Lecture Notes on Infinite-Width Limits of Neural Networks

Cengiz Pehlevan and Blake Bordelon

June 2023

Contents

1 Introduction

Deep learning has indelibly shaped many fields of study and application, driving innovation and pushing the boundaries of what is technologically feasible. Nevertheless, it continues to be viewed as a "black box"—a complex system whose internal workings are largely inscrutable. The desire to peek inside this box has led to an urgent need for a theory of deep learning.

Such a theory, ideally, would exhibit certain critical features. It should be analytically tractable and/or efficiently computable, lending itself to straightforward mathematical treatment or rapid

computational implementation. In addition, it must be mechanistically interpretable, revealing the inner workings of deep learning systems. Moreover, this theory should have strong predictive power, accurately forecasting outcomes in scenarios relevant to practical applications. Lastly, it should provide answers to questions that emerge within various domains, like physics or neuroscience, that have leveraged the power of deep learning.

In the pursuit of constructing this theory, it might be beneficial to draw upon insights from the field of physics—specifically, the physics of disordered systems and statistical mechanics. These areas offer key conceptual tools, such as the notion of a thermodynamic limit where system-size-dependent fluctuations become negligible. This limit allows for the definition of holistic system descriptors that can unveil intriguing aspects of system behavior, including emergent phenomena.

In the context of deep learning, we propose to apply these concepts to the gradient-descent training of neural networks. Our focus is on exploring the "infinite-width" limit, where "width" refers to the number of units in a single layer of a neural network.

We will show that the limit can be taken in many different ways, and the particular way the limit is taken can result in different neural network behavior. We will construct theoretical descriptions of the emergent behavior in these limits.

We will present compelling evidence to suggest that a specific way to take this infinite-width limit offers a robust description of neural network behavior that closely parallels the performance of practical neural networks. Here,"practical" implies networks that perform at a state-of-the-art level and can be trained using reasonable computational resources. Thus, by uncovering these new operational paradigms, we hope to unlock a deeper understanding of these complex systems and push the frontiers of their capabilities.

This set of lectures was prepared for the Princeton Machine Learning Summer School that was held in June 2023. It is accompanied by a set of slides posted on the School website. We focus on the results of [1, 2, 3] and present them in a more pedagogical manner to an audience of graduate students in machine learning.

2 Setting and Notation

In this set of lecture notes we will be concerned with supervised training of multilayer perceptrons (MLP) with (full batch) gradient flow. Extensions including convolutional architectures and gradient descent[1] and [4].

We will consider the learning dynamics of a feedforward MLP of L layers and a scalar output. Given an input $\boldsymbol{x}_\mu \in \mathbb{R}^D$, where $\mu = 1, \ldots, P$, we define the hidden *pre-activation* vectors $\boldsymbol{h}^{(\ell)} \in \mathbb{R}^N$ for layers $\ell \in \{1, ..., L\}$ as

$$f(\boldsymbol{x}_\mu) = \frac{1}{\gamma_0 N} \boldsymbol{w}^{(L)} \cdot \phi(\boldsymbol{h}_\mu^{(L-1)}) \ , \quad \boldsymbol{h}_\mu^{(\ell+1)} = \frac{1}{\sqrt{N}} \boldsymbol{W}^{(\ell+1)} \phi(\boldsymbol{h}_\mu^{(\ell)}) \ , \quad \boldsymbol{h}_\mu^{(1)} = \frac{1}{\sqrt{D}} \boldsymbol{W}^{(1)} \boldsymbol{x}_\mu, \qquad (1)$$

where

$$\boldsymbol{\theta} \equiv \mathrm{Vec}\{\boldsymbol{W}^{(1)}, ..., \boldsymbol{w}^{(L)}\}$$

are the trainable parameters of the network and ϕ is a twice differentiable activation function. We

will sometimes use $\boldsymbol{W}^{(L)} \equiv \boldsymbol{w}^{(L)}$ for notational convenience. The role of the parameter γ_0 will become apparent later. Note that we omitted the bias terms here for notational simplicity. The network parameters $\boldsymbol{\theta}$ will be initialized form a Gaussian distribution with zero mean and unit variance, $\boldsymbol{\theta}_i \sim \mathcal{N}(0,1)$. We explore different scalings in later section.

Given a training set

$$\mathcal{D} \equiv \{(\boldsymbol{x}_\mu, y_\mu)\}_{\mu=1}^{P},$$

we train the network on a loss

$$\mathcal{L}(\mathcal{D};\boldsymbol{\theta}) \equiv \frac{1}{P}\sum_{\mu=1}^{P} l\left(f(\boldsymbol{x}_\mu;\boldsymbol{\theta}), y_\mu\right) = \frac{1}{P}\sum_{\mu=1}^{P} l_\mu \tag{2}$$

with full-batch gradient flow

$$\frac{d\boldsymbol{\theta}}{dt} = -\eta\frac{1}{P}\sum_{\mu=1}^{P}\frac{\partial l_\mu}{\partial \boldsymbol{\theta}} = \eta\frac{1}{P}\sum_{\mu=1}^{P}\Delta_\mu\frac{\partial f(\boldsymbol{x}_\mu;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{3}$$

where η is a learning rate which scales as

$$\eta = \eta_0\gamma_0^2 N, \tag{4}$$

and we defined

$$\Delta_\mu \equiv -\frac{\partial l\left(f(\boldsymbol{x}_\mu;\boldsymbol{\theta}), \boldsymbol{y}_\mu\right)}{\partial f(\boldsymbol{x}_\mu;\boldsymbol{\theta})}. \tag{5}$$

We will consider various limits in our discussion; certain quantities will be taken to zero or infinity. We will use the notation $q \sim \mathcal{O}(N^m)$ to denote the fact that $\lim_{N\to\infty}\frac{q}{N^m} < \infty$.

3 Numerical evidence that the limit is descriptive of practical networks

A recent preprint by our group [3] provides numerical experiments that provides evidence that networks of practical size in the initialization we are studying are already operating close to their limiting behavior with respect to width. For our purposes, by practical we mean that the network is not so wide as to be infeasible to train on a single 80 GB GPU and can perform well on the tasks of interest.

4 Why this scaling of parameters?

In Eq. 1, we made certain choices about how quantities behave as $N \to \infty$. We scaled weights by either $1/N$ or $1/\sqrt{N}$. Why is this particular choice?

4.1 General Scaling

In a series of work, Yang and colleagues [5] provided conditions for how feature-learning networks may arise in the infinite-width limit. Here, we provide an alternative version of that argument, modified from the Appendices of [1].

Infinite-width limit is defined by taking $N \to \infty$. We have to decide how various parameters of the network scale in this limit. Let's start with the most general case. We will assume the following parameterization and initialization

$$f_\mu = \frac{1}{\gamma} h_\mu^{(L)}, \tag{6}$$

$$h_\mu^{(L)} = \frac{1}{N^{a_L}} \boldsymbol{w}^{(L)} \cdot \phi(\boldsymbol{h}_\mu^{(L-1)}) \ , \qquad w_i^{(L)} \sim \mathcal{N}\left(0, \frac{1}{N^{b_L}}\right), \tag{7}$$

$$\boldsymbol{h}_\mu^{(\ell)} = \frac{1}{N^{a_\ell}} \boldsymbol{W}^{(\ell)} \phi(\boldsymbol{h}_\mu^{(\ell-1)}) \ , \qquad W_{ij}^{(\ell)} \sim \mathcal{N}\left(0, \frac{1}{N^{b_\ell}}\right), \tag{8}$$

$$\boldsymbol{h}_\mu^{(1)} = \frac{1}{N^{a_1}} \frac{1}{D^{1/2}} \boldsymbol{W}^{(1)} \boldsymbol{x}_\mu \ , \qquad W_{ij}^{(1)} \sim \mathcal{N}\left(0, \frac{1}{N^{b_1}}\right), \tag{9}$$

and we consider training with gradient flow dynamics

$$\frac{d\boldsymbol{\theta}}{dt} = \eta \frac{1}{P} \sum_{\mu=1}^{P} \Delta_\mu \frac{\partial f_\mu}{\partial \boldsymbol{\theta}}. \tag{10}$$

The learning rate is scaled as

$$\eta = \eta_0 \gamma^2 N^{-c} \tag{11}$$

with $\eta_0 = \mathcal{O}(1)$. The reason for the factor of γ^2 in the learning rate η will be clear later. Lastly, we will scale the feature learning parameter γ as

$$\gamma = \gamma_0 N^d, \tag{12}$$

with $\gamma_0 = \mathcal{O}(1)$. This role of this parameter in switching between lazy and rich training regimes was already emphasized before by other authors [6, 7]. We will find that only $d = \frac{1}{2}$ will allow feature learning. The reason for the slightly odd parameterization of the learning rate and the feature learning parameter will be apparent below.

We will now derive constraints on (a, b, c, d) which give desired large width behavior. We require that in the infinite-width limit have 1. finite preactivations, 2. learning in finite time, 3. feature learning in finite time?.

A note about P: we will keep it $\mathcal{O}(1)$. We will point to where scaling it with N might change our results.

4.1.1 Preactivations are $\mathcal{O}_N(1)$ at initialization

In this section, we identify conditions under which $\boldsymbol{h}^{(\ell)}$ have $\mathcal{O}_N(1)$ entries. Precisely, we will demand the mean-squared values of the entries of $\boldsymbol{h}^{\ell}$ to be $\mathcal{O}_N(1)$, where the averages are taken across random network initializations. This ensures that the network is not diverging or saturated, or have zero output.

We denote averages with respect to initialization weights by

$$\langle q(\boldsymbol{\theta}) \rangle \equiv \mathbb{E}_{\boldsymbol{\theta}} \left[q(\boldsymbol{\theta}) \right]. \tag{13}$$

Later on, when we construct the infinite-width limit through a dynamical mean field theory, we will see that these averages can also be interpreted as averages across units in a network layer.

We first establish the base case for $\boldsymbol{h}^{(1)}$. Note that $\boldsymbol{h}^{(1)}$ is a multivariate Gaussian since it is given by a summation of Gaussian random variables (see (9)). Therefore, its statistics are fully defined by its first two moments:

$$\left\langle h^{(1)}_{\mu,i} \right\rangle = \frac{1}{N^{a_1}} \frac{1}{\sqrt{D}} \sum_k \left\langle W^{(1)}_{ik}(0) \right\rangle x_{\mu,k} = 0. \tag{14}$$

and

$$\left\langle h^{(1)}_{\mu,i} h^{(1)}_{\nu,j} \right\rangle = \frac{1}{N^{2a_1}} \frac{1}{D} \sum_{kk'} \left\langle W^{(1)}_{ik}(0) W^{(1)}_{jk'}(0) \right\rangle x_{\mu,k} x_{\nu,k'} = \delta_{ij} \frac{1}{N^{2a_1+b_1}} \frac{1}{D} \sum_{k=1}^{D} x_{\mu,k} x_{\nu,k}. \tag{15}$$

Assuming that the input norm does not scale with N as $N \to \infty$, we find the constraint that

$$\boxed{2a_1 + b_1 = 0,} \tag{16}$$

for the mean-squared values of preactivations to be $\mathcal{O}_N(1)$.

