Self-Consistent Dynamical Field Theory of Kernel Evolution in Wide Neural Networks

Blake Bordelon & Cengiz Pehlevan
John Paulson School of Engineering and Applied Sciences, Center for Brain Science
Harvard University
Cambridge MA, 02138
blake_bordelon@g.harvard.edu, cpehlevan@g.harvard.edu

Abstract

We analyze feature learning in infinite-width neural networks trained with gradient flow through a self-consistent dynamical field theory. We construct a collection of deterministic dynamical order parameters which are inner-product kernels for hidden unit activations and gradients in each layer at pairs of time points, providing a reduced description of network activity through training. These kernel order parameters collectively define the hidden layer activation distribution, the evolution of the neural tangent kernel, and consequently output predictions. We show that the field theory derivation recovers the recursive stochastic process of infinite-width feature learning networks obtained from Yang & Hu with Tensor Programs [1]. For deep linear networks, these kernels satisfy a set of algebraic matrix equations. For nonlinear networks, we provide an alternating sampling procedure to self-consistently solve for the kernel order parameters. We provide comparisons of the self-consistent solution to various approximation schemes including the static NTK approximation, gradient independence assumption, and leading order perturbation theory, showing that each of these approximations can break down in regimes where general self-consistent solutions still provide an accurate description. Lastly, we provide experiments in more realistic settings which demonstrate that the loss and kernel dynamics of CNNs at fixed feature learning strength is preserved across different widths on a CIFAR classification task.

1 Introduction

Deep learning has emerged as a successful paradigm for solving challenging machine learning and computational problems across a variety of domains [2, 3]. However, theoretical understanding of the training and generalization of modern deep learning methods lags behind current practice. Ideally, a theory of deep learning would be analytically tractable, efficiently computable, capable of predicting network performance and internal features that the network learns, and interpretable through a reduced description involving desirably initialization-independent quantities.

Several recent theoretical advances have fruitfully considered the idealization of *wide neural networks*, where the number of hidden units in each layer is taken to be large. Under certain parameterization, Bayesian neural networks and gradient descent trained networks converge to gaussian processes (NNGPs) [4–6] and neural tangent kernel (NTK) machines [7–9] in their respective infinite-width limits. These limits provide both analytic tractability as well as detailed training and generalization analysis [10–17]. However, in this limit, with these parameterizations, data representations are fixed and do not adapt to data, termed the *lazy regime* of NN training, to contrast it from the *rich regime* where NNs significantly alter their internal features while fitting the data [18, 19]. The fact that the representation of data is fixed renders these kernel-based theories incapable of explaining feature

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

learning, an ingredient which is crucial to the success of deep learning in practice [20, 21]. Thus, alternative theories capable of modeling feature learning dynamics are needed.

Recently developed alternative parameterizations such as the mean field [22] and the μP [1] parameterizations allow feature learning in infinite-width NNs trained with gradient descent. Using the Tensor Programs framework, Yang & Hu identified a stochastic process that describes the evolution of preactivation features in infinite-width μP NNs [1]. In this work, we study an equivalent parameterization to μP with self-consistent dynamical mean field theory (DMFT) and recover the stochastic process description of infinite NNs using this alternative technique. In the same large width scaling, we include a scalar parameter γ_0 that allows smooth interpolation between lazy and rich behavior [18]. We provide a new computational procedure to sample this stochastic process and demonstrate its predictive power for wide NNs.

Our novel contributions in this paper are the following:

1. We develop a path integral formulation of gradient flow dynamics in infinite-width networks in the feature learning regime. Our parameterization includes a scalar parameter γ_0 to allow interpolation between rich and lazy regimes and comparison to perturbative methods.
2. Using a stationary action argument, we identify a set of saddle point equations that the kernels satisfy at infinite-width, relating the stochastic processes that define hidden activation evolution to the kernels and vice versa. We show that our saddle point equations recover at $\gamma_0 = 1$, from an alternative method, the same stochastic process obtained previously with Tensor Programs [1].
3. We develop a polynomial-time numerical procedure to solve the saddle point equations for deep networks. In numerical experiments, we demonstrate that solutions to these self-consistency equations are predictive of network training at a variety of feature learning strengths, widths and depths. We provide comparisons of our theory to various approximate methods, such as perturbation theory.

1.1 Related Works

A natural extension to the lazy NTK/NNGP limit that allows the study of feature learning is to calculate finite width corrections to the infinite-width limit. Finite width corrections to Bayesian inference in wide networks have been obtained with various perturbative [23–29] and self-consistent techniques [30–33]. In the gradient descent based setting, leading order corrections to the NTK dynamics have been analyzed to study finite width effects [34–36, 27]. These methods give approximate corrections which are accurate provided the strength of feature learning is small. In very rich feature learning regimes, however, the leading order corrections can give incorrect predictions [37, 38].

Another approach to study feature learning is to alter NN parameterization in gradient-based learning to allow significant feature evolution even at infinite-width, the *mean field* limit [22, 39]. Works on mean field NNs have yielded formal loss convergence results [40, 41] and shown equivalences of gradient flow dynamics to a partial differential equation (PDE) [42–44].

