phi_\tau)^{1/2}\,\mathrm{d}B_\tau, \qquad \Phi_0 \in \mathrm{SPD}(m), \tag{5.72}$$

with drift b *and diffusion tensor* Σ *as in Theorem 5.9, and with* B_τ *a standard Brownian motion in* $\mathbb{R}^{\bar{m}}$ *in the same coordinates.*

Proof Sketch. Flatten Φ into its $\bar{m}$ upper-triangular entries, as in Section 5.3. The increment expansion in Theorem 5.9, together with the Gaussian conditional-moment bound at the end of its proof, gives the required drift, covariance, and third-moment estimates uniformly on compact subsets of $\mathrm{SPD}(m)$. These conditional estimates depend on the history only through Φ_ℓ, so they are also the Markov-state coefficients required for a diffusion approximation. They match the drift and diffusion covariance in (5.72), hence yield convergence up to exit from regular localization domains lying strictly inside $\mathrm{SPD}(m)$; the precise sufficient result is deferred to Theorem A.4. The additional boundary localization required for the norm-stopped formulation above is not repeated here; compare [34, Appendices A, C, and D]. □

An independent readout converts the terminal Gram matrix into the output covariance. Let $W_d \in \mathbb{R}^{1\times n}$ be an independent readout weight row vector with i.i.d. $\mathcal{N}(0,1)$ entries (independent of the weights used to form h_d^α). Define the (scalar) readout for input α at depth d by

$$f^\alpha := \sqrt{\frac{c}{n}}\, W_d\, \varphi_s(h_d^\alpha), \qquad \alpha = 1, \dots, m. \tag{5.73}$$

Conditional on $\mathcal{F}_d$, the vector $f := (f^1, \dots, f^m) \in \mathbb{R}^m$ is jointly Gaussian with mean zero and covariance

$$\mathrm{Cov}\Big(f^\alpha, f^\beta \mid \mathcal{F}_d\Big) = \frac{c}{n}\left\langle \varphi_s(h_d^\alpha), \varphi_s(h_d^\beta)\right\rangle = \Phi_d^{\alpha\beta}. \tag{5.74}$$

Since this conditional law depends on $\mathcal{F}_d$ only through Φ_d, it follows that

$$f \mid \Phi_d \sim \mathcal{N}(0, \Phi_d). \tag{5.75}$$

If $\Phi_d \xrightarrow{d} \Phi_{\bar{\tau}}$ and $\Phi_{\bar{\tau}}$ is finite, the corresponding limiting readout is therefore a Gaussian mixture with random covariance $\Phi_{\bar{\tau}}$ [34, Theorems 3.2 and 3.9].

We close the section by examining stability in the one-input case. When $m = 1$ (one input / one covariance coordinate), the smooth-shaped drift (5.56) reduces to a logistic-type scalar SDE:

$$\mathrm{d}\Phi_\tau = b\,\Phi_\tau(\Phi_\tau - 1)\,\mathrm{d}\tau + \sqrt{2}\,\Phi_\tau\,\mathrm{d}B_\tau, \qquad b := \frac{\frac{3}{4}\varphi''(0)^2 + \varphi'''(0)}{a^2}, \tag{5.76}$$

where B_τ is a standard one-dimensional Brownian motion.

A useful intuition for the role of b comes from the corresponding ODE $\partial_\tau \Phi_\tau = b\,\Phi_\tau(\Phi_\tau - 1)$, which solves explicitly to

$$\Phi_\tau = \frac{1}{1 - (1 - \Phi_0^{-1})e^{b\tau}}. \tag{5.77}$$

If $b > 0$ and $\Phi_0 > 1$, the ODE explodes at the finite time $\tau_\star = \frac{1}{b}\log\!\big(\frac{\Phi_0}{\Phi_0 - 1}\big)$, whereas for $\Phi_0 \in (0,1)$ it is driven toward 0. If $b < 0$, the drift points toward 1 from both sides, so every positive

ODE trajectory remains bounded; if $b = 0$, every ODE trajectory is constant. For the SDE, the Brownian term at $\Phi_\tau = 1$ is $\sqrt{2}\,\mathrm{d}B_\tau$, so a trajectory that reaches 1 continues to fluctuate rather than staying there. Nevertheless, the same drift mechanism explains why explosion is possible: when $b > 0$, an upward fluctuation above 1 is reinforced by a positive drift that grows quadratically, whereas for $b \le 0$ the drift does not push large values farther outward. The following result makes this intuition precise.

Proposition 5.11 (Finite-Time Explosion Criterion for the Scalar Covariance SDE [34, Proposition 3.7])**.** *Let Φ_τ solve* (5.76) *with $\Phi_0 > 0$ and define the explosion time*

$$\tau^\star := \sup_{M>1} \inf\Big\{\tau \ge 0 :\ \Phi_\tau \ge M \ \text{or}\ \Phi_\tau \le M^{-1}\Big\}. \tag{5.78}$$

Then $\mathbb{P}(\tau^\star = \infty) = 1$ if and only if $b \le 0$.

Thus $b \le 0$ is exactly the almost-sure nonexplosion condition; when $b > 0$, finite-time explosion occurs with positive probability, not necessarily almost surely [34, Section 4].

This is one reason the smooth-shaped covariance SDE limit is naturally formulated in a *local* (stopped) sense: when explosions are possible, one compares the prelimit chain and the limiting diffusion only up to the first exit time from a fixed compact subset of $(0, \infty)$ (or of $\mathrm{SPD}(m)$ in the matrix case) [34, Definition 3.8 and Theorem 3.9].

5.5 Exercises

Exercise 5.12 (Local Truncation Error)**.** Assume (5.4). Show that the one-step error of Euler satisfies

$$\|X_\eta - (X_0 + \eta b(X_0))\| = O(\eta^2). \tag{5.79}$$

Exercise 5.13 (Wick Expansion)**.** Let $g \sim \mathcal{N}(0, \Phi)$ in $\mathbb{R}^m$. Use Isserlis' theorem to compute $\mathbb{E}[g_\alpha g_\beta g_\gamma g_\delta]$ and derive Proposition 5.5.

Exercise 5.14 (Proof of Proposition 5.7)**.** Adapt the one-sample argument of [34, §2.2]. In particular, show that $\Phi_{\ell+1} = \Phi_\ell(1 + \frac{1}{\sqrt{n}} R_\ell)$ for $0 \le \ell \le d-1$, where $\mathbb{E}[R_\ell \mid \mathcal{F}_\ell] = 0$ and $\mathrm{Var}(R_\ell \mid \mathcal{F}_\ell) = 5$. After checking the third-moment remainder, match the zero drift and diffusion coefficient $\sqrt{5}\,\Phi$ in Proposition 5.7; the precise diffusion-approximation result is Theorem A.4.

Chapter 6

Further Topics in Proportional Depth–Width Scaling

This chapter develops three further aspects of the proportional "depth–width" scaling regime. The common mechanism is the accumulation of small layerwise effects when depth and width grow together. Depending on the setting, this produces a covariance SDE at initialization, nontrivial spectral dynamics, or interactions among layer updates after one training step. We study:

(i) how shaping prevents rank collapse in pure attention and yields a covariance SDE in the proportional limit [42];
(ii) what the linear covariance SDE implies for eigenvalues and empirical spectral distributions [32];
(iii) how one gradient step changes features and predictions in a deep linear network [33].

The first two parts concern forward propagation at initialization. The final part turns to training and isolates the layer-interaction term absent from kernel linearization. Parts (ii) and (iii) draw on a recent preprint and a manuscript in preparation, respectively; here we prioritize clean statements and plausibility calculations.

6.1 Shaped Transformers

A Transformer is an architecture that roughly treats a single token (word in a sentence) as a vector similar to a single data point in an MLP, but considers a sequence of tokens as inputs together (like a full sentence or paragraph), and allows for mixing via a self-attention mechanism.

Fix input token vectors $x^1, \dots, x^m \in \mathbb{R}^{n_0}$ and write $h_0 = [\, x^\beta \,]_{\beta=1}^m \in \mathbb{R}^{n_0 \times m}$. The input layer uses an embedding matrix $W_0 \in \mathbb{R}^{n \times n_0}$ and produces the first hidden-state matrix

$$h_1 := \frac{1}{\sqrt{n_0}} W_0 h_0, \qquad h_\ell = \big[\, h_\ell^\beta \,\big]_{\beta=1}^m \in \mathbb{R}^{n \times m}, \quad h_\ell^\beta \in \mathbb{R}^n, \quad \ell \geq 1, \tag{6.1}$$

where the columns correspond to tokens. At layer $\ell \geq 1$, the (single-head) scaled dot-product attention mechanism, with fixed query/key width n_k, first forms the query and key matrices

$$q_\ell = \frac{1}{\sqrt{n}} W_\ell^q h_\ell \in \mathbb{R}^{n_k \times m}, \qquad k_\ell = \frac{1}{\sqrt{n}} W_\ell^k h_\ell \in \mathbb{R}^{n_k \times m}, \tag{6.2}$$

and then the score and attention matrices

$$S_\ell := k_\ell^\top q_\ell \in \mathbb{R}^{m\times m}, \qquad A_\ell := \mathrm{Softmax}\Big(s^{-1}S_\ell - M\Big) \in \mathbb{R}^{m\times m}, \tag{6.3}$$

where M is the mask and $s > 0$ is the temperature. Increasing s shrinks the logits and makes each attention column less concentrated. Standard scaled dot-product attention takes $s \asymp \sqrt{n_k}$, and we will discuss other choices later. Here the Softmax is column-wise, i.e. for $S \in \mathbb{R}^{m\times m}$,

$$\mathrm{Softmax}(S)^{\alpha\beta} \;:=\; \frac{\exp(S^{\alpha\beta})}{\sum_{\gamma=1}^m \exp(S^{\gamma\beta})}. \tag{6.4}$$

Thus the first index α denotes the key token, the second index β denotes the query token, and every column of A_ℓ sums to 1. The corresponding self-attention network recursion is

$$h_{\ell+1} = \frac{1}{\sqrt{n}} W_\ell^v h_\ell \, A_\ell. \tag{6.5}$$

Thus $h_\ell A_\ell$ performs a *weighted average* across the token index. Because tokens are stored as columns, the βth output column is the attention-weighted sum of the value vectors, and no transpose is needed in the recursion. Here $W_\ell^q, W_\ell^k \in \mathbb{R}^{n_k\times n}$ and $W_\ell^v \in \mathbb{R}^{n\times n}$. At initialization, all entries of $W_0, W_\ell^q, W_\ell^k, W_\ell^v$, across weight types and layers, are mutually independent $\mathcal{N}(0,1)$ random variables.

Normally M is the causal mask, with $M^{\alpha\beta} = +\infty$ for $\alpha > \beta$ and 0 otherwise. For simplicity, throughout this section we set $M = 0$ and study unmasked attention.

However, naively increasing the depth of this attention-only recursion leads to a degeneracy known as **rank collapse**. To see the scale of one layer, suppress the layer index and note the exact identity

$$\frac{1}{s}k^\top q = \frac{1}{s}\Big[\, \Big\langle k^\alpha, q^\beta \Big\rangle \,\Big]_{\alpha,\beta=1}^m. \tag{6.6}$$

Under the Gaussian initialization specified above, for fixed α and β the summands $k_a^\alpha q_a^\beta$ are iid and centered across $a \in [n_k]$, conditional on h. Thus $\Big\langle k^\alpha, q^\beta \Big\rangle$ has typical size $\sqrt{n_k}$ when the normalized token norms are of order one. With the *standard* choice $s \asymp \sqrt{n_k}$, the logits are $O(1)$ and

$$A = \mathrm{Softmax}\left(\frac{1}{s}k^\top q\right) \tag{6.7}$$

does not approach the identity as width grows.

Each layer therefore performs non-negligible *token mixing*. In a deep pure self-attention network without shortcuts or MLPs, repeated mixing can drive the token representations toward a token-uniform rank-one configuration; under the conditions of [12, Theorem 2.3], this convergence is doubly exponential in depth. Related signal-propagation analysis connects rank collapse at initialization to vanishing query and key gradients [41, Theorem 3.2].

This is similar to the ReLU correlation degeneration in (5.47). There the scalar correlation map drives every off-diagonal correlation toward 1; here repeated token mixing aligns the token representations. The mechanisms are different, but in both cases normalized representations become progressively less distinguishable with depth.

This motivates an attention analogue of the activation shaping in (5.50) and (5.51): make the token mixing in each layer small enough to accumulate nontrivially over proportional depth.

6.1.1 Shaped Attention and the Covariance SDE Limit

At the shaped temperature, ordinary Softmax approaches the uniform matrix, which still averages the tokens. The shaped Transformer instead uses the same broad strategy as the smooth activation shaping in (5.51): subtract the leading uniform term and add the identity, leaving a width-dependent perturbation of the identity whose small one-step effects accumulate over proportional depth. For attention, the construction in [42, Section 4.2] combines three modifications:

(a) temperature scaling: take s larger than the standard $\sqrt{n_k}$, namely $s = s_0\sqrt{n\, n_k}$;
(b) uniform centering: subtract the uniform Softmax baseline;
(c) identity shift: add the identity so that attention is close to I_m.

Definition 6.1 (Shaped Attention)**.** In the unmasked column-token convention, the shaped attention map is

$$A_\ell^s \;:=\; I_m + \mathrm{Softmax}(s^{-1}S_\ell) - \frac{1}{m}\mathbf{1}\mathbf{1}^\top. \tag{6.8}$$

Here $s = s_0\sqrt{n\, n_k}$ for fixed $s_0 > 0$. The subtraction removes the leading uniform term, while the identity shift prevents mixing between token positions at leading order. Each column of A_ℓ^s still sums to 1, although the uniform centering means that some off-diagonal entries may be negative.

Definition 6.2 (Feature Covariance)**.** The input and hidden token feature covariances are

$$\Phi_0 := \frac{1}{n_0} h_0^\top h_0, \qquad \Phi_\ell := \frac{1}{n} h_\ell^\top h_\ell \in \mathbb{R}^{m\times m}, \quad \ell \ge 1. \tag{6.9}$$

Thus $\Phi_\ell^{\alpha\beta} = n^{-1}\left\langle h_\ell^\alpha, h_\ell^\beta \right\rangle$ for $\ell \ge 1$. The rows of h_1 are iid $\mathcal{N}(0, \Phi_0)$, so the law of large numbers gives $\Phi_1 \to \Phi_0$ entrywise in probability as $n \to \infty$. The hidden token covariance is the main object tracked in the proportional limit.