Now that we have a condition for $\boldsymbol{h}^{(1)}$ to be $\mathcal{O}_N(1)$ in its entries, we proceed to next layer. Note the mean of the entries of $\boldsymbol{h}^{(2)}$ are still zero:

$$\left\langle h^{(2)}_{\mu,i} \right\rangle = \frac{1}{N^{a_2}} \sum_k \left\langle W^{(2)}_{ik}(0) \phi(h^{(1)}_{\mu,k}) \right\rangle = \frac{1}{N^{a_2}} \sum_k \left\langle W^{(2)}_{ik}(0) \right\rangle \left\langle \phi(h^{(1)}_{\mu,k}) \right\rangle = 0. \tag{17}$$

Here, we used the independence of weights at initialization in each layer. The covariance is then

$$\begin{aligned}
\left\langle h^{(2)}_{\mu,i} h^{(2)}_{\nu,j} \right\rangle &= \frac{1}{N^{2a_2}} \sum_{k,k'} \left\langle W^{(2)}_{ik}(0) W^{(2)}_{jk'}(0) \phi(h^{(1)}_{\mu,k}) \phi(h^{(1)}_{\nu,k'}) \right\rangle \\
&= \frac{1}{N^{2a_2}} \sum_{k,k'} \left\langle W^{(2)}_{ik}(0) W^{(2)}_{jk'}(0) \right\rangle \left\langle \phi(h^{(1)}_{\mu,k}) \phi(h^{(1)}_{\nu,k'}) \right\rangle \\
&= \delta_{ij} \frac{1}{N^{2a_2+b_2-1}} \frac{1}{N} \sum_{k=1}^{N} \left\langle \phi(h^{(1)}_{\mu,k}) \phi(h^{(1)}_{\nu,k}) \right\rangle = \delta_{ij} \frac{1}{N^{2a_2+b_2-1}} \left\langle \phi(h^{(1)}_{\mu,1}) \phi(h^{(1)}_{\nu,1}) \right\rangle.
\end{aligned} \tag{18}$$

In the second line, we used the independence of weights at initialization in each layer. In the

third line, we used that $h^{(1)}_{\mu,k}$ are identical in distribution. We demand the mild condition that ϕ is square-integrable with respect to the measure of h^1_k, and is $\mathcal{O}_N(1)$. Altogether, these arguments imply that

$$2a_2 + b_2 = 1. \tag{19}$$

It is easy to see that this argument can be iterated for all layers. Hence, we identify the set of constraints

$$\boxed{2a_\ell + b_\ell = 1, \quad l = 2, \ldots, L.} \tag{20}$$

While we are discussing preactivations, we want to point to a key property of them. We want to also define and discuss the following quantity:

$$\Phi^{(\ell)}_{\mu\nu} \equiv \frac{1}{N}\phi(\boldsymbol{h}^{(\ell)}_\mu) \cdot \phi(\boldsymbol{h}^{(\ell)}_\nu), \quad l = 1, \ldots, L-1 \tag{21}$$

We will call these quantities *feature kernels*.

Proposition 4.1. *Given conditions* (16) *and* (20)*, at initialization, in the infinite-width limit, the feature kernels asymptote to deterministic objects. Further,* $h^{(\ell)}_{i,\mu}$ *are Gaussian distributed:*

$$h^{(\ell)}_{i,\mu} \sim \mathcal{N}\left(0, \Phi^{(\ell-1)}_{\mu\nu}\delta_{ij}\right) \tag{22}$$

and

$$\Phi^{(\ell)}_{\mu\nu} = \mathbb{E}_{h^{(l)}_\mu \sim \mathcal{N}\left(0, \Phi^{(\ell-1)}_{\mu\nu}\right)}\left[\phi(h^{(l)}_\mu)\phi(h^{(l)}_\nu)\right] \tag{23}$$

with initial condition

$$\Phi^{(0)}_{\mu\nu} = \frac{1}{D}\boldsymbol{x}_\mu \cdot \boldsymbol{x}_\nu. \tag{24}$$

Remark 4.1. *The preactivations are Gaussian processes. This result is a slight generalization of the existing work on Neural Network Gaussian Processes [8, 9, 10].*

Proof. A modification of the argument given in `https://jzv.io/assets/pdf/lecture_notes_on_nngp_from_mft.pdf` and [11] can be used to prove this. See Appendix A. The proof introduces techniques which will be used again later. For readers unfamiliar with the techniques commonly employed in statistical mechanics, we strongly recommend a review of the appendix before studying the dynamical mean field theory construction. □

Remark 4.2. *We note that in this limit the neurons in a particular layer* ℓ *are identically and independently distributed.*

4.1.2 Predictions Evolve in $\mathcal{O}_N(1)$ time

Next, we demand that predictions evolve in $\mathcal{O}_N(1)$ time under gradient-flow. This ensures that the network learns in finite time in the infinite-width limit.

By chain rule, the predictions evolve under

$$\frac{df(\boldsymbol{x}_\mu;\boldsymbol{\theta})}{dt} = \frac{\partial f_\mu}{\partial \boldsymbol{\theta}} \cdot \frac{d\boldsymbol{\theta}}{dt} = -\eta \frac{1}{P} \sum_{\nu=1}^{P} \frac{\partial f_\mu}{\partial \boldsymbol{\theta}} \cdot \frac{\partial l_\nu}{\partial \boldsymbol{\theta}} = \eta \frac{1}{P} \sum_{\nu=1}^{P} \frac{\partial f_\mu}{\partial \boldsymbol{\theta}} \cdot \frac{\partial f_\nu}{\partial \boldsymbol{\theta}} \Delta_\nu \tag{25}$$

For the network prediction evolution to be $\mathcal{O}_N(1)$, we demand $\partial_t f_\mu \sim \mathcal{O}_N(1)$. This requires the following quantity to be $\mathcal{O}_N(1)$:

$$\begin{aligned}
&\frac{\gamma^2}{N^c} \frac{\partial f_\mu}{\partial \boldsymbol{\theta}} \cdot \frac{\partial f_\nu}{\partial \boldsymbol{\theta}} = \frac{1}{N^c} \frac{\partial h_\mu^{(L)}}{\partial \boldsymbol{\theta}} \cdot \frac{\partial h_\nu^{(L)}}{\partial \boldsymbol{\theta}} = \frac{1}{N^c} \sum_{\ell=1}^{L} \sum_{i,j} \frac{\partial h_\mu^{(L)}}{\partial W_{ij}^{(\ell)}} \frac{\partial h_\nu^{(L)}}{\partial W_{ij}^{(\ell)}} \\
&= \frac{1}{N^c} \sum_{\ell=1}^{L} \sum_{i,j} \sum_{m,n} \frac{\partial h_\mu^{(L)}}{\partial h_{\mu,m}^{(\ell)}} \frac{\partial h_{\mu,m}^{(\ell)}}{\partial W_{ij}^{(\ell)}} \frac{\partial h_\nu^{(L)}}{\partial h_{\nu,n}^{(\ell)}} \frac{\partial h_{\nu,n}^{(\ell)}}{\partial W_{ij}^{(\ell)}} \\
&= \frac{1}{N^c} \sum_{\ell=2}^{L} \sum_{i,j} \sum_{m,n} \frac{\partial h_\mu^{(L)}}{\partial h_{\mu,m}^{(\ell)}} \frac{\partial h_\nu^{(L)}}{\partial h_{\nu,n}^{(\ell)}} \frac{1}{N^{2a_\ell}} \phi(h_{\mu,j}^{(\ell-1)}) \delta_{im} \phi(h_{\nu,j}^{(\ell-1)}) \delta_{in} + \frac{\partial h_\mu^{(L)}}{\partial h_{\mu,m}^{(1)}} \frac{\partial h_\nu^{(L)}}{\partial h_{\nu,n}^{(1)}} \frac{1}{N^{2a_1} D} x_{\mu,j} \delta_{im} x_{\nu,j} \delta_{in} \\
&= \frac{1}{N^c} \left[\frac{\phi(\boldsymbol{h}_\mu^{(L-1)}) \cdot \phi(\boldsymbol{h}_\nu^{(L-1)})}{N^{2a_L}} + \sum_{\ell=2}^{L-1} \frac{\partial h_\mu^{(L)}}{\partial \boldsymbol{h}_\mu^{(\ell)}} \cdot \frac{\partial h_\nu^{(L)}}{\partial \boldsymbol{h}_\nu^{(\ell)}} \frac{\phi(\boldsymbol{h}_\mu^{(\ell-1)}) \cdot \phi(\boldsymbol{h}_\nu^{(\ell-1)})}{N^{2a_\ell}} + \frac{\partial h_\mu^{(L)}}{\partial \boldsymbol{h}_\mu^{(1)}} \cdot \frac{\partial h_\nu^{(L)}}{\partial \boldsymbol{h}_\nu^{(1)}} \frac{\boldsymbol{x}_\mu \cdot \boldsymbol{x}_\nu}{N^{2a_1} D} \right] \\
&= \frac{1}{N^c} \left[\frac{1}{N^{2a_L - 1}} \Phi_{\mu\nu}^{(L-1)} + \sum_{\ell=2}^{L-1} \frac{1}{N^{2a_\ell - 1}} G_{\mu\nu}^{(\ell)} \Phi_{\mu\nu}^{(\ell-1)} + \frac{1}{N^{2a_1}} G_{\mu\nu}^{(1)} \Phi_{\mu\nu}^{(0)} \right],
\end{aligned} \tag{26}$$

where defined

$$G_{\mu\nu}^{(\ell)} \equiv \frac{1}{N} \boldsymbol{g}_\mu^{(\ell)} \cdot \boldsymbol{g}_\nu^{(\ell)}, \qquad \boldsymbol{g}_\mu^{(\ell)} \equiv \sqrt{N} \frac{dh_\mu^{(L)}}{d\boldsymbol{h}^{(\ell)}}. \tag{27}$$

$\Phi_{\mu\nu}^{(L)}$ concentrate and are $\mathcal{O}_N(1)$ under the assumptions of the previous section. The same can be said about $G_{\mu\nu}^{(\ell)}$. To see this, we start with the last layer and define with a more general scaling

$$g_{\mu,i}^{(L-1)} = N^{a_L + b_L/2} \frac{\partial h_\mu^{(L)}}{\partial h_{\mu,i}^{(L-1)}} = N^{b_L/2} w_i^{(L)} \dot{\phi}(h_i^{(L-1)}). \tag{28}$$

Then,

$$\begin{aligned}
\left\langle g_{\mu,i}^{(L-1)} \right\rangle &= N^{b_L/2} \left\langle w_i^{(L)} \right\rangle \left\langle \dot{\phi}(h_i^{(L-1)}) \right\rangle = 0, \\
\left\langle \left(g_{\mu,i}^{(L-1)} \right)^2 \right\rangle &= N^{b_L/2} \left\langle \left(w_i^{(L)} \right)^2 \right\rangle \left\langle \left(\dot{\phi}(h_i^{(L-1)}) \right)^2 \right\rangle = \left\langle \left(\dot{\phi}(h_i^{(L-1)}) \right)^2 \right\rangle = \mathcal{O}_N(1).
\end{aligned} \tag{29}$$

We can similarly extend this definition to earlier layers $\boldsymbol{g}^\ell = N^{a_L + b_L/2} \frac{\partial h^{L+1}}{\partial \boldsymbol{h}^\ell}$ to see whether $\boldsymbol{g}^\ell$

remains $\mathcal{O}_N(1)$ under its backward-pass recursion

$$\boldsymbol{g}^\ell = \left(\frac{\partial \boldsymbol{h}^{\ell+1}}{\partial \boldsymbol{h}^\ell}\right)^\top \boldsymbol{g}^{\ell+1} = \dot{\phi}(\boldsymbol{h}^\ell) \odot \left[N^{-a_\ell} \boldsymbol{W}^\ell(0)^\top \boldsymbol{g}^\ell\right] \tag{30}$$

Now, letting $\boldsymbol{z}^\ell = N^{-a_\ell}\boldsymbol{W}^\ell(0)^\top \boldsymbol{g}^{\ell+1}$ as in the main text, we have that $\boldsymbol{z}^\ell|\{\boldsymbol{g}^{\ell+1}\}$ is Gaussian with covariance

$$\left\langle z_i^\ell z_j^\ell \right\rangle = \delta_{ij} N^{-2a_\ell - b_\ell} \boldsymbol{g}^{\ell+1} \cdot \boldsymbol{g}^{\ell+1} = \delta_{ij} N^{-2a_\ell - b_\ell + 1} G^{\ell+1}. \tag{31}$$

Under the inductive hypothesis that $G^{\ell+1} \sim \mathcal{O}_N(1)$ and the previous constraint $2a_\ell + b_\ell = 1$, the z variables have $\mathcal{O}_N(1)$ variance. Overall, we can thus ensure that $\Phi^\ell, G^\ell \sim \mathcal{O}_N(1)$ if $2a_\ell + b_\ell = 1$ for $\ell \in \{1, ..., L\}$ and $2a_0 + b_0 = 0$.