Our results are most closely related to a set of recent works which studied infinite-width NNs trained with gradient descent (GD) using the Tensor Programs (TP) framework [1]. We show that our discrete time field theory at unit feature learning strength $\gamma_0 = 1$ recovers the stochastic process which was derived from TP. The stochastic process derived from TP has provided insights into practical issues in NN training such as hyper-parameter search [45]. Computing the exact infinite-width limit of GD has exponential time requirements [1], which we show can be circumvented with an alternating sampling procedure. A projected variant of GD training has provided an infinite-width theory that could be scaled to realistic datasets like CIFAR-10 [46]. Inspired by Chizat and Bach's work on mechanisms of lazy and rich training [18], our theory interpolates between lazy and rich behavior in the mean field limit for varying γ_0 and allows comparison of DMFT to perturbative analysis near small γ_0. Further, our derivation of a DMFT action allows the possibility of pursuing finite width effects.

Our theory is inspired by self-consistent dynamical mean field theory (DMFT) from statistical physics [47–53]. This framework has been utilized in the theory of random recurrent networks [54–59], tensor PCA [60, 61], phase retrieval [62], and high-dimensional linear classifiers [63–66], but has yet to be developed for deep feature learning. By developing a self-consistent DMFT of deep NNs, we gain insight into how features evolve in the rich regime of network training, while retaining many pleasant analytic properties of the infinite-width limit.

2 Problem Setup and Definitions

Our theory applies to infinite-width networks, both fully-connected and convolutional. For notational ease we will relegate convolutional results to later sections. For input $\boldsymbol{x}_\mu \in \mathbb{R}^D$, we define the hidden *pre-activation* vectors $\boldsymbol{h}^\ell \in \mathbb{R}^N$ for layers $\ell \in \{1, ..., L\}$ as

$$f_\mu = \frac{1}{\gamma\sqrt{N}} \boldsymbol{w}^L \cdot \phi(\boldsymbol{h}^L_\mu) \ , \quad \boldsymbol{h}^{\ell+1}_\mu = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell \phi(\boldsymbol{h}^\ell_\mu) \ , \quad \boldsymbol{h}^1_\mu = \frac{1}{\sqrt{D}} \boldsymbol{W}^0 \boldsymbol{x}_\mu, \tag{1}$$

where $\boldsymbol{\theta} = \text{Vec}\{\boldsymbol{W}^0, ..., \boldsymbol{w}^L\}$ are the trainable parameters of the network and ϕ is a twice differentiable activation function. Inspired by previous works on the mechanisms of lazy gradient based training, the parameter γ will control the laziness or richness of the training dynamics [18, 19, 1, 42]. Each of the trainable parameters are initialized as Gaussian random variables with unit variance $W^\ell_{ij} \sim \mathcal{N}(0, 1)$. They evolve under gradient flow $\frac{d}{dt}\boldsymbol{\theta} = -\gamma^2 \nabla_{\boldsymbol{\theta}} \mathcal{L}$. The choice of learning rate γ^2 causes $\frac{d}{dt}\mathcal{L}|_{t=0}$ to be independent of γ. To characterize the evolution of weights, we introduce backpropagation variables $\boldsymbol{g}^\ell_\mu = \gamma\sqrt{N} \frac{\partial f_\mu}{\partial \boldsymbol{h}^\ell_\mu} = \dot{\phi}(\boldsymbol{h}^\ell_\mu) \odot \boldsymbol{z}^\ell_\mu$, where $\boldsymbol{z}^\ell_\mu = \frac{1}{\sqrt{N}} \boldsymbol{W}^{\ell\top} \boldsymbol{g}^{\ell+1}_\mu$ is the *pre-gradient* signal.

The relevant dynamical objects to characterize feature learning are feature and gradient kernels for each hidden layer $\ell \in \{1, ..., L\}$, defined as

$$\Phi^\ell_{\mu\alpha}(t, s) = \frac{1}{N} \phi(\boldsymbol{h}^\ell_\mu(t)) \cdot \phi(\boldsymbol{h}^\ell_\alpha(s)) \ , \quad G^\ell_{\mu\alpha}(t, s) = \frac{1}{N} \boldsymbol{g}^\ell_\mu(t) \cdot \boldsymbol{g}^\ell_\alpha(s). \tag{2}$$

From the kernels $\{\Phi^\ell, G^\ell\}_{\ell=1}^L$, we can compute the *Neural Tangent Kernel* $K^{NTK}_{\mu\alpha}(t, s) = \nabla_\theta f_\mu(t) \cdot \nabla_\theta f_\alpha(s) = \sum_{\ell=0}^L G^{\ell+1}_{\mu\alpha}(t, s) \Phi^\ell_{\mu\alpha}(t, s)$, [7] and the dynamics of the network function f_μ

$$\frac{d}{dt} f_\mu(t) = \sum_{\alpha=1}^P K^{NTK}_{\mu\alpha}(t, t) \Delta_\alpha(t) \ , \quad \Delta_\mu(t) = -\frac{\partial}{\partial f_\mu} \mathcal{L}|_{f_\mu(t)}, \tag{3}$$

where we define base cases $G^{L+1}_{\mu\alpha}(t, s) = 1, \Phi^0_{\mu\alpha}(t, s) = K^x_{\mu\alpha} = \frac{1}{D} \boldsymbol{x}_\mu \cdot \boldsymbol{x}_\alpha$. We note that the above formula holds for any data point μ which may or may not be in the set of P training examples. The above expressions demonstrate that knowledge of the temporal trajectory of the NTK on the $t = s$ diagonal gives the temporal trajectory of the network predictions $f_\mu(t)$.