Lemma 6.3 (A Softmax Expansion Around the Uniform Point)**.** *Let $S \in \mathbb{R}^{m\times m}$ and consider* $\mathrm{Softmax}(s^{-1}S)$ *for large temperature s. For each* column *index $\beta \in [m]$, define the column mean and mean square*

$$\overline{S^\beta} := \frac{1}{m}\sum_{\gamma=1}^m S^{\gamma\beta}, \qquad \overline{(S^\beta)^2} := \frac{1}{m}\sum_{\gamma=1}^m (S^{\gamma\beta})^2. \tag{6.10}$$

Then, entrywise for $\alpha, \beta \in [m]$,

$$\mathrm{Softmax}(s^{-1}S)^{\alpha\beta} = \frac{1}{m} + \frac{1}{ms}(S^{\alpha\beta} - \overline{S^\beta}) + \frac{1}{2ms^2}\left[(S^{\alpha\beta} - \overline{S^\beta})^2 - \left(\overline{(S^\beta)^2} - (\overline{S^\beta})^2\right)\right] + O(s^{-3}), \tag{6.11}$$

where the $O(s^{-3})$ remainder is uniform on bounded sets of S.

Proof. Fix a column index β and set $r_\alpha := S^{\alpha\beta}$. Also write $\bar r := m^{-1}\sum_{\nu=1}^m r_\nu$ and $\overline{r^2} := m^{-1}\sum_{\nu=1}^m r_\nu^2$. Uniformly on bounded sets of S, the following two Taylor expansions hold, and the normalized denominator in the second line is bounded away from zero for all sufficiently large s. Applying $(1+x)^{-1} = 1 - x + x^2 + O(x^3)$ then gives the final line:

$$\begin{aligned} e^{r_\alpha/s} &= 1 + \frac{r_\alpha}{s} + \frac{r_\alpha^2}{2s^2} + O(s^{-3}), \\ \frac{1}{m}\sum_{\nu=1}^m e^{r_\nu/s} &= 1 + \frac{\bar r}{s} + \frac{\overline{r^2}}{2s^2} + O(s^{-3}), \\ \mathrm{Softmax}(s^{-1}S)^{\alpha\beta} &= \frac{1}{m}\left[1 + \frac{r_\alpha - \bar r}{s} + \frac{(r_\alpha - \bar r)^2 - (\overline{r^2} - \bar r^2)}{2s^2}\right] + O(s^{-3}), \end{aligned} \tag{6.12}$$

which is the claimed expansion after substituting the definitions above. □

With $s = s_0\sqrt{n\,n_k}$ and raw scores $S = k^\top q$ of typical size $\Theta(\sqrt{n_k})$, the entries of $s^{-1}S$ are of order $\Theta(n^{-1/2})$. Equivalently, Lemma 6.3 is an expansion in the small parameter $n^{-1/2}$.

More precisely, combining (6.8) with Lemma 6.3 yields the entrywise expansion

$$(A_\ell^s)^{\alpha\beta} = (I_m)^{\alpha\beta} + \underbrace{\frac{S_\ell^{\alpha\beta} - \overline{S_\ell^\beta}}{ms_0\sqrt{n\,n_k}}}_{\text{centered fluctuation}} + \underbrace{\frac{1}{2ms_0^2 n n_k}\left[(S_\ell^{\alpha\beta} - \overline{S_\ell^\beta})^2 - \left(\overline{(S_\ell^\beta)^2} - (\overline{S_\ell^\beta})^2\right)\right]}_{\text{drift-scale correction}} + O_{\mathbb{P}}(n^{-3/2}). \tag{6.13}$$

Conditional on the previous layers, shaped attention is the identity plus a centered fluctuation of order $n^{-1/2}$ and a smaller, generally random correction of order n^{-1}. When this is inserted into the one-step covariance update $\Phi_{\ell+1} = \frac{1}{n}h_{\ell+1}^\top h_{\ell+1}$, the first-order fluctuation contributes to the diffusion-scale term, while the conditional mean of the smaller correction, together with quadratic products of first-order fluctuations, contributes to the drift. The one-step drift and covariance therefore match the coefficients of the candidate SDE, by the same mechanism used in Theorem 5.10; the precise sufficient Markov-chain result is deferred to Theorem A.4.

6.1.2 A Representative SDE Limit Theorem for Shaped Attention

The next theorem is an informal column-token restatement of the no-shortcut case of the unmasked, attention-only covariance limit in [42, Theorem 4.2], rewritten in our Φ-notation.

Theorem 6.4 (Neural Covariance SDE for Shaped Attention (Informal))**.** *Consider a shaped-attention network at initialization in Definition 6.1. Let $d, n \to \infty$ with $d/n \to \bar{\tau} \in (0, \infty)$ and m fixed. The interpolated process $\Phi_\tau^{(n)} := \Phi_{\lfloor n\tau \rfloor}$, $0 \le \tau \le \bar{\tau}$, converges in distribution in the Skorohod topology to the solution of the SDE*

$$\mathrm{d}\Phi_\tau = b(\Phi_\tau)\,\mathrm{d}\tau + \Sigma(\Phi_\tau)^{1/2}\,\mathrm{d}B_\tau, \qquad \Phi_0 = \frac{1}{n_0}h_0^\top h_0, \tag{6.14}$$

where B_τ is a Brownian motion in $\mathbb{R}^{m(m+1)/2}$ (upper-triangular coordinates). The coefficients are as follows. Write $\Phi^{\alpha\bar{x}} := \frac{1}{m}\sum_{\nu=1}^m \Phi^{\alpha\nu}$ and $\bar{\Phi} := \frac{1}{m}\sum_{\nu=1}^m \Phi^{\nu\nu}$. Further define

$$R_1^{\alpha\gamma,\beta\delta}(\Phi) := \Phi^{\alpha\beta}\Big(\Phi^{\gamma\delta} - \Phi^{\gamma\bar{x}} - \Phi^{\delta\bar{x}} + \Phi^{\bar{x}\bar{x}}\Big), \tag{6.15}$$

$$R_2^{\alpha\gamma}(\Phi) := \Phi^{\alpha\alpha}\Big(\Phi^{\gamma\gamma} - 2\Phi^{\gamma\bar{x}} + 2\Phi^{\bar{x}\bar{x}} - \bar{\Phi}\Big), \tag{6.16}$$

where $\Phi^{\bar{x}\bar{x}} := \frac{1}{m^2}\sum_{\nu,\kappa=1}^m \Phi^{\nu\kappa}$. Then the drift can be written as

$$b^{\alpha\beta}(\Phi) = \frac{1}{s_0^2}\left[\frac{1}{m^2}\sum_{\nu,\kappa=1}^m \Phi^{\nu\kappa}\,R_1^{\alpha\nu,\beta\kappa}(\Phi) + \frac{1}{2m}\sum_{\nu=1}^m\Big(\Phi^{\beta\nu}R_2^{\alpha\nu}(\Phi) + \Phi^{\alpha\nu}R_2^{\beta\nu}(\Phi)\Big)\right]. \tag{6.17}$$

Moreover the diffusion has the form

$$\Sigma(\Phi) = \Sigma_{\mathrm{lin}}(\Phi) \;+\; \frac{1}{s_0^2}\,\Sigma_{\mathrm{attn}}(\Phi), \tag{6.18}$$

where $\Sigma_{\mathrm{lin}}^{\alpha\beta,\gamma\delta}(\Phi) = \Phi^{\alpha\gamma}\Phi^{\beta\delta} + \Phi^{\alpha\delta}\Phi^{\beta\gamma}$ *is the linear covariance and*

$$\Sigma_{\mathrm{attn}}^{\alpha\beta,\gamma\delta}(\Phi) := \frac{1}{m^2}\sum_{\nu,\kappa=1}^{m}\begin{pmatrix}\Phi^{\alpha\kappa}\Phi^{\gamma\nu}R_1^{\beta\kappa,\delta\nu}(\Phi) + \Phi^{\alpha\kappa}\Phi^{\delta\nu}R_1^{\beta\kappa,\gamma\nu}(\Phi)\\ + \Phi^{\beta\nu}\Phi^{\gamma\kappa}R_1^{\alpha\nu,\delta\kappa}(\Phi) + \Phi^{\beta\nu}\Phi^{\delta\kappa}R_1^{\alpha\nu,\gamma\kappa}(\Phi)\end{pmatrix}. \tag{6.19}$$

Proof Sketch. For $\ell \geq 1$, condition on the forward filtration $\mathcal{F}_\ell$ from (5.30). Since the layer-ℓ weights are independent of $\mathcal{F}_\ell$, the rows of q_ℓ and k_ℓ are conditionally iid $\mathcal{N}(0, \Phi_\ell)$, and the two row families are conditionally independent. The rows of $n^{-1/2}W_\ell^v h_\ell$ have the same exact conditional law and are conditionally independent of both families. The convergence $\Phi_1 \to \Phi_0$ noted above shows that the single embedding step does not affect the proportional-depth limit.

With $s = s_0\sqrt{n\,n_k}$, the entries of $s^{-1}k_\ell^\top q_\ell$ are of order $n^{-1/2}$. Applying Lemma 6.3 and then the centering and identity shift in (6.8) gives again

$$\begin{aligned}(A_\ell^s)^{\alpha\beta} = (I_m)^{\alpha\beta} &+ \frac{S_\ell^{\alpha\beta} - \overline{S_\ell^\beta}}{ms_0\sqrt{n\,n_k}}\\ &+ \frac{1}{2ms_0^2 n n_k}\left[\left(S_\ell^{\alpha\beta} - \overline{S_\ell^\beta}\right)^2 - \left(\overline{(S_\ell^\beta)^2} - \left(\overline{S_\ell^\beta}\right)^2\right)\right] + O_{\mathbb{P}}(n^{-3/2}).\end{aligned} \tag{6.13}$$

The first correction is a centered fluctuation of order $n^{-1/2}$, while the second is of order n^{-1} and contributes through its conditional mean.

Insert this expansion into the covariance update $\Phi_{\ell+1} = \frac{1}{n}h_{\ell+1}^\top h_{\ell+1}$. For this no-shortcut update, the covariance recursion and its conditional mean over the value weights are exactly

$$\begin{aligned}\Phi_{\ell+1} &= \frac{1}{n^2}(A_\ell^s)^\top h_\ell^\top (W_\ell^v)^\top W_\ell^v h_\ell A_\ell^s,\\ \mathbb{E}\left[\Phi_{\ell+1} \mid \mathcal{F}_\ell, A_\ell^s\right] &= (A_\ell^s)^\top \Phi_\ell A_\ell^s.\end{aligned} \tag{6.20}$$

Indeed, conditioning on $\mathcal{F}_\ell$ and A_ℓ^s leaves W_ℓ^v independent with iid $\mathcal{N}(0,1)$ entries, and $\mathbb{E}[(W_\ell^v)^\top W_\ell^v] = nI_n$. Subtracting Φ_ℓ and expanding the second line of (6.20) around the identity gives

$$\begin{aligned}\mathbb{E}\left[\Phi_{\ell+1} - \Phi_\ell \mid \mathcal{F}_\ell, A_\ell^s\right] = \Phi_\ell\left(A_\ell^s - I_m\right) &+ \left(A_\ell^s - I_m\right)^\top \Phi_\ell\\ &+ \left(A_\ell^s - I_m\right)^\top \Phi_\ell \left(A_\ell^s - I_m\right).\end{aligned} \tag{6.21}$$

After the value-weight average in (6.20), the remaining conditional mean is the fresh query–key average of $(A_\ell^s)^\top \Phi_\ell A_\ell^s$. In the column convention here, the first index of A_ℓ^s is the key and the second is the query, so multiplying the expansion above gives

$$\begin{aligned}\mathbb{E}\left[(A_\ell^s)^{\nu\alpha}(A_\ell^s)^{\kappa\beta} \mid \mathcal{F}_\ell\right] = \delta_{\nu\alpha}\delta_{\kappa\beta} &+ \frac{1}{s_0^2 n}\left[\frac{1}{m^2}R_1^{\alpha\nu,\beta\kappa}(\Phi_\ell)\right.\\ &\left.+ \frac{1}{2m}\left(\delta_{\kappa\beta}R_2^{\alpha\nu}(\Phi_\ell) + \delta_{\nu\alpha}R_2^{\beta\kappa}(\Phi_\ell)\right)\right] + o_{\mathbb{P}}(n^{-1}).\end{aligned} \tag{6.22}$$

The linear means vanish; the linear–linear product is the centered score moment in (6.28) and gives R_1, while the two identity–quadratic products use

$$\frac{1}{n_k}\,\mathbb{E}\left[\left(S_\ell^{\nu\alpha} - \overline{S_\ell^\alpha}\right)^2 - \left(\overline{(S_\ell^\alpha)^2} - \left(\overline{S_\ell^\alpha}\right)^2\right)\,\middle|\,\mathcal{F}_\ell\right] = R_2^{\alpha\nu}(\Phi_\ell). \tag{6.23}$$

Products containing at least three fluctuation-scale factors, together with the Softmax remainder, are $o_{\mathbb{P}}(n^{-1})$; this is the column-index transpose of the pair calculation in [42, Lemmas C.2–C.3]. Thus the tower property and contraction with Φ_ℓ give

$$\begin{aligned}\mathbb{E}\left[\Phi_{\ell+1}^{\alpha\beta} \mid \mathcal{F}_\ell\right] &= \sum_{\nu,\kappa=1}^{m} \Phi_\ell^{\nu\kappa}\mathbb{E}\left[(A_\ell^s)^{\nu\alpha}(A_\ell^s)^{\kappa\beta} \mid \mathcal{F}_\ell\right] \\ &= \Phi_\ell^{\alpha\beta} + \frac{1}{s_0^2 n}\left[\frac{1}{m^2}\sum_{\nu,\kappa=1}^{m}\Phi_\ell^{\nu\kappa}R_1^{\alpha\nu,\beta\kappa}(\Phi_\ell)\right. \\ &\qquad \left. + \frac{1}{2m}\sum_{\nu=1}^{m}\left(\Phi_\ell^{\beta\nu}R_2^{\alpha\nu}(\Phi_\ell) + \Phi_\ell^{\alpha\nu}R_2^{\beta\nu}(\Phi_\ell)\right)\right] + o_{\mathbb{P}}(n^{-1}) \\ &= \Phi_\ell^{\alpha\beta} + \frac{1}{n}b^{\alpha\beta}(\Phi_\ell) + o_{\mathbb{P}}(n^{-1}),\end{aligned} \tag{6.24}$$

where symmetry of Φ_ℓ is used in the R_2 contractions, and the final line is exactly (6.17).