We thus find the following constraints

$$\begin{aligned} &\boxed{2a_\ell + c = 1\ , \quad \ell \in \{2, ..., L\}} \\ &\boxed{2a_1 + c = 0}. \end{aligned} \tag{32}$$

Again this is consistent with the mean-field parameterization in the previous section provided $c = 0$ and $a_1 = 0$ and $a_\ell = \frac{1}{2}$, $\ell = 2, \ldots, L$. We see that for non-zero c, we need non-zero a_1.

4.1.3 Features evolve in $\mathcal{O}_N(1)$ time

Now, we desire that preactivations all evolve by an $\mathcal{O}_N(1)$ amount during network training, which we chose to define as stable feature learning. This requires $\frac{d\boldsymbol{h}_\mu^{(\ell)}}{dt} = \mathcal{O}_N(1)$. To see the scaling this requires, let's start with looking at first layer:

$$\begin{aligned} \frac{d\boldsymbol{h}_\mu^{(1)}}{dt} &= \frac{1}{N^{a_1}} \frac{d\boldsymbol{W}^{(1)}}{dt} \frac{1}{\sqrt{D}} \boldsymbol{x}_\mu = \frac{1}{D} \frac{1}{N^{2a_1}} \frac{\eta_0 \gamma^2}{N^c} \frac{1}{P} \sum_{\nu=1}^P \frac{\Delta_\nu}{\gamma} \frac{\partial h_\nu^{(L)}}{\partial \boldsymbol{h}_\nu^{(1)}} \boldsymbol{x}_\nu \cdot \boldsymbol{x}_\mu \\ &= \frac{1}{N^{2a_1 + c - d + 1/2}} \eta_0 \gamma_0 \frac{1}{P} \sum_{\nu=1}^P \Delta_\nu \boldsymbol{g}_\nu^{(1)} \Phi_{\mu\nu}^{(0)}. \end{aligned} \tag{33}$$

Except the coefficient $N^{-2a_1 - c + d - 1/2}$, everything else is $\mathcal{O}_N(1)$. Given (32), this leads to

$$2a_1 + c - d + 1/2 = 0 \tag{34}$$

Given (32), this leads to

$$d = \frac{1}{2} \tag{35}$$

Now, we can go to the next layers:

$$\frac{d\boldsymbol{h}_\mu^{(\ell)}}{dt} = \frac{1}{N^{a_\ell}}\frac{d\boldsymbol{W}^{(\ell)}}{dt}\phi(\boldsymbol{h}_\mu^{(\ell-1)}) + \frac{1}{N^{a_\ell}}\boldsymbol{W}^{(\ell)}\dot{\phi}(\boldsymbol{h}_\mu^{(\ell-1)}) \odot \frac{d\boldsymbol{h}_\mu^{(\ell-1)}}{dt}$$

$$= \frac{1}{N^{2a_\ell}}\frac{\eta_0\gamma^2}{N^c P}\sum_{\nu=1}^{P}\frac{\Delta_\nu}{\gamma}\frac{\partial h_\nu^L}{\partial \boldsymbol{h}_\nu^{(\ell)}}\phi(\boldsymbol{h}_\nu^{(\ell-1)})\cdot\phi(\boldsymbol{h}_\mu^{(\ell-1)}) \tag{36}$$

$$= \frac{1}{N^{2a_\ell+c-d-1/2}}\eta_0\gamma_0\frac{1}{P}\sum_{\nu=1}^{P}\Delta_\nu \boldsymbol{g}_\nu^{(\ell)}\Phi_{\mu\nu}^{(\ell-1)} + \frac{1}{N^{a_\ell}}\boldsymbol{W}^{(\ell)}\dot{\phi}(\boldsymbol{h}_\mu^{(\ell-1)}) \odot \frac{d\boldsymbol{h}_\mu^{(\ell-1)}}{dt} \tag{37}$$

This implies $2a_\ell + c - d - 1/2 = 0$. Given (32), this leads to

$$d = \frac{1}{2} \tag{38}$$

On the other hand, any choice of $d < \frac{1}{2}$ gives kernel behavior. The choice $d = 0$ corresponds to the NTK parameterization.

4.1.4 Putting Constraints Together

The set of parameterizations which yield $\mathcal{O}(1)$ feature evolution are those for which

1. At initialization, features h are $\mathcal{O}_N(1) \implies 2a_\ell + b_\ell = 1$ for $\ell \in \{2, ..., L\}$ and $2a_1 + b_1 = 0$.
2. Outputs predictions evolve in $\mathcal{O}_N(1)$ time $\implies 2a_\ell + c = 1$, for $\ell \in \{2, ..., L\}$, $2a_1 + c = 0$
3. Features h have $\mathcal{O}_N(1)$ evolution $\implies d = \frac{1}{2}$.

The parameterization discussed in Section 2 satisfies these with $d = \frac{1}{2}, a_\ell = \frac{1}{2}, b_\ell = 0, c = 0$. The quite general requirement for feature learning that $d = \frac{1}{2}$ indicates that $\gamma = \gamma_0\sqrt{N}$ for any choice of a_ℓ, b_ℓ, c as we use in the main text. This indicates that neural network prediction logits at initialization scale as $f_\mu \sim \mathcal{O}(N^{-1/2})$ in the feature learning infinite width limit. The set of parameterizations which meet these three requirements is one dimensional with $d = \frac{1}{2}$, and $(a, b, c) \in \{(a, 1-2a, 1-2a) : a \in \mathbb{R}\}$ for all layers except the first layer which has $(a_1 = a - \frac{1}{2}, b_1 = 1 - 2a)$. Our parameterization corresponds to $a = \frac{1}{2}$.

If one further demands $\mathcal{O}_N(1)$ raw learning rate η, then the parameterization is unique. This requires, $\eta = \eta_0\gamma^2 N^{-c} = \mathcal{O}_N(N^{2d-c}) = \mathcal{O}_N(1) \implies c = 2d = 1$. Under this constraint, $a_\ell = 0$ and $b_\ell = 1$ for $\ell \in \{2, ..., L\}$ and $a_1 = -\frac{1}{2}$ and $b_1 = 1$, which corresponds to a modification of standard parameterization, with first and last layer altered with width. In a computational algorithm, the learning rate would be $\eta = \eta_0\gamma^2 N^{-c} = \eta_0\gamma_0^2 = \mathcal{O}_N(1)$. This is equivalent to the μP parameterization stated in Yang and Hu [5].

5 Construction of a theoretical description in the infinite-width limit

In this section, we introduce the dynamical field theory setup and saddle point equations. The path integral theory we develop is based on the Martin-Siggia-Rose-De Dominicis-Janssen (MSRDJ)

framework [12], of which a useful review for random recurent networks can be found here [13, 14]. Similar computations can be found in recent works which consider typical behavior in high dimensional classification on random data [15, 16].

5.1 Deep Network Field Definitions and Scaling

Some of this section will be repetitive, but it is good to be reminded of things. As discussed before, we consider the following wide network architecture

$$f_\mu = \frac{1}{\gamma_0\sqrt{N}} h_\mu^{(L)} \ , \qquad h_\mu^{(L)} = \frac{1}{\sqrt{N}} \boldsymbol{w}^{(L)} \cdot \phi(\boldsymbol{h}_\mu^{(L-1)})$$
$$\boldsymbol{h}_\mu^{(\ell)} = \frac{1}{\sqrt{N}} \boldsymbol{W}^{(\ell)} \phi(\boldsymbol{h}_\mu^{(\ell-1)}) \ , \qquad \boldsymbol{h}_\mu^{(1)} = \frac{1}{\sqrt{D}} \boldsymbol{W}^{(1)} \boldsymbol{x}_\mu \tag{39}$$

with initialization $\theta_i \sim \mathcal{N}(0,1)$. Let's also remind ourselves

$$\boldsymbol{g}_\mu^{(\ell)} = \sqrt{N} \frac{\partial h_\mu^{(L)}}{\partial \boldsymbol{h}_\mu^{(\ell)}} \tag{40}$$

which admit the recursion and base case

$$\boldsymbol{g}_\mu^{(\ell)} = \sqrt{N} \frac{\partial h_\mu^{(L)}}{\partial \boldsymbol{h}_\mu^{(\ell)}} = \left(\frac{\partial \boldsymbol{h}_\mu^{(\ell+1)}}{\partial \boldsymbol{h}_\mu^{(\ell)}} \right)^\top \left(\sqrt{N} \frac{\partial h_\mu^{(L)}}{\partial \boldsymbol{h}_\mu^{(\ell+1)}} \right) = \dot{\phi}(\boldsymbol{h}_\mu^{(\ell)}) \odot \boldsymbol{z}_\mu^{(\ell)} \ , \ \boldsymbol{z}_\mu^{(\ell)} \equiv \frac{1}{\sqrt{N}} \boldsymbol{W}^{(\ell+1)\top} \boldsymbol{g}_\mu^{(\ell+1)}$$
$$\boldsymbol{g}_\mu^{(L-1)} = \dot{\phi}(\boldsymbol{h}_\mu^{(L-1)}) \odot \boldsymbol{w}^{(L)}, \qquad \boldsymbol{z}_\mu^{(L-1)} \equiv \frac{1}{\sqrt{N}} \boldsymbol{w}^{(L)} g_\mu^{(L)}, \qquad g_\mu^{(L)} \equiv \sqrt{N} \tag{41}$$

and

$$\Phi_{\mu\nu}^{(\ell)}(t,s) = \frac{1}{N} \phi(\boldsymbol{h}_\mu^{(\ell)}(t)) \cdot \phi(\boldsymbol{h}_\nu^{(\ell)}(s)) \ , \qquad G_{\mu\nu}^{(\ell)}(t,s) = \frac{1}{N} \boldsymbol{g}_\mu^{(\ell)}(t) \cdot \boldsymbol{g}_\nu^{(\ell)}(s) \tag{42}$$

.

Using gradient flow with learning rate $\eta = \eta_0 \gamma^2$ on loss function $\mathcal{L} = \frac{1}{P} \sum_{\mu=1}^P \ell(f_\mu, y_\mu)$, and remembering the definition of $\Delta_\mu = -\frac{\partial \mathcal{L}}{\partial f_\mu}$, gradient flow induces the following dynamics

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\eta_0 \gamma}{P} \sum_\mu \Delta_\mu \frac{\partial h_\mu^{(L)}}{\partial \boldsymbol{\theta}} \ , \qquad \frac{df_\mu}{dt} = \frac{\eta_0}{P} \sum_\alpha \Delta_\alpha K_{\mu\alpha}^{NTK}, \quad K_{\mu\alpha}^{NTK} \equiv \frac{\partial h_\mu^{(L)}}{\partial \boldsymbol{\theta}} \cdot \frac{\partial h_\alpha^{(L)}}{\partial \boldsymbol{\theta}}, \tag{43}$$

where $\boldsymbol{K}^{NTK}$ is the Neural Tangent Kernel [17]. For the scaling we are working with, we already showed that (Equation (26))

$$K_{\mu\alpha}^{NTK} = \Phi_{\mu\nu}^{(L-1)} + \sum_{\ell=1}^{L-1} G_{\mu\nu}^{(\ell)} \Phi_{\mu\nu}^{(\ell-1)}. \tag{44}$$

Ultimately, we want to average over initializations to develop our mean field theory. We want to come up with a reformulation of gradient flow equations that depend only on weights at initalization,

but not later. This turns out to be possible through simple manipulations.