Following prior works on infinite-width networks [22, 1, 40, 19], we study the mean field limit

$$N, \gamma \to \infty \ , \quad \gamma_0 = \frac{\gamma}{\sqrt{N}} = \mathcal{O}_N(1) \tag{4}$$

As we demonstrate in the Appendix D and N, this is the only N-scaling which allows feature learning as $N \to \infty$. The $\gamma_0 = 0$ limit recovers the static NTK limit [7]. We discuss other scalings and parameterizations in Appendix N, relating our work to the μP-parameterization and TP analysis of [1], showing they have identical feature dynamics in the infinite-width limit. We also analyze the effect of different hidden layer widths and initialization variances in the Appendix D.8. We focus on equal widths and NTK parameterization (as in eq. (1)) in the main text to reduce complexity.

3 Self-consistent DMFT

Next, we derive our self-consistent DMFT in a limit where $t, P = \mathcal{O}_N(1)$. Our goal is to build a description of training dynamics purely based on representations, and independent of weights. Studying feature learning at infinite-width enjoys several analytical properties:

- The kernel order parameters Φ^ℓ, G^ℓ concentrate over random initializations but are dynamical, allowing flexible adaptation of features to the task structure.
- In each layer ℓ, each neuron's preactivation h^ℓ_i and pregradient z^ℓ_i become i.i.d. draws from a distribution characterized by a set of order parameters $\{\Phi^\ell, G^\ell, A^\ell, B^\ell\}$.
- The kernels are defined as self-consistent averages (denoted by $\langle\rangle$) over this distribution of neurons in each layer $\Phi^\ell_{\mu\alpha}(t, s) = \left\langle \phi(h^\ell_\mu(t)) \phi(h^\ell_\alpha(s)) \right\rangle$ and $G^\ell_{\mu\alpha}(t, s) = \left\langle g^\ell_\mu(t) g^\ell_\alpha(s) \right\rangle$.

The next section derives these facts from a path-integral formulation of gradient flow dynamics.

3.1 Path Integral Construction

Gradient flow after a random initialization of weights defines a high dimensional stochastic process over initalizations for variables $\{\boldsymbol{h}, \boldsymbol{g}\}$. Therefore, we will utilize DMFT formalism to obtain a reduced description of network activity during training. For a simplified derivation of the DMFT for the two-layer ($L = 1$) case, see D.2. Generally, we separate the contribution on each forward/backward pass between the initial condition and gradient updates to weight matrix $\boldsymbol{W}^\ell$, defining new stochastic variables $\boldsymbol{\chi}^\ell, \boldsymbol{\xi}^\ell \in \mathbb{R}^N$ as

$$\boldsymbol{\chi}_\mu^{\ell+1}(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0)\phi(\boldsymbol{h}_\mu^\ell(t)) \ , \quad \boldsymbol{\xi}_\mu^\ell(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0)^\top \boldsymbol{g}_\mu^{\ell+1}(t). \tag{5}$$

We let Z represent the moment generating functional (MGF) for these stochastic fields

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

which requires, by construction the normalization condition $Z[\{\mathbf{0}, \mathbf{0}\}] = 1$. We enforce our definition of $\boldsymbol{\chi}, \boldsymbol{\xi}$ using an integral representation of the delta-function. Thus for each sample $\mu \in [P]$ and each time $t \in \mathbb{R}_+$, we multiply Z by

$$1 = \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \frac{d\boldsymbol{\chi}_\mu^{\ell+1}(t) d\hat{\boldsymbol{\chi}}_\mu^{\ell+1}(t)}{(2\pi)^N} \exp\left(i\hat{\boldsymbol{\chi}}_\mu^{\ell+1}(t) \cdot \left[\boldsymbol{\chi}_\mu^{\ell+1}(t) - \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0)\phi(\boldsymbol{h}_\mu^\ell(t)) \right] \right), \tag{6}$$

for $\boldsymbol{\chi}$ and the respective expression for $\boldsymbol{\xi}$. After making such substitutions, we perform integration over initial Gaussian weight matrices to arrive at an integral expression for Z, which we derive in the appendix D.4. We show that Z can be described by set of order-parameters $\{\Phi^\ell, \hat{\Phi}^\ell, G^\ell, \hat{G}^\ell, A^\ell, B^\ell\}$

$$Z[\{\boldsymbol{j}^\ell, \boldsymbol{v}^\ell\}] \propto \int \prod_{\ell\mu\alpha ts} d\Phi_{\mu\alpha}^\ell(t,s) d\hat{\Phi}_{\mu\alpha}^\ell(t,s) dG_{\mu\alpha}^\ell(t,s) d\hat{G}_{\mu\alpha}^\ell(t,s) dA_{\mu\alpha}^\ell(t,s) dB_{\mu\alpha}^\ell(t,s) \tag{7}$$

$$\times \exp\Big(NS[\{\Phi, \hat{\Phi}, G, \hat{G}, A, B, j, v\}] \Big),$$

$$S = \sum_{\ell\mu\alpha} \int_0^\infty dt \int_0^\infty ds \left[\Phi_{\mu\alpha}^\ell(t,s)\hat{\Phi}_{\mu\alpha}^\ell(t,s) + G_{\mu\alpha}^\ell(t,s)\hat{G}_{\mu\alpha}^\ell(t,s) - A_{\mu\alpha}^\ell(t,s)B_{\mu\alpha}^\ell(t,s) \right]$$

$$+ \ln \mathcal{Z}[\{\Phi, \hat{\Phi}, G, \hat{G}, A, B, j, v\}], \tag{8}$$

where S is the DMFT action and $\mathcal{Z}$ is a single-site MGF, which defines the distribution of fields $\{\chi^\ell, \xi^\ell\}$ over the neural population in each layer. The kernels A and B are related to the correlations between feedforward and feedback signals in the network. We provide a detailed formula for $\mathcal{Z}$ in the Appendix D.4 and show that it factorizes over different layers $\mathcal{Z} = \prod_{\ell=1}^L \mathcal{Z}_\ell$.