After flattening the upper-triangular entries, the two sources of covariance are separated exactly by the conditional covariance formula:

$$\begin{aligned}\operatorname{Cov}\left(\Phi_{\ell+1} - \Phi_\ell \mid \mathcal{F}_\ell\right) = \mathbb{E}\left[\operatorname{Cov}\left(\Phi_{\ell+1} \mid \mathcal{F}_\ell, A_\ell^s\right) \mid \mathcal{F}_\ell\right] \\ + \operatorname{Cov}\left((A_\ell^s)^\top \Phi_\ell A_\ell^s \mid \mathcal{F}_\ell\right).\end{aligned} \tag{6.25}$$

The first term is the Gaussian sample-covariance fluctuation from the value weights, and the second is the fluctuation of the conditional mean caused by shaped attention. The value-weight contribution can be computed exactly. Conditional on $\mathcal{F}_\ell$ and A_ℓ^s, the n output rows are iid centered Gaussian with covariance $(A_\ell^s)^\top \Phi_\ell A_\ell^s$. Their empirical covariance is $\Phi_{\ell+1}$, so Isserlis' identity gives

$$\operatorname{Cov}\left(\Phi_{\ell+1}^{\alpha\beta}, \Phi_{\ell+1}^{\gamma\delta} \mid \mathcal{F}_\ell, A_\ell^s\right) = \frac{1}{n}\,\Sigma_{\text{lin}}^{\alpha\beta,\gamma\delta}\left((A_\ell^s)^\top \Phi_\ell A_\ell^s\right). \tag{6.26}$$

Since $A_\ell^s = I_m + O_{\mathbb{P}}(n^{-1/2})$, averaging this identity over shaped attention gives the leading scaled coefficient $\Sigma_{\text{lin}}(\Phi_\ell)$.

For the attention branch, the score entries are conditionally centered, and independence of the key and query row families gives

$$\frac{1}{n_k}\,\mathbb{E}\left[S_\ell^{\alpha\beta}S_\ell^{\gamma\delta} \mid \mathcal{F}_\ell\right] = \Phi_\ell^{\alpha\gamma}\Phi_\ell^{\beta\delta}, \tag{6.27}$$

$$\frac{1}{n_k}\,\mathbb{E}\left[\left(S_\ell^{\kappa\beta} - \overline{S_\ell^{\beta}}\right)\left(S_\ell^{\nu\delta} - \overline{S_\ell^{\delta}}\right) \mid \mathcal{F}_\ell\right] = R_1^{\beta\kappa,\delta\nu}(\Phi_\ell). \tag{6.28}$$

Indeed, when the scores are expanded as sums over the n_k rows, only equal row indices survive in the first line; the second follows by expanding the two column means over the key index. This also explains the order $R_1^{\beta\kappa,\delta\nu}$: the outside pair records the query covariance, while the centered pair records the key covariance.

At order n^{-1} in the conditional covariance, only the two linear summands $\Phi_\ell(A_\ell^s - I_m)$ and $(A_\ell^s - I_m)^\top\Phi_\ell$ contribute to the attention fluctuation. By (6.13), the leading (κ,β) entry of $A_\ell^s - I_m$ is $\left(S_\ell^{\kappa\beta} - \overline{S_\ell^{\beta}}\right)/(ms_0\sqrt{nn_k})$. Substituting this term and using the conditional score

moments, the covariance of the first linear summand with itself is

$$n\ \mathrm{Cov}\left([\Phi_\ell(A_\ell^s - I_m)]^{\alpha\beta}, [\Phi_\ell(A_\ell^s - I_m)]^{\gamma\delta} \mid \mathcal{F}_\ell\right) = \frac{1}{m^2 s_0^2} \sum_{\kappa,\nu=1}^{m} \Phi_\ell^{\alpha\kappa} \Phi_\ell^{\gamma\nu} R_1^{\beta\kappa,\delta\nu}(\Phi_\ell) + o_{\mathbb{P}}(1). \tag{6.29}$$

After removing the common factor s_0^{-2}, this is the first contraction in Σ_{attn}. The other three pairings among the two linear summands give the remaining three contractions. Although the first-order score fluctuation is conditionally centered, the full $A_\ell^s - I_m$ has a conditional mean of order n^{-1}, which is why the display uses conditional covariance. Combining these four pairings with (6.25) and (6.26), and using that Φ_ℓ is $\mathcal{F}_\ell$-measurable, gives

$$\begin{aligned} &n\ \mathrm{Cov}\left(\Phi_{\ell+1}^{\alpha\beta} - \Phi_\ell^{\alpha\beta}, \Phi_{\ell+1}^{\gamma\delta} - \Phi_\ell^{\gamma\delta} \mid \mathcal{F}_\ell\right) \\ &\quad = \Sigma_{\mathrm{lin}}^{\alpha\beta,\gamma\delta}(\Phi_\ell) + \frac{1}{s_0^2}\Sigma_{\mathrm{attn}}^{\alpha\beta,\gamma\delta}(\Phi_\ell) + o_{\mathbb{P}}(1) = \Sigma^{\alpha\beta,\gamma\delta}(\Phi_\ell) + o_{\mathbb{P}}(1). \end{aligned} \tag{6.30}$$

These leading conditional moments are summarized by the Euler–Maruyama form from (5.53):

$$\Phi_{\ell+1} - \Phi_\ell = \frac{1}{n}\, b(\Phi_\ell) + \frac{1}{\sqrt{n}}\, \Sigma(\Phi_\ell)^{1/2} \xi_\ell + \widetilde{o}_\ell(n^{-1}). \tag{6.31}$$

Conditional on $\mathcal{F}_\ell$, the noise vector ξ_ℓ is centered and has identity covariance in the upper-triangular coordinates; the remainder is understood as in (5.52). The preceding one-step estimates are sufficient to apply the Markov-chain diffusion approximation in Theorem A.4, which completes the proof. □

Remark 6.5 (Residual Shortcuts). The theorem isolates the shaped-attention mechanism through an attention-only recursion. Noci et al. also allow a weighted identity shortcut around the shaped-attention branch, with complementary branch strengths chosen to preserve variance; their theorem shows how these strengths rescale the limiting drift and covariance [42, Theorem 4.2]. We suppress that extension here.

Remark 6.6 (Full Transformer Blocks). A full Transformer block composes shaped attention with a feedforward sublayer. In the same no-shortcut specialization as Theorem 6.4, its column-token recursion can be written as

$$\begin{aligned} \widetilde{h}_\ell &= \frac{1}{\sqrt{n}} W_\ell^v h_\ell A_\ell^s, \\ h_{\ell+1} &= \sqrt{\frac{c}{n}}\, W_\ell^2 \varphi_s\left(\frac{1}{\sqrt{n}} W_\ell^1 \widetilde{h}_\ell\right), \\ \varphi_s(x) &:= s_+ \max\{x, 0\} + s_- \min\{x, 0\}, \qquad s_\pm = 1 + \frac{c_\pm}{\sqrt{n}}. \end{aligned} \tag{6.32}$$

Here $\widetilde{h}_\ell$ is the intermediate state after shaped attention, while h_ℓ and $h_{\ell+1}$ are the states at consecutive full-block boundaries. Thus $\Phi_\ell = n^{-1} h_\ell^\top h_\ell$ records the covariance at those boundaries. The activation is applied entrywise, and the width-dependent normalization $c = c(n)$ is chosen so that $c^{-1} = \mathbb{E}[\varphi_s(g)^2]$ for $g \sim \mathcal{N}(0, 1)$. Thus φ_s is precisely the critically shaped ReLU from (5.50). Here the subscript s refers to the slopes $s_\pm$, not to the attention temperature used in

A_ℓ^s. The matrices $W_\ell^1, W_\ell^2 \in \mathbb{R}^{n\times n}$ have iid $\mathcal{N}(0,1)$ entries; across the two weight types and all layers, they are mutually independent and independent of all query, key, and value weights.

Let b_{ReLU} denote the shaped-ReLU drift in (5.55). Then [42, Theorem 3.2 and Corollary 4.3] gives the corresponding combined limit

$$\mathrm{d}\Phi_\tau = \big[b(\Phi_\tau) + b_{\mathrm{ReLU}}(\Phi_\tau)\big]\,\mathrm{d}\tau + \big[\Sigma(\Phi_\tau) + 2\Sigma_{\mathrm{lin}}(\Phi_\tau)\big]^{1/2}\,\mathrm{d}B_\tau. \tag{6.33}$$

Here b and Σ are the shaped-attention coefficients in Theorem 6.4, while the feedforward sublayer contributes b_{ReLU} and $2\Sigma_{\mathrm{lin}}$ because it contains two fresh Gaussian matrices. Conditional on the completed attention sublayer, and hence on $\widetilde{h}_\ell$, the fluctuation generated by the fresh feedforward weights is centered. The tower property therefore removes its cross-covariance with the preceding attention fluctuation at the limiting covariance scale. The instantaneous covariance matrices add inside the square root; the two square-root diffusion coefficients are not added directly.

6.2 Spectrum Results for the Covariance SDE

We now return to the deep linear network studied in Chapter 5 and ask what its covariance SDE implies for the spectrum. The affine invariance of the linear model leads to explicit eigenvalue dynamics and a tractable mean-field spectral limit. The results are adapted from [32].

6.2.1 Recall: The Linear Covariance SDE

Stack the inputs $x^1, \dots, x^m \in \mathbb{R}^{n_0}$ as the columns of h_0. Let $W_0 \in \mathbb{R}^{n\times n_0}$ and $W_\ell \in \mathbb{R}^{n\times n}$ for $\ell \geq 1$ have i.i.d. $\mathcal{N}(0,1)$ entries. As in Chapter 5, set

$$\begin{aligned} h_1 &= \frac{1}{\sqrt{n_0}} W_0 h_0, \quad h_{\ell+1} = \frac{1}{\sqrt{n}} W_\ell h_\ell, \quad \ell \geq 1, \\ \Phi_0 &= \frac{1}{n_0} h_0^\top h_0, \qquad\quad \Phi_\ell = \frac{1}{n} h_\ell^\top h_\ell, \quad \ell \geq 1. \end{aligned} \tag{6.34}$$

Fix m, let $\bar{m} := m(m+1)/2$, and flatten Φ into its distinct upper-triangular entries, indexed by $\alpha \leq \beta$. As $n \to \infty$, the continuous-time interpolation $\Phi_\tau^{(n)} := \Phi_{\lfloor n\tau \rfloor}$ converges on every bounded time interval to the *linear covariance SDE*

$$\mathrm{d}\Phi_\tau = \Sigma(\Phi_\tau)^{1/2}\,\mathrm{d}B_\tau, \qquad \Sigma^{\alpha\beta,\gamma\delta}(\Phi) = \Phi^{\alpha\gamma}\Phi^{\beta\delta} + \Phi^{\alpha\delta}\Phi^{\beta\gamma}, \quad \alpha \leq \beta,\ \gamma \leq \delta. \tag{6.35}$$

Here B is standard Brownian motion in $\mathbb{R}^{\bar{m}}$; at terminal depths with $d/n \to \bar{\tau}$, the covariance Φ_d therefore converges in distribution to $\Phi_{\bar{\tau}}$. This is precisely the linear covariance limit in Theorem 5.6. In particular, Proposition 5.5 and (5.35) identify the diffusion tensor Σ, while (5.40) gives the SDE in the notation of Chapter 5.

6.2.2 Affine Invariance

A key structural feature of (6.35) is its invariance under congruence transformations $\Phi \mapsto A\Phi A^\top$.

Lemma 6.7 (Affine Invariance of the Covariance Markov Chain)**.** *Let P denote the one-step covariance update $\Phi_\ell \mapsto \Phi_{\ell+1}$ induced by the linear recursion. Then for every $A \in \mathrm{GL}(m,\mathbb{R})$,*

$$AP(\Phi)A^\top \overset{d}{=} P(A\Phi A^\top). \tag{6.36}$$

Consequently, the stochastic flow of (6.35) is affine-invariant: if Φ_τ solves (6.35) with initial condition Φ_0, then $A\Phi_\tau A^\top$ has the same law as the solution started from $A\Phi_0 A^\top$.

Proof. The argument is easiest at the Markov-chain level. Condition on Φ and sample n iid Gaussian vectors $(g_i)_{i=1}^n \subset \mathbb{R}^m$ with $g_i \sim \mathcal{N}(0, \Phi)$. Then the update can be written in distribution as

$$P(\Phi) = \frac{1}{n}\sum_{i=1}^{n} g_i g_i^\top. \tag{6.37}$$

Equivalently, stacking g_i as rows gives $G \in \mathbb{R}^{n\times m}$ with $\mathrm{vec}(G) \sim \mathcal{N}(0, \Phi \otimes I_n)$ and $P(\Phi) = \frac{1}{n}G^\top G$. Now replace Φ by $A\Phi A^\top$. Since $(A\Phi A^\top) \otimes I_n = (A \otimes I_n)(\Phi \otimes I_n)(A^\top \otimes I_n)$, we can couple the Gaussian matrices so that

$$G' = GA^\top \quad \Longrightarrow \quad \mathrm{vec}(G') \sim \mathcal{N}(0, (A\Phi A^\top) \otimes I_n). \tag{6.38}$$

Then

$$P(A\Phi A^\top) \overset{d}{=} \frac{1}{n}(G')^\top G' = \frac{1}{n}AG^\top GA^\top = AP(\Phi)A^\top, \tag{6.39}$$

which is the desired invariance. The invariance of the limiting SDE follows by taking the proportional limit. □

The set $\mathrm{SPD}(m)$ of $m \times m$ symmetric positive-definite matrices is a smooth manifold with tangent space $T_\Phi \mathrm{SPD}(m) \cong \mathrm{Sym}(m)$. This geometry is not needed to use the covariance SDE, but it explains the special structure of its noise.

Proposition 6.8 (Σ^{-1}: Affine-Invariant Metric)**.** *For $\Phi \in \mathrm{SPD}(m)$ and $A, B \in T_\Phi \mathrm{SPD}(m) \cong \mathrm{Sym}(m)$, the affine-invariant Riemannian metric is*

$$g_\Phi(A, B) := \frac{1}{2}\,\mathrm{Tr}(\Phi^{-1} A\, \Phi^{-1} B). \tag{6.40}$$

It is invariant under congruences: for every $P \in \mathrm{GL}(m, \mathbb{R})$,

$$g_{P\Phi P^\top}(PAP^\top, PBP^\top) = g_\Phi(A, B). \tag{6.41}$$

In the distinct upper-triangular coordinates on $\mathrm{Sym}(m)$ used in Chapter 5, the matrix representing this metric is $\Sigma(\Phi)^{-1}$. A proof is given in Appendix A.4; see also [32, Proposition 2.5].

Thus Σ is the cometric associated with (6.40); its multiplicative structure leads to the Dyson-type eigenvalue dynamics studied next.

6.2.3 Geometric Dyson Brownian Motion

For the spectral calculation, it is useful to realize the same diffusion in matrix form. Let S_τ be symmetric Brownian motion on $\mathrm{Sym}(m)$, normalized so that

$$\begin{aligned} \mathrm{d}\langle S^{\alpha\beta}, S^{\gamma\delta}\rangle_\tau &= (\delta_{\alpha\gamma}\delta_{\beta\delta} + \delta_{\alpha\delta}\delta_{\beta\gamma})\,\mathrm{d}\tau, \\ \mathrm{d}\Phi_\tau &= \Phi_\tau^{1/2}\,\mathrm{d}S_\tau\,\Phi_\tau^{1/2}. \end{aligned} \tag{6.42}$$

A direct covariance calculation recovers Σ in (6.35), so the two SDEs are equivalent in law.