The update equations for $\boldsymbol{W}^{(\ell)}$ and $\boldsymbol{h}^{(\ell)}$ give

$$\frac{d}{dt}\boldsymbol{W}^{(\ell)} = \frac{\eta_0\gamma_0}{P}\sum_{\mu=1}^{P}\Delta_\mu \frac{\partial h^{(L)}}{\partial \boldsymbol{h}_\mu^\ell}\phi(\boldsymbol{h}_\mu^{(\ell-1)})^\top = \frac{\eta_0\gamma_0}{\sqrt{N}P}\sum_{\mu=1}^{P}\Delta_\mu \boldsymbol{g}_\mu^{(\ell)}\phi(\boldsymbol{h}_\mu^{(\ell-1)})^\top$$

Integrating in time:

$$\boldsymbol{W}^{(\ell)}(t) = \boldsymbol{W}^{(\ell)}(0) + \frac{\eta_0\gamma_0}{\sqrt{N}P}\int_0^t ds \sum_\mu \Delta_\mu(s)\boldsymbol{g}_\mu^{(\ell)}(s)\phi(\boldsymbol{h}_\mu^{(\ell-1)}(s))^\top \tag{45}$$

Now, noting that $\boldsymbol{h}^{(\ell)}(t) = \frac{1}{\sqrt{N}}\boldsymbol{W}^{(\ell)}(t)\phi(\boldsymbol{h}^{(\ell-1)}(t))$

$$\boldsymbol{h}_\mu^{(\ell)}(t) = \boldsymbol{\chi}_\mu^{(\ell)}(t) + \frac{\eta_0\gamma_0}{P}\sum_\nu \int_0^t ds \Delta_\nu(s)\boldsymbol{g}_\nu^{(\ell)}(s)\Phi_{\mu\nu}^{(\ell-1)}(t,s). \tag{46}$$

where

$$\boldsymbol{\chi}_\mu^{(\ell)}(t) \equiv \frac{1}{\sqrt{N}}\boldsymbol{W}^{(\ell)}(0)\phi(\boldsymbol{h}_\mu^{(\ell-1)}(t)). \tag{47}$$

We can do the same thing for the $\boldsymbol{z}_\mu^\ell(t)$ variables using their iterative definition.

Collecting together:

$$\boldsymbol{h}_\mu^{(\ell)}(t) = \boldsymbol{\chi}_\mu^{\ell}(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu \Delta_\nu(s)\boldsymbol{g}_\nu^{(\ell)}(t)\Phi_{\mu\nu}^{(\ell-1)}(s,t), \qquad \ell = 1,\dots L-1$$

$$\boldsymbol{z}_\mu^{(\ell)}(t) = \boldsymbol{\xi}_\mu^{(\ell)}(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu \Delta_\nu(s)\phi(\boldsymbol{h}_\nu^{(\ell)}(s))G_{\mu\nu}^{(\ell+1)}(s,t)\ ,\ \boldsymbol{g}_\mu^{(\ell)}(t) = \dot\phi(\boldsymbol{h}_\mu^{(\ell)}(t))\odot \boldsymbol{z}_\mu^{(\ell)}(t),$$

$$\ell = 1,\dots L-1$$

$$\frac{\partial f_\mu}{\partial t} = \frac{\eta_0}{P}\sum_\nu \Delta_\nu(t)\left[\Phi_{\mu\nu}^{(L-1)} + \sum_{\ell=1}^{L-1} G_{\mu\nu}^{(\ell)}\Phi_{\mu\nu}^{(\ell-1)}\right] \tag{48}$$

In the above, we implicitly utilize the base cases $G_{\mu\nu}^{(L)}(t,s) = 1$. We also introduced the following random fields $\boldsymbol{\chi}_\mu^\ell(t), \boldsymbol{\xi}_\mu^\ell(t)$ which involve the random initial conditions

$$\boldsymbol{\chi}_\mu^{(\ell)}(t) = \frac{1}{\sqrt{N}}\boldsymbol{W}^{(\ell)}(0)\phi(\boldsymbol{h}_\mu^{(\ell-1)}(t))\ ,\ \boldsymbol{\xi}_\mu^{(\ell)}(t) = \frac{1}{\sqrt{N}}\boldsymbol{W}^{(\ell+1)}(0)^\top \boldsymbol{g}_\mu^{(\ell+1)}(t). \tag{49}$$

We observe that the dynamics of the hidden features is controlled by the factor γ_0. If $\gamma_0 = o(1)$ then we recover static NTK in the limit as $N \to \infty$.

5.2 DMFT for Two Layer NN

To construct our mean field theory, we will compute the moment generating functional for the stochastic processes $\{\boldsymbol{\chi}^{(\ell)}, \boldsymbol{\xi}^{(\ell)}\}_{\ell=1}^{L-1}$

$$Z[\{\boldsymbol{j}^{(\ell)}, \boldsymbol{v}^{(\ell)}\}] = \left\langle \exp\left(\sum_{\ell=1}^{L-1} \sum_{\mu=1}^{P} \int_0^\infty dt \left[\boldsymbol{j}_\mu^{(\ell)}(t) \cdot \boldsymbol{\chi}_\mu^{(\ell)}(t) + \boldsymbol{v}_\mu^{(\ell)}(t) \cdot \boldsymbol{\xi}_\mu^{(\ell)}(t) \right] \right) \right\rangle_{\boldsymbol{\theta}_0 = \mathrm{Vec}\{\boldsymbol{W}^1(0), \ldots \boldsymbol{w}^L(0)\}} \tag{50}$$

Moments of these stochastic fields can be computed through differentiation of Z near zero-source

$$\begin{aligned} &\left\langle \chi_{\mu_1}^{(\ell_1)}(t_1) \ldots \chi_{\mu_n}^{(\ell_n)}(t_n) \xi_{\bar{\mu}_1}^{(\bar{\ell}_1)}(t_1) \ldots \xi_{\bar{\mu}_m}^{(\bar{\ell}_m)}(t_m) \right\rangle \\ &\qquad = \frac{\delta}{\delta j_{\mu_1}^{(\ell_1)}(t_1)} \ldots \frac{\delta}{\delta j_{\mu_n}^{(\ell_n)}(t_n)} \frac{\delta}{\delta v_{\bar{\mu}_1}^{(\bar{\ell}_1)}(t_1)} \ldots \frac{\delta}{\delta v_{\bar{\mu}_m}^{(\bar{\ell}_m)}(t_m)} Z[\{\boldsymbol{j}^\ell, \boldsymbol{v}^\ell\}]|_{\boldsymbol{j}=\boldsymbol{v}=0}. \end{aligned} \tag{51}$$

Here, we provide a warmup problem of a 2 layer network which allows us to illustrate the mechanics of the MSRDJ formalism. Networks with more layers are considered in [1, 4]. We will give the result of that computation in the next section.

Though many of the interesting dynamical aspects of the deep network case are missing in the two layer case, our aim is to show a simple application of the ideas. The fields of interest are $\boldsymbol{\chi}_\mu = \frac{1}{\sqrt{D}} \boldsymbol{W}^{(1)}(0) \boldsymbol{x}_\mu$ and $\boldsymbol{\xi} = \boldsymbol{w}^{(2)}(0)$. Unlike the deeper $L > 2$ case, both of these fields are time invariant since $\boldsymbol{x}_\mu$ does not vary in time. These random fields provide initial conditions for the preactivation and pre-gradient fields $\boldsymbol{h}_\mu(t), \boldsymbol{z}(t) \in \mathbb{R}^N$, which evolve according to

$$\begin{aligned} \boldsymbol{h}_\mu(t) &= \boldsymbol{\chi}_\mu + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\alpha \boldsymbol{g}_\alpha \Phi_{\mu\alpha}^{(0)} \Delta_\alpha(s), \\ \boldsymbol{z}(t) &= \boldsymbol{\xi} + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\alpha \phi(\boldsymbol{h}_\alpha(s)) \Delta_\alpha(s), \\ \boldsymbol{g}_\mu &= \dot{\phi}(\boldsymbol{h}_\mu(t)) \odot \boldsymbol{z}. \end{aligned} \tag{52}$$

where the network predictions evolve as

$$\frac{d}{dt} f_\mu(t) = \frac{\eta_0}{P} \sum_\alpha [\Phi_{\mu\alpha}(t,t) + G_{\mu\alpha}(t,t) \Phi_{\mu\alpha}^{(0)}] \Delta_\alpha(t) \tag{53}$$

for kernels

$$\Phi_{\mu\alpha}(t,t) = \frac{1}{N} \phi(\boldsymbol{h}_\mu(t)) \cdot \phi(\boldsymbol{h}_\alpha(t)), \qquad G_{\mu\alpha}(t,t) = \frac{1}{N} \boldsymbol{g}_\mu(t) \cdot \boldsymbol{g}_\alpha(t). \tag{54}$$

At finite N, the kernels Φ, G will depend on the random initial conditions $\boldsymbol{\chi}, \boldsymbol{\xi}$, leading to a predictor f_μ which varies over initializations. If we can establish that the kernels Φ, G concentrate at infinite-width $N \to \infty$, then Δ_μ are deterministic.

We now study the moment generating function for the fields

$$Z[\{\boldsymbol{j}_\mu\}_{\mu\in[P]}, \boldsymbol{v}] = \left\langle \exp\left(\sum_\mu \boldsymbol{j}_\mu \cdot \boldsymbol{\chi}_\mu + \boldsymbol{\xi}\cdot\boldsymbol{v}\right)\right\rangle_{\boldsymbol{\theta}_0}. \tag{55}$$

To perform the average over $\boldsymbol{\theta}_0 = \{\boldsymbol{W}^{(1)}(0), \boldsymbol{w}^{(2)}(0)\}$, we enforce the definition of $\boldsymbol{\chi}_\mu, \boldsymbol{\xi}$ with delta functions

$$\begin{aligned}
1 &= \int d\boldsymbol{\chi}_\mu\, \delta\left(\boldsymbol{\chi}_\mu - \frac{1}{\sqrt{N}}\boldsymbol{W}^{(1)}(0)\boldsymbol{x}_\mu\right) = \int \frac{d\boldsymbol{\chi}_\mu d\hat{\boldsymbol{\chi}}_\mu}{(2\pi)^N} \exp\left(i\hat{\boldsymbol{\chi}}_\mu \cdot \left(\boldsymbol{\chi}_\mu - \frac{1}{\sqrt{D}}\boldsymbol{W}^{(1)}(0)\boldsymbol{x}_\mu\right)\right) \\
1 &= \int d\boldsymbol{\xi}\ \delta\left(\boldsymbol{\xi} - \boldsymbol{w}^{(2)}(0)\right) = \int \frac{d\boldsymbol{\xi} d\hat{\boldsymbol{\xi}}}{(2\pi)^N} \exp\left(i\hat{\boldsymbol{\xi}}\cdot\left(\boldsymbol{\xi} - \boldsymbol{w}^{(2)}(0)\right)\right).
\end{aligned} \tag{56}$$