3.2 Deriving the DMFT Equations from the Path Integral Saddle Point

As $N \to \infty$, the moment-generating function Z is exponentially dominated by the saddle point of S. The equations that define this saddle point also define our DMFT. We thus identify the kernels that render S locally stationary ($\delta S = 0$). The most important equations are those which define $\{\Phi^\ell, G^\ell\}$

$$\frac{\delta S}{\delta \hat{\Phi}_{\mu\alpha}^\ell(t,s)} = \Phi_{\mu\alpha}^\ell(t,s) + \frac{1}{\mathcal{Z}} \frac{\delta \mathcal{Z}}{\delta \hat{\Phi}_{\mu\alpha}^\ell(t,s)} = \Phi_{\mu\alpha}^\ell(t,s) - \left\langle \phi(h_\mu^\ell(t))\phi(h_\alpha^\ell(s)) \right\rangle = 0,$$

$$\frac{\delta S}{\delta \hat{G}_{\mu\alpha}^\ell(t,s)} = G_{\mu\alpha}^\ell(t,s) + \frac{1}{\mathcal{Z}} \frac{\delta \mathcal{Z}}{\delta \hat{G}_{\mu\alpha}^\ell(t,s)} = G_{\mu\alpha}^\ell(t,s) - \left\langle g_\mu^\ell(t) g_\alpha^\ell(s) \right\rangle = 0, \tag{9}$$

where $\langle\rangle$ denotes an average over the stochastic process induced by $\mathcal{Z}$, which is defined below

$$\{u_\mu^\ell(t)\}_{\mu\in[P], t\in\mathbb{R}_+} \sim \mathcal{GP}(0, \boldsymbol{\Phi}^{\ell-1}) \ , \ \{r_\mu^\ell(t)\}_{\mu\in[P], t\in\mathbb{R}_+} \sim \mathcal{GP}(0, \boldsymbol{G}^{\ell+1}),$$

$$h_\mu^\ell(t) = u_\mu^\ell(t) + \gamma_0 \int_0^t ds \sum_{\alpha=1}^P \left[A_{\mu\alpha}^{\ell-1}(t,s) + \Delta_\alpha(s)\Phi_{\mu\alpha}^{\ell-1}(t,s) \right] z_\alpha^\ell(s)\dot{\phi}(h_\alpha^\ell(s)),$$

$$z_\mu^\ell(t) = r_\mu^\ell(t) + \gamma_0 \int_0^t ds \sum_{\alpha=1}^P \left[B_{\mu\alpha}^\ell(t,s) + \Delta_\alpha(s) G_{\mu\alpha}^{\ell+1}(t,s) \right] \phi(h_\alpha^\ell(s)), \tag{10}$$

where we define base cases $\Phi^0_{\mu\alpha}(t,s) = K^x_{\mu\alpha}$ and $G^{L+1}_{\mu\alpha}(t,s) = 1$, $A^0 = B^L = 0$. We see that the fields $\{h^\ell, z^\ell\}$, which represent the single site preactivations and pre-gradients, are implicit functionals of the mean-zero Gaussian processes $\{u^\ell, r^\ell\}$ which have covariances $\left\langle u^\ell_\mu(t) u^\ell_\alpha(s)\right\rangle = \Phi^{\ell-1}_{\mu\alpha}(t,s)$ and $\left\langle r^\ell_\mu(t) r^\ell_\alpha(s)\right\rangle = G^{\ell+1}_{\mu\alpha}(t,s)$. The other saddle point equations give $A^\ell_{\mu\alpha}(t,s) = \gamma_0^{-1}\left\langle \frac{\delta\phi(h^\ell_\mu(t))}{\delta r^\ell_\alpha(s)}\right\rangle$, $B^\ell_{\mu\alpha}(t,s) = \gamma_0^{-1}\left\langle \frac{\delta g^{\ell+1}_\mu(t)}{\delta u^{\ell+1}_\alpha(s)}\right\rangle$ which arise due to coupling between the feedforward and feedback signals. We note that, in the lazy limit $\gamma_0 \to 0$, the fields approach Gaussian processes $h^\ell \to u^\ell$, $z^\ell \to r^\ell$. Lastly, the final saddle point equations $\frac{\delta S}{\delta \Phi^\ell} = 0$, $\frac{\delta S}{\delta G^\ell} = 0$ imply that $\hat{\Phi}^\ell = \hat{G}^\ell = 0$. The full set of equations that define the DMFT are given in D.7.

This theory is easily extended to more general architectures such as networks with varying widths by layer (App. D.8), trainable bias parameter (App. H), multiple (but $\mathcal{O}_N(1)$) output channels (App. I), convolutional architectures (App. G), networks trained with weight decay (App. J), Langevin sampling (App. K) and momentum (App. L), discrete time training (App. M). In Appendix N, we discuss parameterizations which give equivalent feature and predictor dynamics and show our derived stochastic process is equivalent to the μP scheme of Yang & Hu [1].