Let $\lambda_1(\tau) \le \cdots \le \lambda_m(\tau)$ denote the eigenvalues of Φ_τ.

Theorem 6.9 (Geometric Dyson Brownian Motion)**.** *Let Φ_τ solve (6.35) with $\Phi_0 \in \mathrm{SPD}(m)$. Then the spectrum is almost surely simple for every $\tau > 0$, and the ordered eigenvalues satisfy, locally on $(0, \infty)$,*

$$\mathrm{d}\lambda_i = \sqrt{2}\,\lambda_i\,\mathrm{d}B_\tau^{(i)} \;+\; \sum_{j\neq i} \frac{\lambda_i\lambda_j}{\lambda_i - \lambda_j}\,\mathrm{d}\tau, \qquad i = 1, \dots, m, \tag{6.43}$$

where $(B_\tau^{(i)})_{i=1}^m$ are independent standard Brownian motions. If Φ_0 has simple spectrum, the system holds from $\tau = 0$ and the eigenvalues never collide.

Proof. We sketch a standard Itô argument (cf. eigenvalue computations for Wishart-type processes). The calculation is local on the open set of matrices with simple spectrum. At such a matrix, choose an orthonormal eigenbasis; by the orthogonal invariance in Lemma 6.7, we may compute at $\Phi = \mathrm{diag}(\lambda_1, \dots, \lambda_m)$. For a symmetric perturbation $\Psi \in \mathrm{Sym}(m)$, the first and second Hadamard variation formulae give

$$\langle \nabla\lambda_i(\Phi), \Psi\rangle = \Psi_{ii}, \qquad \nabla^2\lambda_i(\Phi)[\Psi, \Psi] = 2\sum_{j\neq i} \frac{\Psi_{ij}^2}{\lambda_i - \lambda_j}. \tag{6.44}$$

In this basis, (6.42) gives $\mathrm{d}\Phi_{ij} = \sqrt{\lambda_i\lambda_j}\,\mathrm{d}S_{ij}$, and hence

$$\mathrm{d}\langle\Phi_{ii}, \Phi_{jj}\rangle_\tau = 2\lambda_i^2\delta_{ij}\,\mathrm{d}\tau, \qquad \mathrm{d}\langle\Phi_{ij}, \Phi_{ij}\rangle_\tau = \lambda_i\lambda_j\,\mathrm{d}\tau \quad (i \neq j). \tag{6.45}$$

Applying Itô's lemma to $\lambda_i(\Phi_\tau)$ and inserting the above derivatives yields

$$\mathrm{d}\lambda_i = \mathrm{d}\Phi_{ii} + \sum_{j\neq i} \frac{\mathrm{d}\langle\Phi_{ij}, \Phi_{ij}\rangle_\tau}{\lambda_i - \lambda_j}. \tag{6.46}$$

The diagonal quadratic variations show that $\mathrm{d}\Phi_{ii} = \sqrt{2}\,\lambda_i\,\mathrm{d}B^{(i)}$ for independent Brownian motions. Substituting the off-diagonal quadratic variations gives (6.43) up to the first collision. The absence of collisions from a simple initial spectrum, and instantaneous simplicity from an arbitrary positive-definite initial spectrum, are proved in [32, Theorem 1.1 and Appendix B]. □

6.2.4 Mean Field Limit, a Burgers PDE, and a Fixed Point Equation

Write $\mu_i(\tau) := \lambda_i(\tau/m)$ for the eigenvalues on the accelerated spectral clock. Then the empirical spectral measure of the *time-changed* covariance $\Phi_{\tau/m}$ is

$$\rho_\tau^{(m)} := \frac{1}{m}\sum_{i=1}^m \delta_{\mu_i(\tau)}. \tag{6.47}$$

The time change $\tau \mapsto \tau/m$ normalizes the drift in (6.43) to $O(1)$ as $m \to \infty$, and simultaneously shrinks the diffusion by $m^{-1/2}$. The limits here are sequential: first the proportional depth–width limit at fixed m gives the covariance SDE, and only then do we apply this time change and send $m \to \infty$. No simultaneous three-parameter limit is asserted.

The T-transform is useful here because the multiplicative eigenvalue interaction becomes a closed equation for a single analytic function. Define the T-transform

$$G_\tau(z) := \int_{\mathbb{R}} \frac{x}{z - x}\,\rho_\tau(\mathrm{d}x), \qquad z \in \mathbb{C} \setminus \mathbb{R}_{\geq 0}. \tag{6.48}$$

(Equivalently, $G_\tau(z) = z\,g_\tau(z) - 1$ where g_τ is the Stieltjes transform.)

Remark 6.10 (T-Transform Inversion). Let $g_\tau(z) := \int_{\mathbb{R}} \frac{1}{z-x}\,\rho_\tau(\mathrm{d}x)$ be the Stieltjes transform. Whenever ρ_τ admits a density (which we denote by the same symbol) at a point $x \in \mathbb{R}$, the standard Stieltjes inversion formula gives

$$\rho_\tau(x) = -\frac{1}{\pi} \lim_{\varepsilon \downarrow 0} \Im\, g_\tau(x + \mathrm{i}\varepsilon). \tag{6.49}$$

Since $g_\tau(z) = \frac{G_\tau(z)+1}{z}$, we can equivalently write this in terms of the T-transform:

$$\rho_\tau(x) = -\frac{1}{\pi} \lim_{\varepsilon \downarrow 0} \Im \left(\frac{G_\tau(x + \mathrm{i}\varepsilon) + 1}{x + \mathrm{i}\varepsilon} \right). \tag{6.50}$$

The path-level limit and Burgers equation below are proved in [32, Theorem 1.2]; the identity-start fixed point is [32, Corollary 1.3].

Theorem 6.11 (Burgers PDE and Fixed Point)**.** *Assume $\rho_0^{(m)} \xrightarrow{\mathrm{w}} \rho_0$ and that the first moments of $\rho_0^{(m)}$ are uniformly bounded. Then $\rho^{(m)}$ converges in probability, uniformly on compact time intervals in the weak topology, to a deterministic continuous path ρ whose T-transform satisfies the inviscid Burgers-type PDE*

$$\partial_\tau G_\tau(z) = -z\, G_\tau(z)\, \partial_z G_\tau(z). \tag{6.51}$$

If moreover $\rho_0 = \delta_1$, then $G_\tau(z)$ is the physical branch of the fixed point equation

$$G_\tau(z) = \frac{1}{z\, e^{-\tau G_\tau(z)} - 1}. \tag{6.52}$$

Remark 6.12 (Free Log-Normal Fixed Point). The identity-start limit ρ_τ turns out to be the mean-one *free log-normal* law, the multiplicative free-convolution analogue of the classical log-normal law. This identification is most transparent through the S-transform, which turns multiplicative free convolution into multiplication [5]. For the compactly supported law ρ_τ [32, Corollary 4.1], define $\psi_\tau(u) := G_\tau(u^{-1})$, and let χ_τ be its local compositional inverse near the origin. The S-transform of ρ_τ is then $S_{\rho_\tau}(w) := (1+w)\chi_\tau(w)/w$. For z near infinity, taking $w = G_\tau(z)$ gives $\chi_\tau(w) = z^{-1}$, while (6.52) rearranges to $z^{-1} = e^{-\tau w} w/(1+w)$. Consequently, $S_{\rho_\tau}(w) = e^{-\tau w}$, the S-transform of the mean-one free log-normal law [32, Corollary 1.3 and Section 4]. See also [44, Section 16.3] for a physics-oriented treatment.

Remark 6.13 (Physical Branch). In the identity-start case, the fixed-point equation may admit several analytic branches. The T-transform in Theorem 6.11 is selected by $G_\tau(z) \sim z^{-1}$ as $z \to \infty$; this choice is called the *physical branch*.

Proof Sketch for Theorem 6.11. We first derive the Burgers equation. Let $f_z(x) := \frac{x}{z-x}$. On each positive-time interval, and after relabeling the standard Brownian motions $\sqrt{m}\, B^{(i)}_{\tau/m}$ as $B^{(i)}_\tau$, the time-changed form of (6.43) is

$$\mathrm{d}\mu_i(\tau) = \sqrt{\frac{2}{m}}\, \mu_i(\tau)\, \mathrm{d}B^{(i)}_\tau + \frac{1}{m} \sum_{j \neq i} \frac{\mu_i(\tau)\mu_j(\tau)}{\mu_i(\tau) - \mu_j(\tau)}\, \mathrm{d}\tau. \tag{6.53}$$

To display the mean-field mechanism, we retain the leading interaction drift; the martingale and Itô terms vanish in the empirical average as $m \to \infty$. Suppressing the common time argument, differentiate $\langle f_z, \rho_\tau^{(m)} \rangle = \frac{1}{m} \sum_i f_z(\mu_i)$:

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \langle f_z, \rho_\tau^{(m)} \rangle \approx \frac{1}{m^2} \sum_{i \neq j} f_z'(\mu_i) \, \frac{\mu_i \mu_j}{\mu_i - \mu_j}. \tag{6.54}$$

Now symmetrize in (i, j):

$$\sum_{i \neq j} f_z'(\mu_i) \, \frac{\mu_i \mu_j}{\mu_i - \mu_j} = \frac{1}{2} \sum_{i \neq j} \mu_i \mu_j \, \frac{f_z'(\mu_i) - f_z'(\mu_j)}{\mu_i - \mu_j}. \tag{6.55}$$

Passing to the limit $m \to \infty$ heuristically replaces $\frac{1}{m^2} \sum_{i \neq j}$ by $\iint$ against $\rho_\tau \otimes \rho_\tau$, giving

$$\partial_\tau G_\tau(z) = \frac{1}{2} \iint xy \, \frac{f_z'(x) - f_z'(y)}{x - y} \, \rho_\tau(\mathrm{d}x) \rho_\tau(\mathrm{d}y). \tag{6.56}$$

A direct calculation gives the symmetric identity

$$xy \, \frac{f_z'(x) - f_z'(y)}{x - y} = -z \left(f_z(x) \, \partial_z f_z(y) + f_z(y) \, \partial_z f_z(x) \right). \tag{6.57}$$

The two terms agree after integration against $\rho_\tau \otimes \rho_\tau$, so the factor $1/2$ gives precisely $\partial_\tau G_\tau(z) = -z \, G_\tau(z) \, \partial_z G_\tau(z)$.

We next derive the fixed-point equation by characteristics. The idea is to follow a curve along which the two terms in the PDE cancel. Starting from z_0, let $z(\tau)$ solve $\partial_\tau z(\tau) = z(\tau) \, G(\tau, z(\tau))$ with $z(0) = z_0$. Then

$$\frac{\mathrm{d}}{\mathrm{d}\tau} G(\tau, z(\tau)) = \partial_\tau G + \partial_\tau z \, \partial_z G = \partial_\tau G + z G \, \partial_z G = 0, \tag{6.58}$$

so the transform is constant along the curve: $G(\tau, z(\tau)) = G_0(z_0)$. The characteristic equation therefore reduces to the linear ODE $\partial_\tau z(\tau) = G_0(z_0) z(\tau)$, whose solution is $z(\tau) = z_0 \exp(\tau G_0(z_0))$. Now fix the terminal time τ and write $z = z(\tau)$. Since $G_\tau(z) = G_0(z_0)$, the last relation becomes $z = z_0 \exp(\tau G_\tau(z))$, or $z_0 = z \, e^{-\tau G_\tau(z)}$. Substituting this starting point into $G_\tau(z) = G_0(z_0)$ gives the general characteristic identity

$$G_\tau(z) = G_0\Big(z \, e^{-\tau G_\tau(z)} \Big). \tag{6.59}$$

For $\rho_0 = \delta_1$, we have $G_0(z) = (z - 1)^{-1}$, and (6.59) becomes (6.52). This complex-characteristic calculation is formal. □

6.2.5 A Small-τ Approximation in Closed Form

Making the formal first-order replacement $e^{-\tau G_\tau(z)} \approx 1 - \tau G_\tau(z)$ in (6.52) leads to the quadratic approximation [32, Section 4.2]

$$z\tau \, G_\tau(z)^2 - (z - 1) G_\tau(z) + 1 = 0. \tag{6.60}$$

For $\tau > 0$, the physical root is selected by $G_\tau(z) \sim z^{-1}$ as $z \to \infty$; equivalently, the square root is chosen to behave like z at infinity. This gives

$$G_\tau(z) \approx \frac{(z-1) - \sqrt{(z-1)^2 - 4z\tau}}{2z\tau}. \tag{6.61}$$

Let $\lambda_\pm := (\sqrt{1+\tau} \pm \sqrt{\tau})^2$. Since $(z-1)^2 - 4z\tau = (z-\lambda_-)(z-\lambda_+)$, substituting the selected root into (6.50) gives, for $x > 0$,

$$\begin{aligned}
\rho_\tau(x) &\approx -\frac{1}{\pi x} \lim_{\varepsilon \downarrow 0} \Im G_\tau(x + \mathrm{i}\varepsilon) \\
&\approx -\frac{1}{\pi x} \left(-\frac{\sqrt{(x-\lambda_-)(\lambda_+ - x)}}{2x\tau} \right) \mathbb{1}_{\{x \in [\lambda_-, \lambda_+]\}} \\
&= \frac{\sqrt{(x-\lambda_-)(\lambda_+ - x)}}{2\pi\, x^2\, \tau}\, \mathbb{1}_{\{x \in [\lambda_-, \lambda_+]\}}.
\end{aligned} \tag{6.62}$$

Within this approximation, the displayed density has total mass one. More precisely, it is the x^{-1}-reweighting of the Marchenko–Pastur law with variance parameter $1 + \tau$ and aspect ratio $\tau/(1+\tau)$.

6.3 One Step of Feature Learning in the Proportional Limit

The preceding two sections study covariance dynamics at initialization. We now turn to training and ask how a single gradient step changes features when $d/n \to \bar{\tau}$. A one-sample deep linear network is the simplest setting in which simultaneous updates across proportionally many layers interact nontrivially. The broader multi-sample nonlinear theory is the subject of ongoing work [33]. Since that manuscript is still in preparation, the finite-width identities below are derived directly, while the continuum equations serve as a formal guide. Even in this scalar setting, the proportional limit produces nontrivial stochastic dynamics *in depth*.

We begin with a fixed input–output pair (x, y), with nonzero $x \in \mathbb{R}^{n_0}$ and $y \in \mathbb{R}$. We write $h_0 = x$ for the input representation and use the same input boundary as in (5.14):

$$h_0 := x, \qquad h_1 := \frac{1}{\sqrt{n_0}} W_0 x, \qquad h_{\ell+1} := \frac{1}{\sqrt{n}} W_\ell h_\ell, \quad 1 \le \ell \le d-1. \tag{6.63}$$

The scalar readout is

$$f := \frac{1}{\sqrt{n}} W_d h_d. \tag{6.64}$$

Here $W_0 \in \mathbb{R}^{n \times n_0}$, $W_\ell \in \mathbb{R}^{n \times n}$ for $1 \le \ell \le d-1$, and $W_d \in \mathbb{R}^{1 \times n}$ have independent standard Gaussian entries at initialization. To focus on the bulk depth mechanism, we hold W_0 and W_d fixed and train only $W_1, \dots, W_{d-1}$.