Though this step may seem redundant in this example, it will be very helpful in the deep network case, so we pursue it for illustration. After multiplying by these factors of unity and performing the Gaussian integrals, we obtain

$$Z = \int \prod_\mu \frac{d\boldsymbol{\chi}_\mu d\hat{\boldsymbol{\chi}}_\mu}{(2\pi)^N} \frac{d\boldsymbol{\xi} d\hat{\boldsymbol{\xi}}}{(2\pi)^N} \exp\left(-\frac{1}{2}\sum_{\mu\alpha}\hat{\boldsymbol{\chi}}_\mu\cdot\hat{\boldsymbol{\chi}}_\alpha \Phi^{(0)}_{\mu\alpha} + \sum_\mu \boldsymbol{\chi}_\mu\cdot(i\hat{\boldsymbol{\chi}}_\mu + \boldsymbol{j}_\mu) - \frac{1}{2}|\hat{\boldsymbol{\xi}}|^2 + \boldsymbol{\xi}\cdot(i\hat{\boldsymbol{\xi}} + \boldsymbol{v})\right) \tag{57}$$

We now aim enforce the definitions of the kernel order parameters with delta functions

$$\begin{aligned}
1 &= N\int d\Phi_{\mu\alpha}(t,s)\ \delta\left(N\Phi_{\mu\alpha}(t,s) - \phi(\boldsymbol{h}_\mu(t))\cdot\phi(\boldsymbol{h}_\alpha(s))\right) \\
&= \int \frac{d\Phi_{\mu\alpha}(t,s) d\hat{\Phi}_{\mu\alpha}(t,s)}{2\pi i N^{-1}} \exp\left(N\hat{\Phi}_{\mu\alpha}(t,s)\left(N\Phi_{\mu\alpha}(t,s) - \phi(\boldsymbol{h}_\mu(t))\cdot\phi(\boldsymbol{h}_\alpha(s))\right)\right) \\
1 &= N\int dG_{\mu\alpha}(t,s)\ \delta\left(NG_{\mu\alpha}(t,s) - \boldsymbol{g}_\mu(t)\cdot\boldsymbol{g}_\alpha(s)\right) \\
&= \int \frac{dG_{\mu\alpha}(t,s) d\hat{G}_{\mu\alpha}(t,s)}{2\pi i N^{-1}} \exp\left(N\hat{G}_{\mu\alpha}(t,s)\left(NG_{\mu\alpha}(t,s) - \boldsymbol{g}_\mu(t)\cdot\boldsymbol{g}_\alpha(s)\right)\right),
\end{aligned} \tag{58}$$

where the fields $\boldsymbol{h}_\mu(t), \boldsymbol{g}_\mu(t)$ are regarded as functions of $\{\boldsymbol{\chi}_\mu\}_\mu, \boldsymbol{\xi}$ (see Equation (52)) and the $\hat{\Phi}, \hat{G}$ integrals run over the imaginary axis $(-i\infty, i\infty)$. After this step, we can write

$$Z \propto \int \prod_{\mu\alpha ts} d\Phi_{\mu\alpha}(t,s) d\hat{\Phi}_{\mu\alpha}(t,s) dG_{\mu\alpha}(t,s) d\hat{G}_{\mu\alpha}(t,s) \exp\left(NS[\Phi, \hat{\Phi}, G, \hat{G}]\right) \tag{59}$$

where the DMFT action $S[\Phi, \hat{\Phi}, G, \hat{G}]$ is $\mathcal{O}_N(1)$ and has the form

$$S[\Phi, \hat{\Phi}, G, \hat{G}] = \sum_{\mu\alpha}\int dt ds [\Phi_{\mu\alpha}(t,s)\hat{\Phi}_{\mu\alpha}(t,s) + G_{\mu\alpha}(t,s)\hat{G}_{\mu\alpha}(t,s)] + \frac{1}{N}\sum_{i=1}^N \ln \mathcal{Z}[j_i, v_i]. \tag{60}$$

The single site moment generating function $\mathcal{Z}[j, v]$ arises from the factorization of the integrals over

N different fields in the hidden layer and takes the form

$$\mathcal{Z}[j,v] = \int \prod_{\mu} \frac{d\chi_\mu d\hat{\chi}_\mu}{2\pi} \frac{d\xi d\hat{\xi}}{2\pi} \exp\left(-\frac{1}{2} \sum_{\mu\alpha} \hat{\chi}_\mu \hat{\chi}_\alpha \Phi^{(0)}_{\mu\alpha} + (j_\mu + i\hat{\chi}_\mu)\chi_\mu - \frac{1}{2}\hat{\xi}^2 + (v + i\hat{\xi})\xi \right)$$
$$\times \exp\left(- \int_0^\infty dt \int_0^\infty ds \sum_{\mu\alpha} [\hat{\Phi}_{\mu\alpha}(t,s)\phi(h_\mu(t))\phi(h_\alpha(s)) + \hat{G}_{\mu\alpha}(t,s) g_\mu(t) g_\alpha(s)] \right) \quad (61)$$

where, again we must regard $h_\mu(t), g_\mu(t)$ as functions of χ, ξ. The variables in the above are no longer vectors in $\mathbb{R}^N$ but rather are scalars. We can write $\mathcal{Z}[j,v] = \int \prod_\mu d\chi_\mu d\hat{\chi}_\mu d\xi d\hat{\xi} \exp\left(-\mathcal{H}[\{\chi_\mu, \hat{\chi}_\mu\}, \xi, \hat{\xi}, j, v] \right)$ where $\mathcal{H}$ is the logarithm of the integrand above. Since the full MGF takes the form $Z \propto \int d\Phi d\hat{\Phi} dG d\hat{G} \exp\left(N S[\Phi, \hat{\Phi}, G, \hat{G}] \right)$, characterization of the $N \to \infty$ limit requires one to identify the saddle point of S, where $\delta S = 0$ for any variation of these 4 order parameters.

$$\frac{\delta S}{\delta \Phi_{\mu\alpha}(t,s)} = \hat{\Phi}_{\mu\alpha}(t,s) = 0 \ , \ \frac{\delta S}{\delta \hat{\Phi}_{\mu\alpha}(t,s)} = \Phi_{\mu\alpha}(t,s) - \frac{1}{N} \sum_{i=1}^N \langle \phi(h_\mu(t))\phi(h_\alpha(s)) \rangle_i = 0$$
$$\frac{\delta S}{\delta G_{\mu\alpha}(t,s)} = \hat{G}_{\mu\alpha}(t,s) = 0 \ , \ \frac{\delta S}{\delta \hat{G}_{\mu\alpha}(t,s)} = G_{\mu\alpha}(t,s) - \frac{1}{N} \sum_{i=1}^N \langle g_\mu(t) g_\alpha(s) \rangle_i = 0 \quad (62)$$

where the i-th single site average $\langle \rangle_i$ of an observable $O(\chi, \hat{\chi}, \xi, \hat{\xi})$ is defined as

$$\left\langle O(\chi, \hat{\chi}, \xi, \hat{\xi}) \right\rangle_i = \frac{1}{\mathcal{Z}[j_i, v_i]} \int \prod_\mu d\chi_\mu d\hat{\chi}_\mu d\xi d\hat{\xi} \exp\left(-\mathcal{H}[\{\chi_\mu, \hat{\chi}_\mu\}, \xi, \hat{\xi}, j_i, v_i] \right) O(\chi, \hat{\chi}, \xi, \hat{\xi}) \quad (63)$$

Since $\hat{\Phi} = \hat{G} = 0$ the single site MGF reveals that the initial fields are independent Gaussians $\{\chi_\mu\} \sim \mathcal{N}(0, \boldsymbol{K}^x)$ and $\xi \sim \mathcal{N}(0,1)$. At zero source $\boldsymbol{j}, \boldsymbol{v} \to 0$, all single site averages $\langle \rangle_i$ are equivalent and we may merely write $\Phi_{\mu\alpha}(t,s) = \langle \phi(h_\mu(t))\phi(h_\alpha(s)) \rangle$, $G_{\mu\alpha}(t,s) = \langle g_\mu(t) g_\alpha(s) \rangle$, where $\langle \rangle$ is the average over the single site distributions for $\boldsymbol{j}, \boldsymbol{v} \to 0$.

Putting all of the saddle point equations together, we arrive at the following DMFT

$$\{\chi_\mu\}_{\mu \in [P]} \sim \mathcal{N}(0, \boldsymbol{\Phi}^{(0)}) \ , \ \xi \sim \mathcal{N}(0,1)$$
$$h_\mu(t) = \chi_\mu + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\alpha [z(s)\dot{\phi}(h_\alpha(s))] \Phi^{(0)}_{\mu\alpha} \Delta_\alpha(s)$$
$$z(t) = \xi + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\alpha \phi(h_\alpha(s)) \Delta_\alpha(s)$$
$$\Phi_{\mu\alpha}(t,s) = \langle \phi(h_\mu(t))\phi(h_\alpha(s)) \rangle \ , \ G_{\mu\alpha}(t,s) = \left\langle z(t) z(s) \dot{\phi}(h_\mu(t)) \dot{\phi}(h_\alpha(s)) \right\rangle$$
$$\frac{\partial f_\mu}{\partial t} = \frac{\eta_0}{P} \sum_\alpha [\Phi_{\mu\alpha}(t,t) + G_{\mu\alpha}(t,t)\Phi^{(0)}_{\mu\alpha}] \Delta_\alpha(t) \quad (64)$$

We see that for $L = 2$ networks, it suffices to solve for the kernels on the time-time diagonal. Further in this two layer case χ, ξ are independent and do not vary in time. These facts will not hold in general for $L \geq 2$ networks, which requires a more intricate analysis as we show in the next

section.

5.3 DMFT for $L > 2$

We have the following complete DMFT equations. The full derivation is given in appendices of [18]. See also slides:

$$
\begin{aligned}
\{u_\mu^{(\ell)}(t)\}_{\mu\in[P],t\in\mathbb{R}_+} &\sim \mathcal{GP}(0, \mathbf{\Phi}^{\ell-1}) \ , \ \{r_\mu^{(\ell)}(t)\}_{\mu\in[P],t\in\mathbb{R}_+} \sim \mathcal{GP}(0, \boldsymbol{G}^{(\ell+1)}), \\
h_\mu^{(\ell)}(t) &= u_\mu^{(\ell)}(t) + \gamma_0 \int_0^t ds \sum_\alpha \left[A_{\mu\alpha}^{(\ell-1)}(t,s) + \Delta_\alpha(s) \Phi_{\mu\alpha}^{(\ell-1)}(t,s) \right] \dot\phi(h_\alpha^{(\ell)}(s)) z_\alpha^{(\ell)}(s) \\
z_\mu^{(\ell)}(t) &= r_\mu^{(\ell)}(t) + \gamma_0 \int_0^t \sum_\alpha \left[B_{\mu\alpha}^{\ell}(t,s) + \Delta_\alpha(s) G_{\mu\alpha}^{(\ell+1)}(t,s) \right] \phi(h_\alpha^{(\ell)}(s)) \\
\Phi_{\mu\alpha}^{(\ell)}(t,s) &= \left\langle \phi(h_\mu^{(\ell)}(t)) \phi(h_\alpha^{(\ell)}(s)) \right\rangle \ , \ G_{\mu\alpha}^{(\ell)}(t,s) = \left\langle g_\mu^{(\ell)}(t) g_\alpha^{(\ell)}(s) \right\rangle \\
A_{\mu\alpha}^{(\ell)}(t,s) &= \gamma_0^{-1} \left\langle \frac{\delta \phi(h_\mu^{(\ell)}(t))}{\delta r_\alpha^{(\ell)}(s)} \right\rangle \ , \ B_{\mu\alpha}^{(\ell)}(t,s) = \gamma_0^{-1} \left\langle \frac{\delta g_\mu^{(\ell+1)}(t)}{\delta u_\alpha^{(\ell+1)}(s)} \right\rangle, \\
\frac{d}{dt} f_\mu(t) &= \frac{\eta_0}{P} \sum_\alpha [\Phi_{\mu\alpha}(t,t) + G_{\mu\alpha}(t,t) \Phi_{\mu\alpha}^{(0)}] \Delta_\alpha(t)
\end{aligned}
\tag{65}
$$

where we define base cases $\Phi_{\mu\alpha}^{(0)}(t,s) = \frac{1}{\sqrt{D}} \boldsymbol{x}_\mu \cdot \boldsymbol{x}_\nu$, and $G_{\mu\alpha}^{(L)}(t,s) = 1$, $A^{(1)} = B^{(L)} = 0$. Note that A and B are new fields that do not appear in two layers.