4 Solving the Self-Consistent DMFT

The saddle point equations obtained from the field theory discussed in the previous section must be solved self-consistently. By this we mean that, given knowledge of the kernels, we can characterize the distribution of $\{h^\ell, z^\ell\}$, and given the distribution of $\{h^\ell, z^\ell\}$, we can compute the kernels [67, 64]. In the Appendix B, we provide Algorithm 1, a numerical procedure based on this idea to efficiently solve for the kernels with an alternating Monte-Carlo strategy. The output of the algorithm are the dynamical kernels $\Phi^\ell_{\mu\alpha}(t,s), G^\ell_{\mu\alpha}(t,s), A^\ell_{\mu\alpha}(t,s), B^\ell_{\mu\alpha}(t,s)$, from which any network observable can be computed as we discuss in Appendix D. We provide an example of the solution to the saddle point equations compared to training a finite NN in Figure 1. We plot Φ^ℓ, G^ℓ at the end of training and the sample-trace of these kernels through time. Additionally, we compare the kernels of finite width N network to the DMFT predicted kernels using a cosine-similarity alignment metric $A(\boldsymbol{\Phi}^{DMFT}, \boldsymbol{\Phi}^{NN}) = \frac{\text{Tr}\, \boldsymbol{\Phi}^{DMFT}\boldsymbol{\Phi}^{NN}}{|\boldsymbol{\Phi}^{DMFT}||\boldsymbol{\Phi}^{NN}|}$. Additional examples are in Appendix Figures 6 and Figure 7.

4.1 Deep Linear Networks: Closed Form Self-Consistent Equations

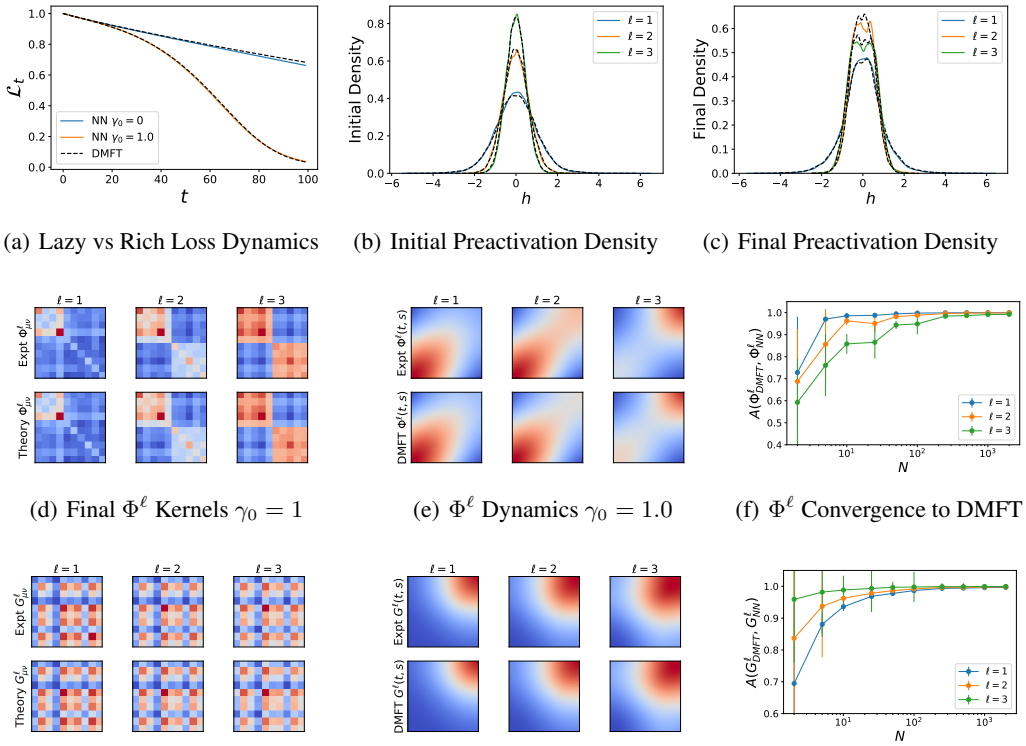
Deep linear networks ($\phi(h) = h$) are of theoretical interest since they are simpler to analyze than nonlinear networks but preserve non-trivial training dynamics and feature learning [68–72, 25, 32, 23]. In a deep linear network, we can simplify our saddle point equations to algebraic formulas that close in terms of the kernels $H^\ell_{\mu\alpha}(t,s) = \left\langle h^\ell_\mu(t) h^\ell_\alpha(s)\right\rangle$, $G^\ell(t,s) = \left\langle g^\ell(t) g^\ell(s)\right\rangle$ [1]. This is a significant simplification since it allows solution of the saddle point equations without a sampling procedure.