For the backward pass, define the scaled output sensitivity

$$g_\ell := \sqrt{n}\, \frac{\partial f}{\partial h_\ell} \in \mathbb{R}^n, \qquad 1 \le \ell \le d. \tag{6.65}$$

Then $g_d = W_d^\top$, and the chain rule gives

$$g_\ell = \frac{1}{\sqrt{n}} W_\ell^\top g_{\ell+1}, \qquad \ell = d-1, \dots, 1. \tag{6.66}$$

6.3.1 Scalar Kernels and Depth-Time SDEs at Initialization

All Gram objects reduce to scalars. Define

$$\Phi_0 := \frac{1}{n_0}\|x\|^2, \qquad \Phi_\ell := \frac{1}{n}\|h_\ell\|^2, \qquad G_\ell := \frac{1}{n}\|g_\ell\|^2, \qquad Q_\ell := \frac{1}{\sqrt{n}}\langle h_\ell, g_\ell\rangle. \tag{6.67}$$

The last three quantities are defined for $1 \le \ell \le d$.

Set the depth-time scaling $\tau = \ell/n$ and assume $d/n \to \bar{\tau} \in (0,\infty)$. We use $\mathcal{F}^h$ for the forward pass, $\mathcal{F}^g$ for the backward pass, and $\mathcal{F}^{\Delta h}$ below for the feature change. In the one-sample linear setting, the forward kernel converges to a geometric Brownian motion in the proportional limit.

Remark 6.14 (Recall: Geometric Brownian Motion for the Forward Kernel Φ). In the deep linear case, let $\mathcal{F}_\ell^h := \sigma(h_0,\dots,h_\ell)$ be the forward filtration. For $0 \le \ell \le d-1$, conditioning on $\mathcal{F}_\ell^h$ gives

$$\Phi_{\ell+1} \mid \mathcal{F}_\ell^h \overset{d}{=} \Phi_\ell \frac{\chi_n^2}{n}, \tag{6.68}$$

where $\chi_n^2 \sim \chi^2(n)$ is independent of $\mathcal{F}_\ell^h$. As recalled in (5.23), when $d,n \to \infty$ with $d/n \to \bar{\tau}$ the interpolated process $\Phi_\tau^{(n)} := \Phi_{\lfloor n\tau \rfloor}$ converges to the SDE

$$\mathrm{d}\Phi_\tau = \sqrt{2}\,\Phi_\tau\,\mathrm{d}B_\tau^\Phi, \qquad \Phi_0 = \frac{1}{n_0}\|x\|^2. \tag{6.69}$$

The linear forward and backward recursions also imply that Q is exactly constant in depth. Indeed, using $h_{\ell+1} = \frac{1}{\sqrt{n}} W_\ell h_\ell$ and $g_\ell = \frac{1}{\sqrt{n}} W_\ell^\top g_{\ell+1}$,

$$Q_{\ell+1} = \frac{1}{\sqrt{n}}\langle h_{\ell+1}, g_{\ell+1}\rangle = \frac{1}{n}\langle W_\ell h_\ell, g_{\ell+1}\rangle = \frac{1}{n}\left\langle h_\ell, W_\ell^\top g_{\ell+1}\right\rangle = \frac{1}{\sqrt{n}}\langle h_\ell, g_\ell\rangle = Q_\ell. \tag{6.70}$$

Thus $Q_\ell \equiv Q$ for all ℓ. At the readout boundary this common value is the output itself, while the backward kernel has the terminal behavior

$$Q = Q_d = f = \sqrt{\Phi_d}\,Z, \qquad G_d = \frac{1}{n}\|W_d\|^2 \xrightarrow{\mathbb{P}} 1, \tag{6.71}$$

where $Z := \left\langle h_d, W_d^\top \right\rangle / \|h_d\| \sim \mathcal{N}(0,1)$ is independent of the forward trajectory. Thus the limiting constant $Q = \sqrt{\Phi_{\bar{\tau}}}Z$ remains coupled to the terminal forward kernel, even though G_d converges to one.

To analyze the backward kernel G, complete the forward pass and then expose the backward variables from the readout toward the input. Define the reverse filtration

$$\mathcal{F}_\ell^g := \sigma\big(\mathcal{F}_d^h,\ \{g_k\}_{k=\ell}^d\big), \qquad 1 \le \ell \le d. \tag{6.72}$$

Thus $\mathcal{F}_{\ell+1}^g \subset \mathcal{F}_\ell^g$. The nontrivial transition is therefore $g_\ell \mid \mathcal{F}_{\ell+1}^g$ for $1 \le \ell \le d-1$.

Conditional on $\mathcal{F}_d^h$, the layerwise Gaussian residual blocks are independent. The variables $g_{\ell+1},\dots,g_d$ use only higher-layer residual blocks, so adjoining them to $\mathcal{F}_d^h$ to form $\mathcal{F}_{\ell+1}^g$ reveals W_ℓ only through the forward constraint $W_\ell h_\ell = \sqrt{n}\,h_{\ell+1}$. Writing $P_{h_\ell}^\perp := I_n - h_\ell h_\ell^\top / \|h_\ell\|^2$, the one-sided Gaussian conditioning identity in Lemma 2.26 leaves the $P_{h_\ell}^\perp$ block fresh. Multiplying its transpose by $g_{\ell+1}/\sqrt{n}$ gives

$$g_\ell = \frac{1}{\sqrt{n}} W_\ell^\top g_{\ell+1} \mid \mathcal{F}_{\ell+1}^g \overset{d}{=} \frac{Q_{\ell+1}}{\Phi_\ell}\frac{h_\ell}{\sqrt{n}} + \sqrt{G_{\ell+1}}\,P_{h_\ell}^\perp \xi_\ell. \tag{6.73}$$

Here $\xi_\ell \sim \mathcal{N}(0, I_n)$ is independent of $\mathcal{F}^g_{\ell+1}$.

The two terms in this representation are orthogonal, while $P^\perp_{h_\ell}$ has rank $n-1$. Thus, writing $\chi^2_{n-1} := \left\| P^\perp_{h_\ell} \xi_\ell \right\|^2$, we have $\chi^2_{n-1} \sim \chi^2(n-1)$ with a conditional law independent of $\mathcal{F}^g_{\ell+1}$, and we obtain the exact one-step formula

$$G_\ell = \frac{1}{n} \|g_\ell\|^2 = G_{\ell+1} \frac{\chi^2_{n-1}}{n} + \frac{Q^2_{\ell+1}}{n\,\Phi_\ell}. \tag{6.74}$$

From $\mathbb{E}[\chi^2_{n-1}] = n-1$ and $\mathrm{Var}(\chi^2_{n-1}) = 2(n-1)$ we read off the conditional drift and conditional variance:

$$\mathbb{E}\Big[\, G_\ell \,\Big|\, \mathcal{F}^g_{\ell+1} \,\Big] = G_{\ell+1} + \frac{1}{n}\left(\frac{Q^2_{\ell+1}}{\Phi_\ell} - G_{\ell+1} \right), \tag{6.75}$$

$$\mathrm{Var}\Big(G_\ell \,\Big|\, \mathcal{F}^g_{\ell+1} \Big) = \frac{2}{n}\left(1 - \frac{1}{n}\right) G^2_{\ell+1}. \tag{6.76}$$

For this reverse recursion, let $\widetilde{o}^g_{\ell+1}(n^{-1})$ denote the scalar analogue of the conditional remainder in (5.52), with conditioning on $\mathcal{F}^g_{\ell+1}$ and interpreted locally after stopping G on bounded sets. For $n \geq 2$, define the standardized chi-square innovation $\xi^G_{\ell,n} := (\chi^2_{n-1} - (n-1))/\sqrt{2(n-1)}$ and write the one-step recursion in Euler form as

$$\begin{aligned} G_\ell - G_{\ell+1} &= \frac{1}{n}\left(\frac{Q^2}{\Phi_\ell} - G_{\ell+1} \right) + \sqrt{\frac{2}{n}}\sqrt{1 - \frac{1}{n}}\, G_{\ell+1}\, \xi^G_{\ell,n} \\ &= \frac{1}{n}\left(\frac{Q^2}{\Phi_\ell} - G_{\ell+1} \right) + \frac{\sqrt{2}}{\sqrt{n}}\, G_{\ell+1}\, \xi^G_{\ell,n} + \widetilde{o}^g_{\ell+1}(n^{-1}). \end{aligned} \tag{6.77}$$

Conditional on $\mathcal{F}^g_{\ell+1}$, the innovation $\xi^G_{\ell,n}$ has mean zero and variance one, its law does not depend on the filtration, and $\xi^G_{\ell,n} \xrightarrow{d} \mathcal{N}(0,1)$. The remainder in the last line is precisely $\sqrt{2/n}\left(\sqrt{1 - 1/n} - 1\right) G_{\ell+1} \xi^G_{\ell,n}$; on stopped sets where $G_{\ell+1}$ is bounded, it has conditional mean zero and conditional second moment $O(n^{-3})$, and hence has the stated conditional order.

Finally, (6.77) has the form of an Euler–Maruyama step when the layers are traversed from the readout toward the input. At the continuum level, set $s = \bar{\tau} - \tau$ and $\widehat{G}_s := G_{\bar{\tau}-s}$, so that $s = 0$ corresponds to the readout. In this clock, the moment calculation suggests a standard Itô equation without requiring a nonstandard forward-depth convention.

Proposition 6.15 (Formal Limit for (Q, G) at Initialization (One Sample, Linear))**.** *Assume $d, n \to \infty$ with $d/n \to \bar{\tau} \in (0, \infty)$. Then $Q_\tau \equiv Q = \sqrt{\Phi_{\bar{\tau}}}\, Z$ is constant in depth. Conditional on the realized forward path and the terminal variable Z from (6.71), the Euler expansion* (6.77) *suggests the formal reverse-depth limit*

$$\mathrm{d}\widehat{G}_s = \left(\frac{Q^2}{\Phi_{\bar{\tau}-s}} - \widehat{G}_s \right) \mathrm{d}s + \sqrt{2}\, \widehat{G}_s\, \mathrm{d}B^G_s, \qquad \widehat{G}_0 = 1. \tag{6.78}$$

The Brownian motion B^G is independent of the forward path and Z. The exact finite-width content is (6.71), (6.74), and (6.77); the stochastic equation is its formal diffusion summary.

6.3.2 One Training Step: Overlap Processes R^h, R^g, and $\dot{\Phi}$

We now take one simultaneous gradient-descent step, with all gradients evaluated at the initialized network, while keeping the proportional depth scaling. Let $L(f, y)$ be a differentiable loss and write

$$r := \partial_f L(f, y). \tag{6.79}$$

(For the mean-squared error $L(f, y) = \frac{1}{2}(f - y)^2$ this is the usual residual $r = f - y$.) We use the learning-rate scaling $\eta = \eta_0/d$.

For $\ell = 1, \dots, d-1$ the gradient of the scalar output f with respect to W_ℓ is

$$\nabla_{W_\ell} f = \frac{1}{n}\, g_{\ell+1}\, h_\ell^\top, \tag{6.80}$$

so the simultaneous hidden-layer update is

$$W_\ell^+ := W_\ell - \eta\, r\, \nabla_{W_\ell} f = W_\ell - \frac{\eta\, r}{n}\, g_{\ell+1} h_\ell^\top \qquad \text{(linear, one sample)}. \tag{6.81}$$

Let h_ℓ^+ denote the resulting forward features and let Δh_ℓ denote their change:

$$h_1^+ = h_1, \qquad h_{\ell+1}^+ = \frac{1}{\sqrt{n}} W_\ell^+ h_\ell^+, \qquad \Delta h_\ell := h_\ell^+ - h_\ell. \tag{6.82}$$

To track this feature change, define the scalar overlaps

$$R_\ell^h := \frac{1}{n}\, \langle \Delta h_\ell, h_\ell \rangle\,, \qquad R_\ell^g := \frac{1}{\sqrt{n}}\, \langle \Delta h_\ell, g_\ell \rangle\,, \qquad \dot{\Phi}_\ell := \frac{1}{n}\, \|\Delta h_\ell\|^2\,. \tag{6.83}$$

Thus $\dot{\Phi}_\ell$ is *not* a time derivative of Φ_ℓ; it is the squared size of the feature change. If $\Phi_\ell^+ := n^{-1} \left\| h_\ell^+ \right\|^2$, then the actual kernel change is exactly

$$\Phi_\ell^+ - \Phi_\ell = 2R_\ell^h + \dot{\Phi}_\ell. \tag{6.84}$$

Expanding the updated forward recursion gives

$$\Delta h_{\ell+1} = \frac{1}{\sqrt{n}} W_\ell \Delta h_\ell - \eta\, r\big(\Phi_\ell + R_\ell^h\big) \frac{1}{\sqrt{n}} g_{\ell+1}, \qquad \Delta h_1 = 0. \tag{6.85}$$

The extra R_ℓ^h comes from the interaction between the weight change and the feature change within the same finite step.

The overlap R^g satisfies an exact recursion. Pair (6.85) with $g_{\ell+1}$ and use $g_\ell = \frac{1}{\sqrt{n}} W_\ell^\top g_{\ell+1}$:

$$R_{\ell+1}^g = \frac{1}{\sqrt{n}}\, \langle \Delta h_{\ell+1}, g_{\ell+1} \rangle = \frac{1}{\sqrt{n}}\, \langle \Delta h_\ell, g_\ell \rangle - \eta\, r\big(\Phi_\ell + R_\ell^h\big) \frac{1}{n}\, \|g_{\ell+1}\|^2\,. \tag{6.86}$$

Since $G_{\ell+1} = \frac{1}{n}\, \|g_{\ell+1}\|^2$, we obtain the exact increment

$$R_{\ell+1}^g - R_\ell^g = -\eta\, r\big(\Phi_\ell + R_\ell^h\big) G_{\ell+1}. \tag{6.87}$$

Passing to depth-time $\tau = \ell/n$ formally suggests the continuum equation

$$\mathrm{d}R_\tau^g = -\frac{\eta_0}{\bar{\tau}}\, r\big(\Phi_\tau + R_\tau^h\big) G_\tau\, \mathrm{d}\tau, \qquad R_0^g = 0. \tag{6.88}$$

Equivalently, integrating (6.88) gives the integral form

$$R_\tau^g = -\frac{\eta_0}{\bar\tau}\, r \int_0^\tau (\Phi_s + R_s^h) G_s \,\mathrm{d}s. \tag{6.89}$$

The remaining overlaps R^h and $\dot\Phi$ pick up $O(n^{-1/2})$ fluctuations from the random matrix product. Their one-step laws require conditioning on both the initialized forward and backward passes. The active matrix W_ℓ is then constrained by $W_\ell h_\ell = \sqrt{n}\, h_{\ell+1}$ and $W_\ell^\top g_{\ell+1} = \sqrt{n}\, g_\ell$. For this analysis, use the increasing filtration

$$\mathcal{F}_0^{\Delta h} := \mathcal{F}_1^g, \qquad \mathcal{F}_\ell^{\Delta h} := \sigma(\mathcal{F}_1^g, \{\Delta h_k\}_{1\le k\le \ell}), \qquad 1 \le \ell \le d. \tag{6.90}$$

All initialized gradients and weight changes are measurable with respect to $\mathcal{F}_1^g$. Given this field, layerwise conditional independence and the change recursion imply that each Δh_k with $k \le \ell$ uses only residual blocks from matrices W_j with $j < \ell$. Thus adjoining the earlier feature changes reveals no additional action of the active matrix.