5.4 Extensions

See slides and [4] for extensions to other learning rules. [1] also discusses extensions to gradient descent, weight decay, multiple output channels, varying widths, convolutional networks, momentum, and Langevin dynamics (Bayesian networks).

5.5 Relation to some other work

Previous work studied the same limit and weight scaling we are studying in two layer networks and dervied a mean field theory description of learning dynamics [19, 20]. This mean field theory is given by a partial differential equation (PDE) that tracks the evolution of the density of parameters (weights). A PDE version of our DMFT can be derived in two layers, see appendices of [1], however this PDE tracks the density of $\boldsymbol{h}$ and z.

6 Linear Networks

When activations are taken to be linear, the DMFT simplifies and can be closed under deterministic equations [18]. Here, we will work out the simplest case of two-layer networks, MSE loss and $\Phi^{(0)} = \mathbf{I}$ here, which is exactly solvable [18].

For two layer linear networks, (64) becomes

$$
\begin{aligned}
\{\chi_\mu\}_{\mu\in[P]} &\sim \mathcal{N}(0,\mathbf{\Phi}^{(0)})\ ,\ \xi\sim\mathcal{N}(0,1)\\
h_\mu(t) &= \chi_\mu + \frac{\eta_0\gamma_0}{P}\int_0^t ds\, g(s)\sum_\alpha \Phi^{(0)}_{\mu\alpha}\Delta_\alpha(s)\\
g(t) &= \xi + \frac{\eta_0\gamma_0}{P}\int_0^t ds\sum_\alpha h_\alpha(s)\Delta_\alpha(s)\\
H_{\mu\alpha}(t,s) &= \langle h_\mu(t)h_\alpha(s)\rangle\ ,\ G(t,s)=\langle g(t)g(s)\rangle\,, \qquad \Delta_\mu = -\frac{\partial l_\mu}{\partial f_\mu},\\
\frac{df_\mu}{dt} &= \frac{\eta_0}{P}\sum_\alpha [H_{\mu\alpha}(t,t)+G(t,t)\Phi^{(0)}_{\mu\alpha}]\Delta_\alpha(t)
\end{aligned}
\tag{66}
$$

We will introduce a vector notation for this section $\boldsymbol{h}$, $\boldsymbol{f}$ and $\boldsymbol{\Delta}$, whose elements are $h_{\mu=1,\dots,P}$, $f_{\mu=1,\dots,P}$ and $\Delta_{\mu=1,\dots,P}$, and a matrix notation for $\boldsymbol{H}=\left\langle \boldsymbol{h}\boldsymbol{h}^\top\right\rangle$.

In this case, we can have a differential version of these equations:

$$
\begin{aligned}
\frac{d\boldsymbol{h}(t)}{dt} &= \frac{\eta_0\gamma_0}{P}g(t)\mathbf{\Phi}^{(0)}\boldsymbol{\Delta},\\
\frac{dg(t)}{dt} &= \frac{\eta_0\gamma_0}{P}\boldsymbol{h}\cdot\boldsymbol{\Delta}\\
\boldsymbol{H}(t) &= \left\langle \boldsymbol{h}(t)\boldsymbol{h}(t)^\top\right\rangle, \qquad G(t)=\left\langle g(t)^2\right\rangle,\\
\frac{d\boldsymbol{f}(t)}{dt} &= \frac{\eta_0}{P}\left[\boldsymbol{H}(t)+G(t)\mathbf{\Phi}^{(0)}\right]\boldsymbol{\Delta},
\end{aligned}
\tag{67}
$$

$$
\boldsymbol{h}(0)\sim\mathcal{N}(0,\mathbf{\Phi}^{(0)}),\quad \boldsymbol{g}(0)\sim\mathcal{N}(0,1),\quad \boldsymbol{H}(0)=\mathbf{\Phi}^{(0)},\quad G(0)=1,\quad \boldsymbol{f}(0)=\boldsymbol{0}.
\tag{68}
$$

If we further commit to MSE loss, we can get rid of the stochasticity fully, and obtain a set of equations that only involve $\boldsymbol{H}$, G and $\boldsymbol{f}$. In this case:

$$
\Delta_\mu = y_\mu - f_\mu,\quad \implies \quad \boldsymbol{\Delta}=\boldsymbol{y}-\boldsymbol{f},\quad \dot{\boldsymbol{\Delta}}=-\dot{\boldsymbol{f}}.
\tag{69}
$$

Now, note that:

$$
\begin{aligned}
\frac{d}{dt}\boldsymbol{H}(t) &= \left\langle \dot{\boldsymbol{h}}(t)\boldsymbol{h}(t)^\top\right\rangle + \left\langle \boldsymbol{h}(t)\dot{\boldsymbol{h}}(t)^\top\right\rangle = \frac{\eta_0\gamma_0}{P}\mathbf{\Phi}^{(0)}\boldsymbol{\Delta}\left\langle g(t)\boldsymbol{h}(t)^T\right\rangle + \frac{\eta_0\gamma_0}{P}\left\langle g(t)\boldsymbol{h}(t)\right\rangle\boldsymbol{\Delta}^\top\mathbf{\Phi}^{(0)},\\
\frac{d}{dt}G(t) &= 2\left\langle \dot{g}(t)g(t)\right\rangle = 2\frac{\eta_0\gamma_0}{P}\left\langle g(t)\boldsymbol{h}(t)\right\rangle\cdot\boldsymbol{\Delta},
\end{aligned}
\tag{70}
$$

and

$$
\frac{d}{dt}\left\langle g(t)\boldsymbol{h}(t)\right\rangle = \left\langle \dot{g}(t)\boldsymbol{h}(t)\right\rangle + \left\langle g(t)\dot{\boldsymbol{h}}(t)\right\rangle = \frac{\eta_0\gamma_0}{P}\boldsymbol{H}(t)\boldsymbol{\Delta} + \frac{\eta_0\gamma_0}{P}G(t)\mathbf{\Phi}^{(0)}\boldsymbol{\Delta} = \gamma_0\frac{d\boldsymbol{f}(t)}{dt}
\tag{71}
$$

Noting that $\left\langle g(0)\boldsymbol{h}(0)\right\rangle = \boldsymbol{f}(0)=0$, we can conclude that $\left\langle g(t)\boldsymbol{h}(t)\right\rangle=\gamma_0\boldsymbol{f}(t)$.

Collecting all these facts together, we are left with the following set of equations

$$\begin{aligned}
\frac{d\boldsymbol{H}(t)}{dt} &= \frac{\eta_0\gamma_0^2}{P}\boldsymbol{\Phi}^{(0)}(\boldsymbol{y}-\boldsymbol{f})\boldsymbol{f}^\top + \frac{\eta_0\gamma_0^2}{P}\boldsymbol{f}(\boldsymbol{y}-\boldsymbol{f})^\top\boldsymbol{\Phi}^{(0)},\\
\frac{dG(t)}{dt} &= 2\frac{\eta_0\gamma_0^2}{P}\boldsymbol{f}\cdot(\boldsymbol{y}-\boldsymbol{f}),\\
\frac{d\boldsymbol{f}(t)}{dt} &= \frac{\eta_0}{P}\left[\boldsymbol{H}(t)+G(t)\boldsymbol{\Phi}^{(0)}\right](\boldsymbol{y}-\boldsymbol{f}),\\
\boldsymbol{H}(0) &= \boldsymbol{\Phi}^{(0)}, \qquad G(0)=1, \qquad \boldsymbol{f}(0)=\boldsymbol{0}.
\end{aligned} \tag{72}$$

We repeat that we obtained a deterministic set of equations.

We can further simplify this set of equations if we commit to $\boldsymbol{\Phi}^{(0)}=\mathbf{I}$. In this case we get:

$$\begin{aligned}
\frac{d\boldsymbol{H}(t)}{dt} &= \frac{\eta_0\gamma_0^2}{P}(\boldsymbol{y}-\boldsymbol{f})\boldsymbol{f}^\top + \frac{\eta_0\gamma_0^2}{P}\boldsymbol{f}(\boldsymbol{y}-\boldsymbol{f})^\top,\\
\frac{dG(t)}{dt} &= 2\frac{\eta_0\gamma_0^2}{P}\boldsymbol{f}\cdot(\boldsymbol{y}-\boldsymbol{f}),\\
\frac{d\boldsymbol{f}(t)}{dt} &= \frac{\eta_0}{P}\left[\boldsymbol{H}(t)+G(t)\mathbf{I}\right](\boldsymbol{y}-\boldsymbol{f}),\\
\boldsymbol{H}(0) &= \mathbf{I}, \qquad G(0)=1, \qquad \boldsymbol{f}(0)=\boldsymbol{0}.
\end{aligned} \tag{73}$$

Now, we assume an änsatz of the form:

$$\boldsymbol{H}(t) = (H_y(t)-1)\,\frac{\boldsymbol{y}\boldsymbol{y}^\top}{|\boldsymbol{y}|^2} + \mathbf{I}, \qquad \boldsymbol{f}(t) = f_y(t)\frac{\boldsymbol{y}}{|\boldsymbol{y}|} \tag{74}$$

where $H_y(t)\equiv\frac{\boldsymbol{y}^\top\boldsymbol{H}(t)\boldsymbol{y}}{|\boldsymbol{y}|^2}$. Plugging these we get:

$$\begin{aligned}
\frac{dH_y(t)}{dt} &= \frac{2\eta_0\gamma_0^2}{P}(y-f_y)f_y,\\
\frac{dG(t)}{dt} &= \frac{2\eta_0\gamma_0^2}{P}(y-f_y)f_y,\\
\frac{df_y(t)}{dt} &= \frac{\eta_0}{P}\left(H_y(t)+G(t)\right)(y-f_y),\\
H_y(0) &= 1, \qquad G(0)=1, \qquad f_y(0)=0.
\end{aligned} \tag{75}$$

where $y\equiv|\boldsymbol{y}|$. Note that since $\dot{H}_y(t)=\dot{G}(t)$ and $H_y(0)=G(0)$, $H_y(t)=G(t)$. Hence, we are left with

$$\begin{aligned}
\frac{dH_y(t)}{dt} &= \frac{2\eta_0\gamma_0^2}{P}(y-f_y)f_y,\\
\frac{df_y(t)}{dt} &= \frac{2\eta_0}{P}H_y(t)(y-f_y),\\
H_y(0) &= 1, \qquad f_y(0)=0.
\end{aligned} \tag{76}$$

These equations can further be reduced to a one dimensional system by noting that there exists a

conservation law:

$$L \equiv H_y(t)^2 - \gamma_0^2 f_y(t)^2,$$
$$\frac{dL}{dt} = 0, \quad L(0) = 1, \quad \implies \quad L(t) = 1. \tag{77}$$

As $t \to \infty$, we expect $\boldsymbol{f}(t) \to \boldsymbol{y}$ or $f_t(t) \to y$. Therefore, we see that

$$\lim_{t\to\infty} \boldsymbol{H}(t) = \mathbf{I} + \frac{1}{y^2}\left(\sqrt{1+\gamma_0^2} - 1\right)\boldsymbol{y}\boldsymbol{y}^\top. \tag{78}$$

Further,we can substitute $H_y(t) = \sqrt{1+\gamma_0^2 f_y(t)}$ to obtain

$$\frac{df_y(t)}{dt} = \frac{2\eta_0}{P}\sqrt{1+\gamma_0^2 f_y(t)}\,(y - f_y), \qquad f_y(0) = 0. \tag{79}$$

For $\gamma_0 \to 0$, this gives the lazy NTK limit dynamics as expected

$$\frac{df_y(t)}{dt} \approx \frac{2\eta_0}{P}(y - f_y), \quad f_y(0) = 0, \quad \implies \quad \boldsymbol{f}(t) = (1 - e^{-2\eta_0 t/P})\boldsymbol{y}. \tag{80}$$

See slides for simulations and other comments.