To describe the result, we first introduce a vectorization notation $\boldsymbol{h}^\ell = \text{Vec}\{h^\ell_\mu(t)\}_{\mu\in[P], t\in\mathbb{R}_+}$. Likewise we convert kernels $\boldsymbol{H}^\ell = \text{Mat}\{H^\ell_{\mu\alpha}(t,s)\}_{\mu,\alpha\in[P], t,s\in\mathbb{R}_+}$ into matrices. The inner product under this vectorization is defined as $\boldsymbol{a}\cdot\boldsymbol{b} = \int_0^\infty dt \sum_{\mu=1}^P a_\mu(t) b_\mu(t)$. In a practical computational implementation, the theory would be evaluated on a grid of T time points with discrete time gradient descent, so these kernels $\boldsymbol{H}^\ell \in \mathbb{R}^{PT\times PT}$ would indeed be matrices of the appropriate size. The fields $\boldsymbol{h}^\ell, \boldsymbol{g}^\ell$ are linear functionals of independent Gaussian processes $\boldsymbol{u}^\ell, \boldsymbol{r}^\ell$, giving $(\mathbf{I} - \gamma_0^2 \boldsymbol{C}^\ell \boldsymbol{D}^\ell)\boldsymbol{h}^\ell = \boldsymbol{u}^\ell + \gamma_0 \boldsymbol{C}^\ell \boldsymbol{r}^\ell$, $(\mathbf{I} - \gamma_0^2 \boldsymbol{D}^\ell \boldsymbol{C}^\ell)\boldsymbol{g}^\ell = \boldsymbol{r}^\ell + \gamma_0 \boldsymbol{D}^\ell \boldsymbol{u}^\ell$. The matrices $\boldsymbol{C}^\ell$ and $\boldsymbol{D}^\ell$ are causal integral operators which depend on $\{\boldsymbol{A}^{\ell-1}, \boldsymbol{H}^{\ell-1}\}$ and $\{\boldsymbol{B}^\ell, \boldsymbol{G}^{\ell+1}\}$ respectively which we define in Appendix F. The saddle point equations which define the kernels are

$$\boldsymbol{H}^\ell = \left\langle \boldsymbol{h}^\ell \boldsymbol{h}^{\ell\top}\right\rangle = (\mathbf{I} - \gamma_0^2 \boldsymbol{C}^\ell \boldsymbol{D}^\ell)^{-1}[\boldsymbol{H}^{\ell-1} + \gamma_0^2 \boldsymbol{C}^\ell \boldsymbol{G}^{\ell+1} \boldsymbol{C}^{\ell\top}]\left[(\mathbf{I} - \gamma_0^2 \boldsymbol{C}^\ell \boldsymbol{D}^\ell)^{-1}\right]^\top$$
$$\boldsymbol{G}^\ell = \left\langle \boldsymbol{g}^\ell \boldsymbol{g}^{\ell\top}\right\rangle = \left(\mathbf{I} - \gamma_0^2 \boldsymbol{D}^\ell \boldsymbol{C}^\ell\right)^{-1}\left[\boldsymbol{G}^{\ell+1} + \gamma_0^2 \boldsymbol{D}^\ell \boldsymbol{H}^{\ell-1} \boldsymbol{D}^{\ell\top}\right]\left[\left(\mathbf{I} - \gamma_0^2 \boldsymbol{D}^\ell \boldsymbol{C}^\ell\right)^{-1}\right]^\top . \quad (11)$$

Examples of the predictions obtained by solving these systems of equations are provided in Figure 2. We see that these DMFT equations describe kernel evolution for networks of a variety of depths and that the change in each layer's kernel increases with the depth of the network.

Unlike many prior results [68–71], our DMFT does not require any restrictions on the structure of the input data but hold for any $\boldsymbol{K}^x, \boldsymbol{y}$. However, for whitened data $\boldsymbol{K}^x = \mathbf{I}$ we show in



(a) Lazy vs Rich Loss Dynamics (b) Initial Preactivation Density (c) Final Preactivation Density

(d) Final Φ^ℓ Kernels $\gamma_0 = 1$ (e) Φ^ℓ Dynamics $\gamma_0 = 1.0$ (f) Φ^ℓ Convergence to DMFT

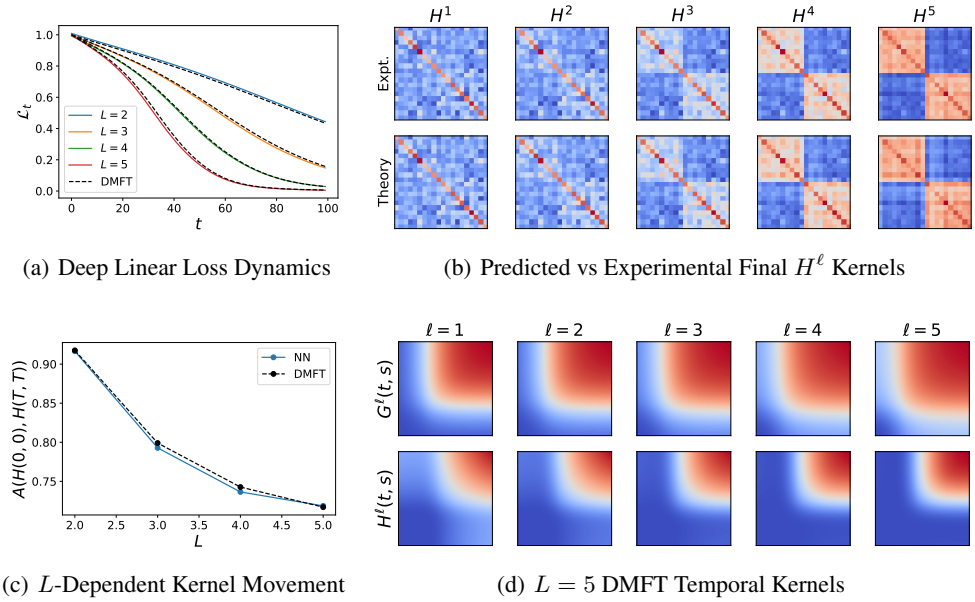
(g) Final G^ℓ kernels $\gamma_0 = 1.0$ (h) G^ℓ Dynamics $\gamma_0 = 1.0$ (i) G^ℓ Convergence to DMFT