The two-sided decomposition in Lemma A.5 shows that the conditional law is a possibly degenerate Gaussian with

$$\begin{aligned}
\mathbb{E}\Big[\Delta h_{\ell+1} \,\Big|\, \mathcal{F}_\ell^{\Delta h}\Big] &= \frac{R_\ell^h}{\Phi_\ell} h_{\ell+1} + \frac{1}{\sqrt{n}\, G_{\ell+1}} \left(R_{\ell+1}^g - \frac{Q}{\Phi_\ell} R_\ell^h\right) g_{\ell+1}, \\
\mathrm{Cov}\Big(\Delta h_{\ell+1} \,\Big|\, \mathcal{F}_\ell^{\Delta h}\Big) &= \left(\dot\Phi_\ell - \frac{(R_\ell^h)^2}{\Phi_\ell}\right) P_{g_{\ell+1}}^\perp, \qquad 1 \le \ell \le d-1.
\end{aligned} \tag{6.91}$$

Here $P_{g_{\ell+1}}^\perp$ is the orthogonal projector defined analogously to $P_{h_\ell}^\perp$ above. The exact recursion (6.87) makes $R_{\ell+1}^g$ measurable with respect to $\mathcal{F}_\ell^{\Delta h}$, so its appearance in the conditional mean is not circular.

The whole forward trajectory belongs to $\mathcal{F}_\ell^{\Delta h}$, so the forward increment $\Phi_{\ell+1} - \Phi_\ell$ is fixed under this conditioning. Taking scalar moments of (6.91) gives the exact conditional mean increments

$$\begin{aligned}
\mathbb{E}\Big[R_{\ell+1}^h - R_\ell^h \,\Big|\, \mathcal{F}_\ell^{\Delta h}\Big] &= \frac{R_\ell^h}{\Phi_\ell}(\Phi_{\ell+1} - \Phi_\ell) + \frac{Q}{nG_{\ell+1}} \left(R_{\ell+1}^g - \frac{Q}{\Phi_\ell} R_\ell^h\right), \\
\mathbb{E}\Big[\dot\Phi_{\ell+1} - \dot\Phi_\ell \,\Big|\, \mathcal{F}_\ell^{\Delta h}\Big] &= \frac{(R_\ell^h)^2}{\Phi_\ell^2}(\Phi_{\ell+1} - \Phi_\ell) \\
&\quad + \frac{1}{n}\left(\frac{(R_{\ell+1}^g)^2}{G_{\ell+1}} - \frac{Q^2 (R_\ell^h)^2}{G_{\ell+1}\Phi_\ell^2} - \dot\Phi_\ell + \frac{(R_\ell^h)^2}{\Phi_\ell}\right).
\end{aligned} \tag{6.92}$$

The corresponding exact conditional covariances are

$$\begin{aligned}
n\ \mathrm{Var}\Big(R_{\ell+1}^h \,\Big|\, \mathcal{F}_\ell^{\Delta h}\Big) &= \left(\dot\Phi_\ell - \frac{(R_\ell^h)^2}{\Phi_\ell}\right)\left(\Phi_{\ell+1} - \frac{Q^2}{nG_{\ell+1}}\right), \\
n\ \mathrm{Cov}\Big(R_{\ell+1}^h, \dot\Phi_{\ell+1} \,\Big|\, \mathcal{F}_\ell^{\Delta h}\Big) &= \frac{2R_\ell^h}{\Phi_\ell}\left(\dot\Phi_\ell - \frac{(R_\ell^h)^2}{\Phi_\ell}\right)\left(\Phi_{\ell+1} - \frac{Q^2}{nG_{\ell+1}}\right), \\
n\ \mathrm{Var}\Big(\dot\Phi_{\ell+1} \,\Big|\, \mathcal{F}_\ell^{\Delta h}\Big) &= 2\left(1 - \frac{1}{n}\right)\left(\dot\Phi_\ell - \frac{(R_\ell^h)^2}{\Phi_\ell}\right)^2 \\
&\quad + \frac{4(R_\ell^h)^2}{\Phi_\ell^2}\left(\dot\Phi_\ell - \frac{(R_\ell^h)^2}{\Phi_\ell}\right)\left(\Phi_{\ell+1} - \frac{Q^2}{nG_{\ell+1}}\right).
\end{aligned} \tag{6.93}$$

Here $\Phi_{\ell+1} - Q^2/(nG_{\ell+1}) = n^{-1}\left\|P^{\perp}_{g_{\ell+1}} h_{\ell+1}\right\|^2$ is nonnegative. The part of the conditional mean parallel to $g_{\ell+1}$ contributes to the drifts but not to these covariances, because the fresh Gaussian block lies in $g^{\perp}_{\ell+1}$.

On bounded state sets, substituting (6.87) into (6.92) shows that replacing $R^g_{\ell+1}$ by R^g_ℓ changes both conditional mean increments by $O(n^{-2})$; in the second identity, the pure square of the $O(n^{-1})$ overlap difference contributes only $O(n^{-3})$. We include these terms in $\widetilde{o}^{\Delta h}_\ell(n^{-1})$, used componentwise in the sense of the conditional remainder in (5.52), with conditioning on $\mathcal{F}^{\Delta h}_\ell$ and locally where Φ_ℓ and $G_{\ell+1}$ stay away from zero. The forward one-step law gives $\Phi_{\ell+1} - \Phi_\ell = o_{\mathbb{P}}(1)$, while the alignment and finite-rank corrections in (6.93) are $O_{\mathbb{P}}(n^{-1})$ locally, so the exact identities reduce to the leading conditional covariance at Euler scale:

$$n\,\mathrm{Cov}\left(\begin{pmatrix} R^h_{\ell+1} \\ \dot\Phi_{\ell+1}\end{pmatrix} \,\middle|\, \mathcal{F}^{\Delta h}_\ell\right) = \begin{pmatrix} \Phi_\ell\dot\Phi_\ell - (R^h_\ell)^2 & 2R^h_\ell\left(\dot\Phi_\ell - \dfrac{(R^h_\ell)^2}{\Phi_\ell}\right) \\ 2R^h_\ell\left(\dot\Phi_\ell - \dfrac{(R^h_\ell)^2}{\Phi_\ell}\right) & 2\left(\dot\Phi^2_\ell - \dfrac{(R^h_\ell)^4}{\Phi^2_\ell}\right)\end{pmatrix} + o_{\mathbb{P}}(1). \tag{6.94}$$

By Cauchy–Schwarz, $(R^h_\ell)^2 \le \Phi_\ell\dot\Phi_\ell$, so the radicands below are nonnegative.

Combining the exact conditional means in (6.92) with a factorization of the leading covariance in (6.94) gives the Euler expansions

$$\begin{aligned} R^h_{\ell+1} - R^h_\ell &= \frac{R^h_\ell}{\Phi_\ell}(\Phi_{\ell+1} - \Phi_\ell) + \frac{1}{\sqrt{n}}\sqrt{\Phi_\ell\dot\Phi_\ell - (R^h_\ell)^2}\,\xi^h_\ell \\ &\quad + \frac{Q}{nG_{\ell+1}}\left(R^g_\ell - \frac{Q}{\Phi_\ell}R^h_\ell\right) + \widetilde{o}^{\Delta h}_\ell(n^{-1}), \end{aligned} \tag{6.95}$$

$$\begin{aligned} \dot\Phi_{\ell+1} - \dot\Phi_\ell &= \frac{(R^h_\ell)^2}{\Phi^2_\ell}(\Phi_{\ell+1} - \Phi_\ell) + \frac{2R^h_\ell}{\sqrt{n}\,\Phi_\ell}\sqrt{\Phi_\ell\dot\Phi_\ell - (R^h_\ell)^2}\,\xi^h_\ell + \sqrt{\frac{2}{n}}\left(\dot\Phi_\ell - \frac{(R^h_\ell)^2}{\Phi_\ell}\right)\xi^{\dot\Phi}_\ell \\ &\quad + \frac{1}{n}\left(\frac{(R^g_\ell)^2}{G_{\ell+1}} - \frac{Q^2}{G_{\ell+1}}\frac{(R^h_\ell)^2}{\Phi^2_\ell} - \dot\Phi_\ell + \frac{(R^h_\ell)^2}{\Phi_\ell}\right) + \widetilde{o}^{\Delta h}_\ell(n^{-1}). \end{aligned} \tag{6.96}$$

In (6.95) and (6.96), ξ^h_ℓ is the Gaussian fluctuation shared by both increments, whereas $\xi^{\dot\Phi}_\ell$ records the remaining fluctuation in the squared feature change and appears only in the second. Once the quantities in $\mathcal{F}^{\Delta h}_\ell$ are held fixed, these two random variables are independent in the Euler representation: ξ^h_ℓ is standard Gaussian, while $\xi^{\dot\Phi}_\ell$ converges to a standard Gaussian as $n \to \infty$. The exact $\dot\Phi$ increment also contains a smaller term proportional to $((\xi^h_\ell)^2 - 1)/n$; its conditional second moment is $O(n^{-2})$, so it is included in $\widetilde{o}^{\Delta h}_\ell(n^{-1})$. Thus (6.92) and (6.93) are exact finite-width identities, whereas (6.95) and (6.96) are local conditional asymptotic expansions.

At a formal continuum level, the Euler expansions suggest the following coupled equations:

$$\mathrm{d}R^h_\tau = \sqrt{2}\,R^h_\tau\,\mathrm{d}B^\Phi_\tau + \sqrt{\Phi_\tau\dot\Phi_\tau - (R^h_\tau)^2}\,\mathrm{d}B^h_\tau + \left(\frac{Q}{G_\tau}\,R^g_\tau - \frac{Q^2}{\Phi_\tau\,G_\tau}\,R^h_\tau\right)\mathrm{d}\tau, \tag{6.97}$$

$$\begin{aligned} \mathrm{d}\dot\Phi_\tau &= \sqrt{2}\,\frac{(R^h_\tau)^2}{\Phi_\tau}\,\mathrm{d}B^\Phi_\tau + \frac{2R^h_\tau}{\Phi_\tau}\sqrt{\Phi_\tau\dot\Phi_\tau - (R^h_\tau)^2}\,\mathrm{d}B^h_\tau + \sqrt{2}\left(\dot\Phi_\tau - \frac{(R^h_\tau)^2}{\Phi_\tau}\right)\mathrm{d}B^{\dot\Phi}_\tau \\ &\quad + \left(\frac{(R^g_\tau)^2}{G_\tau} - \frac{Q^2}{G_\tau}\,\frac{(R^h_\tau)^2}{\Phi^2_\tau} - \dot\Phi_\tau + \frac{(R^h_\tau)^2}{\Phi_\tau}\right)\mathrm{d}\tau, \end{aligned} \tag{6.98}$$

In this formal covariance factorization, B^h and $B^{\dot\Phi}$ may be taken independent of each other and of the forward driver B^Φ in (6.69). In particular, this factorization gives $\mathrm{d}\langle R^h, \dot\Phi\rangle_\tau = 2R^h_\tau \dot\Phi_\tau \,\mathrm{d}\tau$. Together with (6.88) and the formal backward equation (6.78), these relations motivate the consolidated summary below.

Theorem 6.16 (Scalar Proportional-Limit Kernel Dynamics (Formal))**.** *In the proportional regime $d, n \to \infty$ with $d/n \to \bar\tau \in (0,\infty)$, the preceding calculations formally suggest the following six-equation continuum system. For the backward kernel, write $s = \bar\tau - \tau$ and $\widehat G_s = G_{\bar\tau - s}$:*

$$\mathrm{d}\Phi_\tau = \sqrt{2}\,\Phi_\tau \,\mathrm{d}B^\Phi_\tau, \qquad \Phi_0 = \frac{1}{n_0}\|x\|^2, \qquad \text{[forward in } \tau\text{]}, \tag{6.99}$$

$$Q_\tau = Q \text{ (constant in depth)}, \qquad Q = \sqrt{\Phi_{\bar\tau}}Z, \qquad \text{[terminal coupling]}, \tag{6.100}$$

$$\mathrm{d}\widehat G_s = \left(\frac{Q^2}{\Phi_{\bar\tau - s}} - \widehat G_s\right)\mathrm{d}s + \sqrt{2}\,\widehat G_s \,\mathrm{d}B^G_s, \qquad \widehat G_0 = 1, \qquad \text{[reverse depth } s\text{]}, \tag{6.101}$$

$$\mathrm{d}R^g_\tau = -\frac{\eta_0}{\bar\tau}\, r(\Phi_\tau + R^h_\tau) G_\tau \,\mathrm{d}\tau, \qquad R^g_0 = 0, \qquad \text{[forward in } \tau\text{]}. \tag{6.102}$$

The two feature-change overlaps also use the forward clock τ:

$$\mathrm{d}R^h_\tau = \sqrt{2}\,R^h_\tau \,\mathrm{d}B^\Phi_\tau + \sqrt{\Phi_\tau\dot\Phi_\tau - (R^h_\tau)^2}\,\mathrm{d}B^h_\tau + \left(\frac{Q}{G_\tau}R^g_\tau - \frac{Q^2}{\Phi_\tau G_\tau}R^h_\tau\right)\mathrm{d}\tau, \qquad R^h_0 = 0, \tag{6.103}$$

$$\begin{aligned}\mathrm{d}\dot\Phi_\tau = {}& \sqrt{2}\,\frac{(R^h_\tau)^2}{\Phi_\tau}\,\mathrm{d}B^\Phi_\tau + \frac{2R^h_\tau}{\Phi_\tau}\sqrt{\Phi_\tau\dot\Phi_\tau - (R^h_\tau)^2}\,\mathrm{d}B^h_\tau + \sqrt{2}\left(\dot\Phi_\tau - \frac{(R^h_\tau)^2}{\Phi_\tau}\right)\mathrm{d}B^{\dot\Phi}_\tau \\ & + \left(\frac{(R^g_\tau)^2}{G_\tau} - \frac{Q^2(R^h_\tau)^2}{G_\tau\Phi^2_\tau} - \dot\Phi_\tau + \frac{(R^h_\tau)^2}{\Phi_\tau}\right)\mathrm{d}\tau, \qquad \dot\Phi_0 = 0.\end{aligned} \tag{6.104}$$

Here $Z \sim \mathcal{N}(0,1)$ is independent of the forward path. The Brownian motions B^h and $B^{\dot\Phi}$ may be taken independent of each other and of B^Φ. The backward driver B^G is independent of the forward path and Z; no independence from B^h or $B^{\dot\Phi}$ is asserted.