Acknowledgments

CP thanks Boris Hanin for invitation to teach at the Princeton Machine Learning Summer School in June 2023. CP also thanks GPT-4 for help with proofreading and editing the introduction. CP and BB thank Jacob Zavatone-Veth for discussions.

A Proof of Proposition 4.1

This section is heavily influenced by Jacob Zavatone-Veth's notes `https://jzv.io/assets/pdf/lecture_notes_on_nngp_from_mft.pdf`. We will consider the cumulant generating function of preactivations

$$\log Z \equiv \log\left\langle \exp\left(i \sum_{\ell=1}^{L}\sum_{\mu=1}^{P} \boldsymbol{b}_\mu^{(\ell)} \cdot \boldsymbol{h}_\mu^{(\ell)} \right)\right\rangle \tag{81}$$

and show that in the $N \to \infty$ limit with the assumed weight scalings, this is a cumulant generating function for Gaussian fields. Our derivation uses some standard but useful tricks from statistical mechanics.

To reduce the notational burden of the following calculations, we introduce the following defi-

nitions

$$\boldsymbol{W}^{(L)} \equiv (\boldsymbol{w}^{(L)})^{\top}, \qquad \phi_{\ell}(h) \equiv \begin{cases} \phi(h), & \ell = 1, \ldots, L-1 \\ \frac{h}{\sqrt{D}}, & \ell = 0 \end{cases} . \tag{82}$$

Also, note that we are performing averages with respect to weights at initialization. We would normaly denote this with notation $\boldsymbol{W}^{(\ell)}(0)$, but we are suppressing the time dependence for notational convenience.

We first start by noting that the average in (81) is with respect to weights at initialization, but we are interested in distribution of the preactivations. To get at preactivation statistics, we multiply Z by "1":

$$1 = \int d\boldsymbol{h}_{\mu}^{(\ell)} \delta \left(\boldsymbol{h}_{\mu}^{(\ell)} - \frac{1}{N^{a_\ell}} \boldsymbol{W}^{(\ell)} \phi_{\ell-1}(h_{\mu}^{(\ell-1)}) \right), \qquad \ell = 1, \ldots, L \tag{83}$$

meaning,

$$Z = \int \left[\prod_{\mu=1}^{P} \prod_{\ell=1}^{L} d\boldsymbol{h}_{\mu}^{(\ell)} \right] \exp \left(i \sum_{\ell=1}^{L} \sum_{\mu=1}^{P} \boldsymbol{b}_{\mu}^{(\ell)} \cdot \boldsymbol{h}_{\mu}^{(\ell)} \right) \left\langle \prod_{\mu=1}^{P} \prod_{\ell=1}^{L} \delta \left(\boldsymbol{h}_{\mu}^{(\ell)} - \frac{1}{N^{a_\ell}} \boldsymbol{W}^{(\ell)} \phi_{\ell-1}(h_{\mu}^{(\ell-1)}) \right) \right\rangle \tag{84}$$

If we could perform the weight averages over the delta functions, we would end up with the measure we are looking for. This turns out to be slightly complicated, but manageable as we will see. A helpful trick is the following:

$$\delta \left(\boldsymbol{h}_{\mu}^{(\ell)} - \frac{1}{N^{a_\ell}} \boldsymbol{W}^{(\ell)} \phi_{\ell-1}(h_{\mu}^{(\ell-1)}) \right) = \int \frac{d\hat{\boldsymbol{h}}_{\mu}^{(\ell)}}{(2\pi)^N} \exp \left(i \hat{\boldsymbol{h}}_{\mu}^{(\ell)} \cdot \left(\boldsymbol{h}_{\mu}^{(\ell)} - \frac{1}{N^{a_\ell}} \boldsymbol{W}^{(\ell)} \phi_{\ell-1}(h_{\mu}^{(\ell-1)}) \right) \right) \tag{85}$$

Inserting this into Z to replace the delta functions, we get

$$\begin{aligned} Z = \int \left[\prod_{\mu=1}^{P} \prod_{\ell=1}^{L} d\boldsymbol{h}_{\mu}^{(\ell)} \frac{d\hat{\boldsymbol{h}}_{\mu}^{(\ell)}}{(2\pi)^N} \right] \exp \left(i \sum_{\ell=1}^{L} \sum_{\mu=1}^{P} \left(\boldsymbol{b}_{\mu}^{(\ell)} + \hat{\boldsymbol{h}}_{\mu}^{(\ell)} \right) \cdot \boldsymbol{h}_{\mu}^{(\ell)} \right) \\ \times \prod_{\ell=1}^{L} \left\langle \prod_{\mu=1}^{P} \exp \left(-i \hat{\boldsymbol{h}}_{\mu}^{(\ell)} \cdot \frac{1}{N^{a_\ell}} \boldsymbol{W}^{(\ell)} \phi_{\ell-1}(\boldsymbol{h}_{\mu}^{(\ell-1)}) \right) \right\rangle_{\boldsymbol{W}^{(\ell)}} \end{aligned} \tag{86}$$

Here, we introduced the notation $\langle \cdot \rangle_{\boldsymbol{W}^{(\ell)}}$ for averages with respect to $\boldsymbol{W}^{(\ell)}$. Note that these averages

are Gaussian, separate in weights, and can be performed in closed form:

$$\begin{aligned}
&\left\langle \prod_{\mu=1}^{P} \exp\left(-i\hat{\boldsymbol{h}}_\mu^{(\ell)} \cdot \frac{1}{N^{a_\ell}} \boldsymbol{W}^{(\ell)} \phi_{\ell-1}(\boldsymbol{h}_\mu^{(\ell-1)})\right) \right\rangle_{\boldsymbol{W}^{(\ell)}} \\
&\quad = \prod_{i,j=1}^{N} \int \frac{dW_{ij}^{(\ell)}}{\sqrt{2\pi/N^{b_\ell}}} \exp\left(-\sum_{i,j=1}^{N} \frac{W_{ij}^{(\ell)}}{2/N^{b_\ell}} - i\frac{W_{ij}^{(\ell)}}{N^{a_\ell}} \sum_{\mu=1}^{P} \hat{h}_{\mu,i}^{(\ell)} \phi_{\ell-1}(h_{\mu,j}^{(\ell-1)})\right) \\
&\quad = \prod_{i,j=1}^{N} \exp\left(-\frac{1}{2N^{2a_\ell+b_\ell}} \sum_{\mu,\nu=1}^{P} \hat{h}_{\mu,i}^{(\ell)} \phi_{\ell-1}(h_{\mu,j}^{(\ell-1)}) \hat{h}_{\nu,i}^{(\ell)} \phi_{\ell-1}(h_{\nu,j}^{(\ell-1)})\right) \\
&\quad = \exp\left(-\frac{1}{2N^{2a_\ell+b_\ell}} \sum_{\mu,\nu=1}^{P} \hat{\boldsymbol{h}}_\mu^{(\ell)} \cdot \hat{\boldsymbol{h}}_\nu^{(\ell)} \phi_{\ell-1}(\boldsymbol{h}_\mu^{(\ell-1)}) \cdot \phi_{\ell-1}(\boldsymbol{h}_\nu^{(\ell-1)})\right)
\end{aligned} \tag{87}$$

Now we remember the constraints (16) and (20). Putting everything together, we get:

$$\begin{aligned}
Z = \int \left[\prod_{\mu=1}^{P}\prod_{\ell=1}^{L} d\boldsymbol{h}_\mu^{(\ell)} \frac{d\hat{\boldsymbol{h}}_\mu^{(\ell)}}{(2\pi)^N}\right] \exp\left(i\sum_{\ell=1}^{L}\sum_{\mu=1}^{P}\left(\boldsymbol{b}_\mu^{(\ell)} + \hat{\boldsymbol{h}}_\mu^{(\ell)}\right) \cdot \boldsymbol{h}_\mu^{(\ell)}\right) \\
\times \exp\left(-\frac{1}{2}\sum_{\mu,\nu=1}^{P} \hat{\boldsymbol{h}}_\mu^{(\ell)} \cdot \hat{\boldsymbol{h}}_\nu^{(\ell)} \left(\frac{1}{N^{1-\delta_{1,\ell}}} \phi_{\ell-1}(\boldsymbol{h}_\mu^{(\ell-1)}) \cdot \phi_{\ell-1}(\boldsymbol{h}_\nu^{(\ell-1)})\right)\right).
\end{aligned} \tag{88}$$

The feature kernels appeared in the above expression:

$$\begin{aligned}
\Phi_{\mu\nu}^{(\ell)} &\equiv \frac{1}{N}\phi(\boldsymbol{h}_\mu^{(\ell)}) \cdot \phi(\boldsymbol{h}_\nu^{(\ell)}), \qquad \ell = 1, \ldots, L \\
\Phi_{\mu\nu}^{(0)} &\equiv \frac{1}{D}\boldsymbol{x}_\mu \cdot \boldsymbol{x}_\nu.
\end{aligned} \tag{89}$$

We will now formally introduce them again using delta functions and their Fourier transforms, except $\Phi_{\mu\nu}^{(0)}$ which we can simply replace because it doesn't include any variables that are integrated over. The statements below are for $\ell = 1, \ldots, L$:

$$\begin{aligned}
1 &= \int d\Phi_{\mu\nu}^{(\ell)}\, \delta\left(\Phi_{\mu\nu}^{(\ell)} - \frac{1}{N}\phi(\boldsymbol{h}_\mu^{(\ell)}) \cdot \phi(\boldsymbol{h}_\nu^{(\ell)})\right), \\
&= \int d\Phi_{\mu\nu}^{(\ell)} \frac{d\hat{\Phi}_{\mu\nu}^{(\ell)}}{2\pi} \exp\left(i\hat{\Phi}_{\mu\nu}^{(\ell)}\left(\Phi_{\mu\nu}^{(\ell)} - \frac{1}{N}\phi(\boldsymbol{h}_\mu^{(\ell)}) \cdot \phi(\boldsymbol{h}_\nu^{(\ell)})\right)\right) \\
&= \int d\Phi_{\mu\nu}^{(\ell)} \frac{d\hat{\Phi}_{\mu\nu}^{(\ell)}}{4\pi/N} \exp\left(\frac{i}{2}\hat{\Phi}_{\mu\nu}^{(\ell)}\left(N\Phi_{\mu\nu}^{(\ell)} - \phi(\boldsymbol{h}_\mu^{(\ell)}) \cdot \phi(\boldsymbol{h}_\nu^{(\ell)})\right)\right).
\end{aligned} \tag{90}$$