Figure 1: Neural network feature learning dynamics is captured by self-consistent dynamical mean field theory (DMFT). (a) Training loss curves on a subsample of $P = 10$ CIFAR-10 training points in a depth 4 ($L = 3$, $N = 2500$) tanh network ($\phi(h) = \tanh(h)$) trained with MSE. Increasing γ_0 accelerates training. (b)-(c) The distribution of preactivations at the beginning and end of training matches predictions of the DMFT. (d) The final Φ^ℓ (at $t = 100$) kernel order parameters match the finite width network. (e) The temporal dynamics of the sample-traced kernels $\sum_\mu \Phi^\ell_{\mu\mu}(t, s)$ matches experiment and reveals rich dynamics across layers. (f) The alignment $A(\boldsymbol{\Phi}^\ell_{DMFT}, \boldsymbol{\Phi}^\ell_{NN})$, defined as cosine similarity, of the kernel $\Phi^\ell_{\mu\alpha}(t, s)$ predicted by theory (DMFT) and width N networks for different N but fixed $\gamma_0 = \gamma/\sqrt{N}$. Errorbars show standard deviation computed over 10 repeats. Around $N \sim 500$ DMFT begins to show near perfect agreement with the NN. (g)-(i) The same plots but for the gradient kernel $\boldsymbol{G}^\ell$. Whereas finite width effects for $\boldsymbol{\Phi}^\ell$ are larger at later layers ℓ since variance accumulates on the forward pass, fluctuations in $\boldsymbol{G}^\ell$ are large in early layers.

Appendix F.1.1, F.2 that our DMFT learning curves interpolate between NTK dynamics and the sigmoidal trajectories of prior works [68, 69] as γ_0 is increased. For example, in the two layer ($L = 1$) linear network with $\boldsymbol{K}^x = \mathbf{I}$, the dynamics of the error norm $\Delta(t) = ||\boldsymbol{\Delta}(t)||$ takes the form $\frac{\partial}{\partial t}\Delta(t) = -2\sqrt{1 + \gamma_0^2(y - \Delta(t))^2}\Delta(t)$ where $y = ||\boldsymbol{y}||$. These dynamics give the linear convergence rate of the NTK if $\gamma_0 \to 0$ but approaches logistic dynamics of [69] as $\gamma_0 \to \infty$. Further, $\boldsymbol{H}(t) = \left\langle \boldsymbol{h}^1(t)\boldsymbol{h}^1(t)^\top \right\rangle \in \mathbb{R}^{P\times P}$ only grows in the $\boldsymbol{y}\boldsymbol{y}^\top$ direction with $H_y(t) = \frac{1}{y^2}\boldsymbol{y}^\top \boldsymbol{H}(t)\boldsymbol{y} = \sqrt{1 + \gamma_0^2(y - \Delta(t))^2}$. At the end of training $\boldsymbol{H}(t) \to \mathbf{I} + \frac{1}{y^2}[\sqrt{1 + \gamma_0^2 y^2} - 1]\boldsymbol{y}\boldsymbol{y}^\top$, recovering the rank one spike which was recently obtained in the small initialization limit [73]. We show this one dimensional system in Figure 8.

4.2 Feature Learning with L2 Regularization

As we show in Appendix J, the DMFT can be extended to networks trained with weight decay $\frac{d\boldsymbol{\theta}}{dt} = -\gamma^2 \nabla_{\boldsymbol{\theta}} \mathcal{L} - \lambda \boldsymbol{\theta}$. If neural network is homogenous in its parameters so that $f(c\boldsymbol{\theta}) = c^\kappa f(\boldsymbol{\theta})$ (examples include networks with linear, ReLU, quadratic activations), then the final network predictor



(a) Deep Linear Loss Dynamics

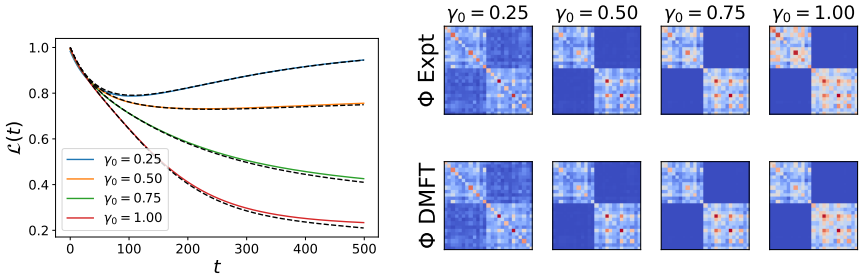
(b) Predicted vs Experimental Final H^ℓ Kernels

(c) L-Dependent Kernel Movement

(d) $L = 5$ DMFT Temporal Kernels

Figure 2: Deep linear network with the full DMFT. (a) The train loss for NNs of varying L. (b) For a $L = 5, N = 1000$ NN, the kernels H^ℓ at the end of training compared to DMFT theory on $P = 20$ datapoints. (c) The average displacement of feature kernels for different depth networks at same γ_0 value. For equal values of γ_0, deeper networks exhibit larger changes to their features, manifested in lower alignment with their initial $t = 0$ kernels $\boldsymbol{H}$. (d) The solution to the temporal components of the $G^\ell(t, s)$ and $\sum_\mu H^\ell_{\mu\mu}(t, s)$ kernels obtained from the self-consistent equations.