To read off the effect of the step on the prediction, let $f^+ := \frac{1}{\sqrt{n}} W_d h^+_d$. The readout is fixed and $g_d = W_d^\top$, so telescoping (6.87) from $R^g_1 = 0$ gives

$$f^+ - f = \frac{1}{\sqrt{n}}\langle \Delta h_d, g_d\rangle = R^g_d = -\eta r \sum_{\ell=1}^{d-1}\left(\Phi_\ell + R^h_\ell\right) G_{\ell+1}. \tag{6.105}$$

Thus the terminal overlap R^g_d gives the exact prediction change, while (6.84) recovers the feature-kernel change from R^h_ℓ and $\dot\Phi_\ell$.

The factor $\Phi_\ell + R^h_\ell = n^{-1}\left\langle h_\ell, h^+_\ell\right\rangle$ is the overlap between the original and updated features. Replacing it by Φ_ℓ gives the linearized prediction update $-\eta r\sum_{\ell=1}^{d-1}\Phi_\ell G_{\ell+1}$. Indeed, $\|\nabla_{W_\ell} f\|^2_{\mathrm{F}} = \Phi_\ell G_{\ell+1}$, so summing these squared gradient norms gives the initialized scalar NTK restricted to the trained hidden matrices; compare Proposition 2.25. The remaining $R^h_\ell G_{\ell+1}$ contribution records interactions among the simultaneous layer updates: an earlier feature change is passed through later matrices that have also been updated. All gradients are still evaluated at initialization. The network is linear in its input, but not jointly linear in all its weights.

Extensions to multiple inputs and nonlinear activations remain part of the ongoing work [33].

Appendix A

Technical Complements

A.1 Formal Derivation of the Transport PDE

Here is the standard weak-form derivation of (1.11) from the mean-field system in (1.8), with the velocity field b defined in (1.10).

Let $\rho_t^{(n)}$ be the empirical measure from (1.9), and let $q : \mathbb{R} \to \mathbb{R}$ be a smooth compactly supported test function. Then

$$\int q(x)\, \rho_t^{(n)}(dx) = \frac{1}{n} \sum_{i=1}^{n} q(X_i(t)). \tag{A.1}$$

Differentiate in time and use the chain rule with (1.8):

$$\begin{aligned} \frac{d}{dt} \int q(x)\, \rho_t^{(n)}(dx) &= \frac{1}{n} \sum_{i=1}^{n} q'(X_i(t))\, \partial_t X_i(t) \\ &= \int q'(x)\, b(x, \rho_t^{(n)})\, \rho_t^{(n)}(dx). \end{aligned} \tag{A.2}$$

Passing formally to the limit gives

$$\frac{d}{dt} \int q(x)\, \rho_t(dx) = \int q'(x)\, b(x, \rho_t)\, \rho_t(dx). \tag{A.3}$$

This is precisely the distributional formulation of (1.11); it does not require ρ_t to have a density. The equation is a prototype of a McKean–Vlasov PDE because the velocity field depends on the law ρ_t itself.

A.2 A Sufficient Condition for Stochastic Equicontinuity

This technical lemma supplies the tightness condition used in the function-space strengthening following Theorem 2.3. Let f_n denote the width-n network in (2.1).

Lemma A.1 (Hölder Activation Implies Tightness)**.** *Assume that for some $\gamma \in (0, 1]$ and $C_\varphi > 0$,*

$$|\varphi(u) - \varphi(v)| \le C_\varphi |u - v|^\gamma \qquad \forall u, v \in \mathbb{R}. \tag{A.4}$$

Fix a compact $K \subset \mathbb{R}^{n_0}$. For every $p \geq 2$, there is a constant $C_{p,K} < \infty$ such that, for all n and all $x, x' \in K$,

$$\mathbb{E}\big[|f_n(x) - f_n(x')|^p\big] \leq C_{p,K} \, \|x - x'\|^{\gamma p}. \tag{A.5}$$

In particular, choose p so that $\gamma p > n_0$ and set $\beta := \gamma p - n_0 > 0$. Then the right-hand side of (A.5) is $C_{p,K} \, \|x - x'\|^{n_0+\beta}$, and $\{f_n|_K\}_{n\geq 1}$ is tight in $C(K)$.

Proof. Let $\mathcal{F} := \sigma(h_i(x), h_i(x') : i \in [n])$. By the conditional Gaussian output-layer identity (2.3), the increment $f_n(x) - f_n(x')$, conditional on $\mathcal{F}$, is centered Gaussian with variance $n^{-1}\sum_{i=1}^n |\varphi(h_i(x)) - \varphi(h_i(x'))|^2$. Hence the Gaussian moment formula and convexity, with C_p denoting the pth absolute moment of a standard Gaussian, give

$$\begin{aligned}
\mathbb{E}\big[|f_n(x) - f_n(x')|^p\big] &= C_p \, \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n |\varphi(h_i(x)) - \varphi(h_i(x'))|^2\right)^{p/2}\right] \\
&\leq C_p \, \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n |\varphi(h_i(x)) - \varphi(h_i(x'))|^p\right] \\
&\leq C_p C_\varphi^p \, \mathbb{E}\left[\left|\frac{\langle w, x - x'\rangle}{\sqrt{n_0}}\right|^{\gamma p}\right] \leq C_{p,K} \, \|x - x'\|^{\gamma p}.
\end{aligned} \tag{A.6}$$

The same conditional Gaussian identity at one fixed $x_0 \in K$, together with $|\varphi(u)| \leq |\varphi(0)| + C_\varphi |u|^\gamma$, gives $\sup_n \mathbb{E}[|f_n(x_0)|^p] < \infty$. Choosing p with $\gamma p > n_0$ and applying the Kolmogorov–Chentsov tightness criterion proves the final assertion; each f_n has continuous sample paths because φ is Hölder continuous. □

Remark A.2. Combined with the finite-dimensional convergence in Theorem 2.3, this tightness yields weak convergence in $C(K)$ to a continuous version of the limiting Gaussian process.

A.3 Convergence in the Skorohod Topology

Many processes in these notes arise by *piecewise-constant* interpolation of a layer-indexed Markov chain (e.g. $\Phi_\tau^{(n)} := \Phi_{\lfloor \tau n \rfloor}$). Such interpolations are càdlàg—right-continuous with left limits—and therefore live naturally in a Skorohod path space. This appendix records the definitions needed to interpret convergence statements like

$$X^{(n)} \xrightarrow{d} X \qquad \text{in } D([0,T], S), \tag{A.7}$$

where S is a metric state space. We follow [34, Appendix A] (see also standard references such as Ethier–Kurtz and Kallenberg).

Let (S, d_S) be a complete separable metric space, and write $D([0,T], S)$ for the set of càdlàg functions $x : [0,T] \to S$. When $S = \mathbb{R}^k$, we use the Euclidean distance $d_S(x,y) = \|x - y\|$.

To compare two càdlàg paths up to small shifts of their jump times, let Λ_T be the set of increasing homeomorphisms $\lambda : [0,T] \to [0,T]$ with $\lambda(0) = 0$ and $\lambda(T) = T$.

Skorohod J_1 metric. For $x, y \in D([0,T], S)$ define the (equivalent) J_1 metric

$$d_{J_1}(x,y) := \inf_{\lambda \in \Lambda_T} \left\{ \sup_{t\in[0,T]} d_S\big(x(\lambda(t)), y(t)\big) \;\vee\; \sup_{t\in[0,T]} |\lambda(t) - t| \right\}. \tag{A.8}$$

Then $x_n \to x$ in the Skorohod J_1 topology if and only if there exist $\lambda_n \in \Lambda_T$ such that

$$\sup_{t\in[0,T]} |\lambda_n(t) - t| \to 0 \qquad \text{and} \qquad \sup_{t\in[0,T]} d_S\big(x_n(\lambda_n(t)), x(t)\big) \to 0, \tag{A.9}$$

which is the same criterion stated as [34, (A.1)].

The Skorohod space accommodates the piecewise-constant approximations used here. Its topology allows small shifts in jump times. When the limiting path is continuous, as for the diffusions considered in these notes, convergence in this topology is equivalent to uniform convergence of paths.

Equipped with the J_1 topology, $D([0,T], S)$ is a Polish space (separable and completely metrizable), so weak convergence of probability measures on $D([0,T], S)$ is well-defined. We write $X^{(n)} \xrightarrow{d} X$ when the law of $X^{(n)}$ converges weakly to the law of X as random elements of $D([0,T], S)$.

Definition A.3 (Local Skorohod Convergence)**.** Let $U \subseteq \mathbb{R}^k$ be open, and write $D \Subset U$ when D is open and bounded and $\overline{D}$ is a compact subset of U. Let $X^{(n)}$ be càdlàg U-valued processes, and let X be a continuous process defined up to its lifetime in U. For $D \Subset U$ and $Y = X^{(n)}$ or $Y = X$, define the exit time

$$\tau_D(Y) := \inf\{t \in [0,T] : Y_t \notin D\}, \tag{A.10}$$

with the convention $\inf \varnothing := T$. The stopped paths are frozen at their exit values and are therefore defined on all of $[0,T]$, even if X later leaves U or explodes. We say $X^{(n)}$ converges *locally* to X in distribution if, for every such D at which the stopping map is almost surely continuous under X,

$$X^{(n)}_{\cdot\wedge\tau_D(X^{(n)})} \xrightarrow{d} X_{\cdot\wedge\tau_D(X)} \qquad \text{in } D([0,T], \mathbb{R}^k). \tag{A.11}$$

A standard sufficient condition is $\tau_D(X) = \tau_{\overline{D}}(X)$ almost surely.

This formulation is useful when the limiting diffusion can leave compact sets or explode. The stopping times in Theorem 5.10 localize the process against norm growth. A compact localization inside $\mathrm{SPD}(m)$ must in addition stay a positive distance from the singular boundary.

A standard generator route combines a Feller construction—see [15, Chapter 8, Theorem 2.5] under smoother coefficient assumptions—with the discrete-time criterion [28, Theorem 17.28]; [34, Appendix A] develops this route for the Neural Covariance SDE. For the present notes, the following "moment" version is a convenient sufficient condition.

Theorem A.4 (Diffusion Approximation for Markov Chains)**.** *Let $U \subseteq \mathbb{R}^k$ be open, and for each n let $(X_j^{(n)})_{j\ge 0}$ be a time-homogeneous Markov chain taking values in U. Define the piecewise-constant interpolation $X_t^{(n)} := X^{(n)}_{\lfloor nt \rfloor}$. Assume $X_0^{(n)} \xrightarrow{d} X_0$, where X_0 takes values in U. Assume there exist functions $b : U \to \mathbb{R}^k$ and $\Sigma : U \to \mathbb{R}^{k\times k}$ such that for every compact $K \subset U$,*

$$\sup_{x\in K} \Big\| n\,\mathbb{E}[X_1^{(n)} - X_0^{(n)} \mid X_0^{(n)} = x] - b(x) \Big\| \to 0, \tag{A.12}$$

$$\sup_{x\in K} \Big\| n\ \mathrm{Cov}(X_1^{(n)} - X_0^{(n)} \mid X_0^{(n)} = x) - \Sigma(x) \Big\| \to 0, \tag{A.13}$$

$$\sup_{x\in K} n\,\mathbb{E}\big[\|X_1^{(n)} - X_0^{(n)}\|^3 \mid X_0^{(n)} = x\big] \to 0. \tag{A.14}$$

Here the conditional moments refer to the one-step transition kernel started from x. Let $\sigma : U \to \mathbb{R}^{k\times k}$ satisfy $\sigma(x)\sigma(x)^\top = \Sigma(x)$, and assume that b and σ are locally Lipschitz. Let X be the maximal solution, started from X_0, of

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t. \tag{A.15}$$

Then, for each $T > 0$, $X^{(n)} \xrightarrow{d} X$ locally in the sense of Definition A.3.

Proof. Fix an admissible localization domain $D \Subset U$ from Definition A.3, let P_n be the transition kernel of $X^{(n)}$, and let $\Delta_n(x)$ have the one-step increment law started from x. We use the localized Markov-chain-to-diffusion theorem [47, Theorem 11.2.3]; compare the generator argument in [34, Appendix A, especially Proposition A.6]. The coefficient criterion [47, Lemma 11.2.1] reduces its local-characteristic hypotheses to the following moment checks. The first-moment assumption gives the drift. If $\mu_n(x) := \mathbb{E}[\Delta_n(x)]$, then the raw second moment is $n\,\mathrm{Cov}(\Delta_n(x)) + n\mu_n(x)\mu_n(x)^\top$. The second term is $O(n^{-1})$ locally uniformly because $n\mu_n$ is locally bounded, so the limiting second moment is Σ. Finally, the third-moment assumption and Markov's inequality give $n\mathbb{P}(\|\Delta_n(x)\| > \varepsilon) \to 0$ locally uniformly for every $\varepsilon > 0$; the same bound shows that truncation at a fixed jump size does not change the first two limits. The cited coefficient criterion therefore gives, for $f \in C_c^\infty(U)$,

$$\begin{aligned} A_n f(x) &:= n\int_U (f(y) - f(x))P_n(x,\mathrm{d}y) \\ &\longrightarrow b(x)^\top \nabla f(x) + \frac{1}{2}\,\mathrm{Tr}\Big(\Sigma(x)D^2 f(x)\Big) =: Af(x), \end{aligned} \tag{A.16}$$

locally uniformly in x.

The localized diffusion-approximation theorem [47, Theorem 11.2.3] now gives Skorohod convergence up to exit from D. The convergence of the initial laws identifies the starting distribution, while local Lipschitz continuity of b and σ gives uniqueness of the limiting SDE up to exit from D. The regularity condition on D permits passage to the stopped paths in Definition A.3. Since the argument applies to every admissible $D \Subset U$, it yields the claimed local convergence. □

A.4 Affine-Invariant Metric for the Covariance SDE

Proof of Proposition 6.8. Congruence invariance follows from $(P\Phi P^\top)^{-1} = P^{-\top}\Phi^{-1}P^{-1}$ and cyclicity of the trace. It remains to identify the metric in the coordinates of (6.35). For $A \in \mathrm{Sym}(m)$, let $F_A(\Phi) := \mathrm{Tr}(A\Phi)$, so that $\mathrm{d}F_A(\Phi)[V] = \mathrm{Tr}(AV)$. Applying F_A and F_B to (6.42), the normalization of S_τ gives

$$\mathrm{d}\langle F_A(\Phi_\tau), F_B(\Phi_\tau)\rangle = 2\,\mathrm{Tr}(A\Phi_\tau B\Phi_\tau)\,\mathrm{d}\tau. \tag{A.17}$$

The metric-dual tangent vector of $\mathrm{d}F_A(\Phi)$ is $2\Phi A\Phi$, since

$$g_\Phi(2\Phi A\Phi, V) = \mathrm{Tr}(AV). \tag{A.18}$$

Writing g_Φ^{-1} for the cometric, we therefore have

$$g_\Phi^{-1}(\mathrm{d}F_A, \mathrm{d}F_B) = g_\Phi(2\Phi A\Phi, 2\Phi B\Phi) = 2\,\mathrm{Tr}(A\Phi B\Phi), \tag{A.19}$$

so the covariance tensor of the SDE is the cometric associated with g_Φ.