This gives us

$$Z = \int \left[\prod_{\mu=1}^{P}\prod_{\nu=1}^{P}\prod_{\ell=1}^{L} d\Phi_{\mu\nu}^{(\ell)} \frac{d\hat{\Phi}_{\mu\nu}^{(\ell)}}{4\pi/N}\right] \left[\prod_{\mu=1}^{P}\prod_{\ell=1}^{L} d\boldsymbol{h}_{\mu}^{(\ell)} \frac{d\hat{\boldsymbol{h}}_{\mu}^{(\ell)}}{(2\pi)^N}\right] \exp\left(\frac{i}{2}\sum_{\mu,\nu=1}^{P} \hat{\Phi}_{\mu\nu}^{(\ell)}\left(N\Phi_{\mu\nu}^{(\ell)} - \phi(\boldsymbol{h}_{\mu}^{(\ell)})\cdot\phi(\boldsymbol{h}_{\nu}^{(\ell)})\right)\right)$$
$$\times \exp\left(i\sum_{\ell=1}^{L}\sum_{\mu=1}^{P}\left(\boldsymbol{b}_{\mu}^{(\ell)} + \hat{\boldsymbol{h}}_{\mu}^{(\ell)}\right)\cdot\boldsymbol{h}_{\mu}^{(\ell)} - \frac{1}{2}\sum_{\mu,\nu=1}^{P}\hat{\boldsymbol{h}}_{\mu}^{(\ell)}\cdot\hat{\boldsymbol{h}}_{\nu}^{(\ell)}\Phi_{\mu\nu}^{(\ell-1)}\right). \tag{91}$$

We rewrite this in the following form

$$Z = \int \left[\prod_{\mu=1}^{P}\prod_{\nu=1}^{P}\prod_{\ell=1}^{L} \frac{d\Phi_{\mu\nu}^{(\ell)} d\hat{\Phi}_{\mu\nu}^{(\ell)}}{4\pi/N}\right] e^{NS(\Phi,\hat{\Phi})}, \tag{92}$$

where

$$S(\Phi,\hat{\Phi}) = \frac{i}{2}\hat{\Phi}_{\mu\nu}^{(\ell)}\Phi_{\mu\nu}^{(\ell)} + \frac{1}{N}\sum_{l=1}^{L}\sum_{j=1}^{N}\log z_j^{(\ell)}\left(b_{\mu,j}^{(\ell)}, \boldsymbol{\Phi}^{(\ell-1)}, \hat{\boldsymbol{\Phi}}^{(\ell)}\right) \tag{93}$$

and

$$z_j^{(\ell)}\left(b_{\mu,j}^{(\ell)}, \boldsymbol{\Phi}^{(\ell-1)}, \hat{\boldsymbol{\Phi}}^{(\ell)}\right) \equiv \int \left[\prod_{\mu=1}^{P}\frac{dh_{\mu,j}^{(\ell)} d\hat{h}_{\mu,j}^{(\ell)}}{(2\pi)^N}\right] \exp\left(-\frac{i}{2}\sum_{\mu,\nu=1}^{P}\hat{\Phi}_{\mu\nu}^{(\ell)}\phi(h_{\mu,j}^{(\ell)})\phi(h_{\nu,j}^{(\ell)})\right)$$
$$\times \exp\left(-\frac{1}{2}\sum_{\mu,\nu=1}^{P}\hat{h}_{\mu,j}^{(\ell)}\hat{h}_{\nu,j}^{(\ell)}\Phi_{\mu\nu}^{(\ell-1)} + i\sum_{\mu=1}^{P}\left(b_{\mu,j}^{(\ell)} + \hat{h}_{\mu,j}^{(\ell)}\right)h_{\mu,j}^{(\ell)}\right). \tag{94}$$

We expect S to be $\mathcal{O}_N(1)$. Hence, in the large-N limit, we can evaluate Z using a saddle point approximation and

$$\log Z = NS(\Phi^*, \hat{\Phi}^*). \tag{95}$$

Saddle point equations are given by

$$\frac{\partial S}{\partial \Phi_{\mu\nu}^{(\ell)}} = 0, \qquad \frac{\partial S}{\partial \hat{\Phi}_{\mu\nu}^{(\ell)}} = 0, \qquad \ell = 1,\ldots,L, \qquad \mu,\nu = 1,\ldots,P. \tag{96}$$

Evaluating the first of these, we get

$$i\hat{\Phi}_{\mu\nu}^{(\ell)} = \frac{1}{N}\sum_{j}\left\langle \hat{h}_{\mu,j}^{(\ell+1)}\hat{h}_{\nu,j}^{(\ell+1)}\right\rangle_j, \; i\hat{\Phi}_{\mu\nu}^{(L)} = 0 \tag{97}$$

where we introduced

$$\langle\cdot\rangle_j \equiv \frac{1}{z_j^{(\ell)}\left(b_{\mu,j}^{(\ell)}, \mathbf{\Phi}^{(\ell-1)}, \hat{\mathbf{\Phi}}^{(\ell)}\right)} \int \left[\prod_{\mu=1}^{P} \frac{dh_{\mu,j}^{(\ell)} d\hat{h}_{\mu,j}^{(\ell)}}{(2\pi)^N}\right] (\cdot) \exp\left(-\frac{i}{2}\sum_{\mu,\nu=1}^{P} \hat{\Phi}_{\mu\nu}^{(\ell)} \phi(h_{\mu,j}^{(\ell)})\phi(h_{\nu,j}^{(\ell)})\right)$$
$$\times \exp\left(-\frac{1}{2}\sum_{\mu,\nu=1}^{P} \hat{h}_{\mu,j}^{(\ell)}\hat{h}_{\nu,j}^{(\ell)}\Phi_{\mu\nu}^{(\ell-1)} + i\sum_{\mu=1}^{P}\left(b_{\mu,j}^{(\ell)} + \hat{h}_{\mu,j}^{(\ell)}\right) h_{\mu,j}^{(\ell)}\right) \tag{98}$$

The second saddle point equation gives

$$\Phi_{\mu\nu}^{(\ell)} = \frac{1}{N}\sum_{j=1}^{N} \left\langle \phi(h_{\mu,j}^{(\ell)})\phi(h_{\nu,j}^{(\ell)})\right\rangle_j \tag{99}$$

A consistent solution to (97) is given by

$$\hat{\Phi}_{\mu\nu}^{(\ell)} = 0 \tag{100}$$

One can see this by the following argument. Suppose one is interested in calculating a moment that involves a finite number, say $k \sim \mathcal{O}_N(1)$, of hs (and their powers). We can achieve this only by turning on the corresponding bs in the moment generating function, and keeping the rest of the bs zero. This means, in the sum (97), there are $N-k$ averages, where (98) is evaluated with $b=0$. These terms are identical and equal to 0, because the h integral gives a $\delta(\hat{h})$ when $\hat{\Phi}_{\mu\nu}^{(\ell)} = 0$. The remaining k terms are $\mathcal{O}_N(1/N)$ and asymptote to zero, leading to a consistent solution.

Note that under this condition,

$$\begin{aligned} z_j^{(\ell)}\left(b_{\mu,j}^{(\ell)}, \mathbf{\Phi}^{(\ell-1)}, \mathbf{0}\right) &= \int \left[\prod_{\mu=1}^{P} \frac{dh_{\mu,j}^{(\ell)} d\hat{h}_{\mu,j}^{(\ell)}}{(2\pi)^N}\right] \exp\left(-\frac{1}{2}\sum_{\mu,\nu=1}^{P} \hat{h}_{\mu,j}^{(\ell)}\hat{h}_{\nu,j}^{(\ell)}\Phi_{\mu\nu}^{(\ell-1)} + i\sum_{\mu=1}^{P}\left(b_{\mu,j}^{(\ell)} + \hat{h}_{\mu,j}^{(\ell)}\right) h_{\mu,j}^{(\ell)}\right) \\ &= \int \left[\frac{\prod_{\mu=1}^{P} dh_{\mu,j}^{(\ell)}}{\sqrt{(2\pi)^N \det \mathbf{\Phi}^{(\ell-1)}}}\right] \exp\left(-\frac{1}{2}\sum_{\mu,\nu=1}^{P} h_{\mu,j}^{(\ell)}\left(\Phi^{(\ell-1)}\right)_{\mu\nu}^{-1} h_{\nu,j}^{(\ell)} + i\sum_{\mu=1}^{P} b_{\mu,j}^{(\ell)} h_{\mu,j}^{(\ell)}\right) \end{aligned} \tag{101}$$

which is the moment generating function of a multivariate Gaussian. Note that when $b_{\mu,j}^{(\ell)} = 0$, $h_{\mu,j}^{(\ell)}$ are identical in distribution across the j index. Then,

$$\log Z = \sum_{\ell=1}^{L}\sum_{j=1}^{N} \log z_j^{(\ell)}\left(b_{\mu,j}^{(\ell)}, \mathbf{\Phi}^{(\ell-1)}, \mathbf{0}\right) \tag{102}$$

What does this reveal?

$$h_{\mu,j}^{(\ell)} \sim \mathcal{N}\left(0, \Phi_{\mu\nu}^{(\ell-1)}\delta_{ij}\right) \tag{103}$$

and

$$\Phi^{(\ell)}_{\mu\nu} = \mathbb{E}_{h^{(l)}_{\mu} \sim \mathcal{N}\left(0, \Phi^{(\ell-1)}_{\mu\nu}\right)} \left[\phi(h^{(l)}_{\mu}) \phi(h^{(l)}_{\nu}) \right] \tag{104}$$

with initial condition

$$\Phi^{(0)}_{\mu\nu} = \frac{1}{D} \boldsymbol{x}_{\mu} \cdot \boldsymbol{x}_{\nu}. \tag{105}$$

References

[1] Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. In *Advances in Neural Information Processing Systems*, 2022.

[2] Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *arXiv preprint arXiv:2304.03408*, 2023.

[3] Nikhil Vyas, Alexander Atanasov, Blake Bordelon, Depen Morwani, Sabarish Sainathan, and Cengiz Pehlevan. Feature-learning networks are consistent across widths at realistic scales. *arXiv preprint arXiv:2305.18411*, 2023.

[4] Blake Bordelon and Cengiz Pehlevan. The influence of learning rule on representation dynamics in wide neural networks. In *International Conference on Learning Representations (ICLR)*, 2023.

[5] Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.

[6] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020.

[8] Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.

[9] Christopher KI Williams. Computing with infinite networks. *Advances in neural information processing systems*, pages 295–301, 1997.

[10] 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*, 2018.

[11] Kai Segadlo, Bastian Epping, Alexander van Meegen, David Dahmen, Michael Krämer, and Moritz Helias. Unified field theoretical approach to deep and recurrent neuronal networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(10):103401, 2022.

[12] Paul Cecil Martin, ED Siggia, and HA Rose. Statistical dynamics of classical systems. *Physical Review A*, 8(1):423, 1973.

[13] A Crisanti and H Sompolinsky. Path integral approach to random neural networks. *Physical Review E*, 98(6):062120, 2018.

[14] Moritz Helias and David Dahmen. *Statistical field theory for neural networks*, volume 970. Springer, 2020.

[15] Elisabeth Agoritsas, Giulio Biroli, Pierfrancesco Urbani, and Francesco Zamponi. Out-of-equilibrium dynamical mean-field equations for the perceptron model. *Journal of Physics A: Mathematical and Theoretical*, 51(8):085002, 2018.

[16] Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. *Advances in Neural Information Processing Systems*, 33:9540–9550, 2020.

[17] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580. Curran Associates, Inc., 2018.

[18] Blake Bordelon and Cengiz Pehlevan. Learning curves for stochastic gradient descent on structured features. *International Conference on Learning Representations*, 2022.

[19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

[20] Grant M Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *arXiv preprint arXiv:1805.00915*, 2018.