We now pass explicitly to the distinct upper-triangular coordinates. For $\alpha \le \beta$, let $x^{\alpha\beta}(\Phi) := \Phi^{\alpha\beta}$ and $C_{\alpha\beta} := (E_{\alpha\beta} + E_{\beta\alpha})/2$, where $E_{\alpha\beta}$ is the standard matrix unit. Then $\mathrm{d}x^{\alpha\beta}(V) = V^{\alpha\beta} = \mathrm{Tr}(C_{\alpha\beta}V)$, and hence

$$\begin{aligned} g_\Phi^{-1}(\mathrm{d}x^{\alpha\beta}, \mathrm{d}x^{\gamma\delta}) &= 2\,\mathrm{Tr}(C_{\alpha\beta}\Phi C_{\gamma\delta}\Phi) \\ &= \Phi^{\alpha\gamma}\Phi^{\beta\delta} + \Phi^{\alpha\delta}\Phi^{\beta\gamma} = \Sigma^{\alpha\beta,\gamma\delta}(\Phi). \end{aligned} \tag{A.20}$$

Thus $\Sigma(\Phi)$ is exactly the matrix of the cometric in the coordinate coframe $(\mathrm{d}x^{\alpha\beta})_{\alpha\le\beta}$. The metric matrix in the dual coordinate tangent basis is therefore $\Sigma(\Phi)^{-1}$. □

A.5 Gaussian Conditioning

This section collects two useful Gaussian conditioning identities. The first is a *one-sided* conditioning identity used when we condition only on the forward information $W\varphi$ (see Lemma 2.26). The second is a *two-sided* projection decomposition used when we condition simultaneously on forward and backward information $(W\varphi,\ W^\top g)$ (see Lemma A.5). Formally, the two-sided statement reduces to the one-sided one when the backward information is absent (e.g. by setting $g = 0$ in Lemma A.5, which makes $P_g = 0$).

A.5.1 One-Sided Gaussian Conditioning

This subsection provides a proof of Lemma 2.26.

Proof of Lemma 2.26. Write the rows of W as $w_1, \dots, w_n \in \mathbb{R}^{1\times p}$. Since the rows are independent and identically distributed, it suffices to prove the statement for a single row $w \sim \mathcal{N}(0, I_p)$.

Let $\varphi \in \mathbb{R}^{p\times q}$ be deterministic and define the orthogonal projector $P_\varphi = \varphi(\varphi^\top\varphi)^\dagger\varphi^\top$ onto $\mathrm{col}(\varphi)$, with $P_\varphi^\perp = I_p - P_\varphi$. Decompose

$$w = wP_\varphi + wP_\varphi^\perp. \tag{A.21}$$

The random vector wP_φ is measurable with respect to $w\varphi$ because P_φ depends on φ only and wP_φ is the (unique minimum-norm) linear function of $w\varphi$ matching the constraint in least squares. Moreover, $wP_\varphi^\perp$ is a centered Gaussian with covariance $P_\varphi^\perp$, and it is independent of $w\varphi$ because

$$\mathrm{Cov}(wP_\varphi^\perp,\ w\varphi) = P_\varphi^\perp\varphi = 0, \tag{A.22}$$

so joint Gaussianity implies independence.

Let $\widetilde{w} \sim \mathcal{N}(0, I_p)$ be an independent copy of w. Then $\widetilde{w}P_\varphi^\perp \sim \mathcal{N}(0, P_\varphi^\perp)$ and is independent of $w\varphi$. Therefore,

$$w \,|\, \sigma(w\varphi) \ \overset{d}{=}\ wP_\varphi + \widetilde{w}P_\varphi^\perp. \tag{A.23}$$

Stacking the rows back together yields $W \,|\, \sigma(W\varphi) \ \overset{d}{=}\ WP_\varphi + \widetilde{W}P_\varphi^\perp$ with $\widetilde{W}$ an independent copy of W, as claimed. □

A.5.2 Two-Sided Gaussian Conditioning

Lemma A.5 (Gaussian Conditioning (Two-Sided Projection Decomposition))**.** *Let $W \in \mathbb{R}^{n\times n}$ have iid $\mathcal{N}(0,1)$ entries and let $\varphi, g \in \mathbb{R}^{n\times m}$ be deterministic. Define the orthogonal projections onto the column spans*

$$P_\varphi = \varphi(\varphi^\top\varphi)^\dagger\varphi^\top, \qquad P_g = g(g^\top g)^\dagger g^\top, \tag{A.24}$$

where $(\cdot)^\dagger$ is the Moore–Penrose pseudo-inverse. Then

$$W \mid (W\varphi,\ W^\top g) \ \overset{d}{=}\ P_g W + W P_\varphi - P_g W P_\varphi + P_g^\perp \widetilde{W} P_\varphi^\perp, \tag{A.25}$$

where $P^\perp = I_n - P$ and $\widetilde{W}$ is an independent copy of W.

This subsection proves Lemma A.5. The statement is a convenient matrix form of linear-Gaussian regression: conditioning a Gaussian matrix W on the linear observations $W\varphi$ and $W^\top g$ amounts to removing the components of W that are "seen" along the spans of φ and g, and leaving an independent Gaussian on the orthogonal complement.

Proof of Lemma A.5. Write the Frobenius inner product as $\langle A, B\rangle_F := \mathrm{Tr}(A^\top B)$. Let $\mathcal{U} := \mathrm{col}(\varphi) \subset \mathbb{R}^n$ and $\mathcal{V} := \mathrm{col}(g) \subset \mathbb{R}^n$ and denote the corresponding orthogonal projections by P_φ and P_g.

Decompose W into four blocks with respect to the orthogonal splittings $\mathbb{R}^n = \mathcal{V} \oplus \mathcal{V}^\perp$ (rows) and $\mathbb{R}^n = \mathcal{U} \oplus \mathcal{U}^\perp$ (columns):

$$W = \underbrace{P_g W P_\varphi}_{\text{seen by both}} + \underbrace{P_g W P_\varphi^\perp}_{\text{seen by } W^\top g} + \underbrace{P_g^\perp W P_\varphi}_{\text{seen by } W\varphi} + \underbrace{P_g^\perp W P_\varphi^\perp}_{\text{unseen}}. \tag{A.26}$$

The key observations are:

- $W\varphi$ depends only on the columns of W in $\mathcal{U}$, i.e. on WP_φ (equivalently the sum of the first and third blocks);
- $W^\top g$ depends only on the rows of W in $\mathcal{V}$, i.e. on $P_g W$ (equivalently the sum of the first and second blocks);
- the "unseen" block $P_g^\perp W P_\varphi^\perp$ is independent of $(WP_\varphi,\ P_g W)$ because W has iid Gaussian entries and these are orthogonal linear functionals of $\mathrm{vec}(W)$.

More formally, $\mathrm{vec}(W) \sim \mathcal{N}(0, I_{n^2})$ and the maps

$$L_1 : W \mapsto WP_\varphi, \qquad L_2 : W \mapsto P_g W, \qquad L_3 : W \mapsto P_g^\perp W P_\varphi^\perp \tag{A.27}$$

are linear. One checks that L_3 is orthogonal (as a linear map on $\mathbb{R}^{n\times n}$ with Frobenius inner product) to both L_1 and L_2: for any matrices A, B,

$$\langle P_g^\perp A P_\varphi^\perp,\ BP_\varphi\rangle_F = \mathrm{Tr}\big((P_g^\perp A P_\varphi^\perp)^\top B P_\varphi\big) = 0, \qquad \langle P_g^\perp A P_\varphi^\perp,\ P_g B\rangle_F = 0, \tag{A.28}$$

since $P_\varphi^\perp P_\varphi = 0$ and $P_g P_g^\perp = 0$. Therefore $L_3(W)$ is independent of $(L_1(W), L_2(W))$, and its conditional law given $(W\varphi, W^\top g)$ is unchanged.

It remains to identify the conditional mean. Given $W\varphi$ and $W^\top g$, the blocks $P_g^\perp W P_\varphi$ and $P_g W P_\varphi^\perp$ are fixed by the constraints, and the intersection block $P_g W P_\varphi$ is fixed twice but consistently. A short algebraic rearrangement gives

$$P_g W + W P_\varphi - P_g W P_\varphi = P_g W P_\varphi + P_g W P_\varphi^\perp + P_g^\perp W P_\varphi, \tag{A.29}$$

which is exactly the sum of the blocks determined by $(W^\top g, W\varphi)$. Thus the conditional law is "deterministic part + independent unseen Gaussian block", i.e.

$$W \mid (W\varphi, W^\top g) \stackrel{d}{=} (P_g W + W P_\varphi - P_g W P_\varphi) + P_g^\perp \widetilde{W} P_\varphi^\perp, \tag{A.30}$$

where $\widetilde{W}$ is an independent copy of W. □

Acknowledgements

I gratefully acknowledge the contributions of the student scribes—Sidharth Bajaj, Jonathan Borenstein, Edward Chang, Majid Ghasemi, Disen Liao, Ruihan Lin, Xiaoxi Luo, Marty Mukherjee, Lucas Noritomi-Hartwig, Gustavo Sutter, Kevin Zhao, Daniel Zhang, and Randolph (Yifei) Zhang—whose careful scribing and clarifications made these lecture notes possible. I also acknowledge GPT Pro and, later, Sol Ultra, in assisting me in compiling, organizing, and editing the scribed and handwritten materials.

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization, 2018.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel, Switzerland, 2005.

[3] Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024.

[4] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[5] Hari Bercovici and Dan-Virgil Voiculescu. Lévy-hinčin type theorems for multiplicative and additive free convolution. *Pacific journal of mathematics*, 153(2):217–248, 1992.

[6] Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pages 32240–32256, 2022.

[7] Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. In *Advances in Neural Information Processing Systems*, 2023.

[8] Guillaume Chaintron and Alexandre Diez. Propagation of chaos: A review of models, methods and applications. *Kinetic and Related Models*, 15(6):895–1012, 2022.

[9] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, pages 6572–6583, 2018.

[10] Lenaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[11] Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, volume 22, pages 342–350, 2009.

[12] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2793–2803. PMLR, 2021.

[13] Monroe D. Donsker. An invariance principle for certain probability limit theorems. In *Four Papers on Probability*, number 6 in Memoirs of the American Mathematical Society, pages 1–12. American Mathematical Society, Providence, RI, 1951.

[14] Simon Shaolei Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks, 2018.

[15] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, 2009.

[16] Margalit Glasgow, Denny Wu, and Joan Bruna. Propagation of chaos in one-hidden-layer neural networks beyond logarithmic time, 2025.

[17] Moritz Haas, Sebastian Bordt, Ulrike von Luxburg, and Leena Chennuru Vankadara. On the surprising effectiveness of large learning rates under standard width scaling, 2025.

[18] Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations (ICLR)*, 2020.

[19] Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020.

[20] Soufiane Hayou and Greg Yang. Width and depth limits commute in residual networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 12700–12723. PMLR, 2023.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[23] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically, 2017.

[24] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. Also appeared at NeurIPS 2022 (Chinchilla).

[25] Pierre-Emmanuel Jabin and Zhenfu Wang. Quantitative estimates of propagation of chaos for stochastic systems with lipschitz interactions. *The Annals of Probability*, 46(3):1546–1586, 2018.

[26] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pages 8571–8580, 2018.

[27] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[28] Olav Kallenberg. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, Cham, third edition, 2021.

[29] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[30] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1711.00165.

[31] Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[32] Mufan Li, Jaume de Dios Pont, Mihai Nica, and Daniel M. Roy. Geometric Dyson Brownian motions and the Free Log-Normal limit for a non-square Gaussian matrix product, 2026. arXiv:2606.30831.

[33] Mufan Li, Lorenzo Noci, and Boris Hanin. Feature learning in the proportional limit. In preparation, 2026.

[34] Mufan B. Li, Mihai Nica, and Daniel M. Roy. The neural covariance SDE: Shaped infinite depth-and-width networks at initialization. In *Advances in Neural Information Processing Systems*, volume 35, pages 10795–10808, 2022.

[35] James Martens, Andrew Ballard, Guillaume Desjardins, Grégoire Swirszcz, Valentin Dalibard, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Rapid training of deep neural networks without skip connections or normalization layers using deep kernel shaping, 2021.

[36] Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1804.11271.

[37] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[38] Radford M. Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer, New York, NY, USA, 1996.

[39] Quynh Nguyen, Marco Mondelli, and Guillermo Montúfar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep ReLU networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.

[40] Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles, 2017.

[41] Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *Advances in Neural Information Processing Systems*, volume 35, pages 27198–27211, 2022.

[42] Lorenzo Noci, Chuning Li, Mufan B. Li, Bobby He, Thomas Hofmann, Chris J. Maddison, and Daniel M. Roy. The Shaped Transformer: Attention models in the infinite depth-and-width limit. In *Advances in Neural Information Processing Systems*, volume 36, pages 54250–54281, 2023.

[43] Valentin V. Petrov. *Sums of Independent Random Variables*, volume 82 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag, Berlin and Heidelberg, 1975.

[44] Marc Potters and Jean-Philippe Bouchaud. *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists*. Cambridge University Press, Cambridge, UK, 2020.

[45] Grant M. Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error, 2018.

[46] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[47] Daniel W Stroock and SR Srinivasa Varadhan. *Multidimensional diffusion processes*, volume 233. Springer Science & Business Media, 1997.

[48] Alain-Sol Sznitman. Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX—1989*, volume 1464 of *Lecture Notes in Mathematics*, pages 165–251. Springer, Berlin, Heidelberg, 1991.

[49] Eugene P. Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67(2):325–327, 1958.

[50] Greg Yang and Edward Hu. Tensor programs IV: Feature learning in infinite-width neural networks, 2021.

[51] Greg Yang, Edward Hu, Igor Babuschkin, Saurabh Garg, Hongyu Zhang, Jeffrey Dean, and Noam Shazeer. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. In *International Conference on Learning Representations (ICLR)*, 2022.

[52] Guodong Zhang, Aleksandar Botev, and James Martens. Deep learning without shortcuts: Shaping the kernel with tailored rectifiers. In *International Conference on Learning Representations (ICLR)*, 2022.

[53] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep ReLU networks, 2018.