is a kernel regressor with the final NTK $\lim_{t\to\infty} f(\boldsymbol{x}, t) = \boldsymbol{k}(\boldsymbol{x})^\top [\boldsymbol{K} + \lambda\kappa \mathbf{I}]^{-1} \boldsymbol{y}$ where $K(\boldsymbol{x}, \boldsymbol{x}')$ is the *final*-NTK, $[\boldsymbol{k}(\boldsymbol{x})]_\mu = K(\boldsymbol{x}, \boldsymbol{x}_\mu)$ and $[\boldsymbol{K}]_{\mu\alpha} = K(\boldsymbol{x}_\mu, \boldsymbol{x}_\alpha)$. We note that the effective regularization $\lambda\kappa$ increases with depth L. In NTK parameterization, weight decay in infinite width homogenous networks gives a trivial fixed point $K(\boldsymbol{x}, \boldsymbol{x}') \to 0$ and consequently a zero predictor $f \to 0$ [74]. However, as we show in Figure 3, increasing feature learning γ_0 can prevent convergence to the trivial fixed point, allowing a non-zero fixed point for K, f even at infinite width. The kernel and function dynamics can be predicted with DMFT. The fixed point is a nontrivial function of the hyperparameters $\lambda, \kappa, L, \gamma_0$.



(a) Loss for varying γ_0

(b) Final Φ Kernels

Figure 3: Width $N = 1000$ ReLU networks trained with L2 regularization have nontrivial fixed point in DMFT limit ($\gamma_0 > 0$). (a) Training loss dynamics for a $L = 1$ ReLU network with $\lambda = 1$. In $\gamma_0 \to 0$ limit the fixed point is trivial $f = K = 0$. The final loss is a decreasing function of γ_0. (b) The final kernel is more aligned with target with increasing γ_0. Networks with homogenous activations enjoy a representer theorem at infinite-width as we show in Appendix J.

5 Approximation Schemes

We now compare our exact DMFT with approximations of prior works, providing an explanation of when these approximations give accurate predictions and when they break down.

5.1 Gradient Independence Ansatz

We can study the accuracy of the ansatz $\boldsymbol{A}^\ell = \boldsymbol{B}^\ell = 0$, which is equivalent to treating the weight matrices $\boldsymbol{W}^\ell(0)$ and $\boldsymbol{W}^\ell(0)^\top$ which appear in forward and backward passes respectively as independent Gaussian matrices. This assumption was utilized in prior works on signal propagation in deep networks in the lazy regime [75–79]. A consequence of this approximation is the Gaussianity and statistical independence of χ^ℓ and ξ^ℓ (conditional on $\{\boldsymbol{\Phi}^\ell, \boldsymbol{G}^\ell\}$) in each layer as we show in Appendix O. This ansatz works very well near $\gamma_0 \approx 0$ (the static kernel regime) since $\frac{d\boldsymbol{h}}{d\boldsymbol{r}}, \frac{d\boldsymbol{z}}{d\boldsymbol{u}} \sim \mathcal{O}(\gamma_0)$ or around initialization $t \approx 0$ but begins to fail at larger values of γ_0, t (Figure 4).

5.2 Perturbation theory in γ_0 at infinite-width

In the $\gamma_0 \to 0$ limit, we recover static kernels, giving linear dynamics identical to the NTK limit [7]. Corrections to this lazy limit can be extracted at small but finite γ_0. This is conceptually similar to recent works which consider perturbation series for the NTK in powers of $1/N$ [35, 27, 28] (though not identical, see Appendix P.7 for finite N effects). We expand all observables $q(\gamma_0)$ in a power series in γ_0, giving $q(\gamma_0) = q^{(0)} + \gamma_0 q^{(1)} + \gamma_0^2 q^{(2)} + ...$ and compute corrections up to $\mathcal{O}(\gamma_0^2)$. We show that the $\mathcal{O}(\gamma_0)$ and $\mathcal{O}(\gamma_0^3)$ corrections to kernels vanish, giving leading order expansions of the form $\boldsymbol{\Phi} = \boldsymbol{\Phi}^0 + \gamma_0^2 \boldsymbol{\Phi}^2 + \mathcal{O}(\gamma_0^4)$ and $\boldsymbol{G} = \boldsymbol{G}^0 + \gamma_0^2 \boldsymbol{G}^2 + \mathcal{O}(\gamma_0^4)$ (see Appendix P.2). Further, we show that the NTK has relative change at leading order which scales linearly with depth $|\Delta K^{NTK}|/|K^{NTK,0}| \sim \mathcal{O}_{\gamma_0,L}(L\gamma_0^2) = \mathcal{O}_{N,\gamma,L}(\frac{\gamma^2 L}{N})$, which is consistent with finite width effective field theory at $\gamma = \mathcal{O}_N(1)$ [26–28] (Appendix P.6). Further, at the leading order correction, all temporal dependencies are controlled by $P(P+1)$ functions $v_\alpha(t) = \int_0^t ds \Delta_\alpha^0(s)$ and $v_{\alpha\beta}(t) = \int_0^t ds \Delta_\alpha^0(s) \int_0^s ds' \Delta_\beta^0(s')$, which is consistent with those derived for finite width NNs using a truncation of the Neural Tangent Hierarchy [34, 35, 27]. To lighten notation, we focus our main text comparison of our non-perturbative DMFT to perturbation theory in the deep linear case. Full perturbation theory is in Appendix P.2.

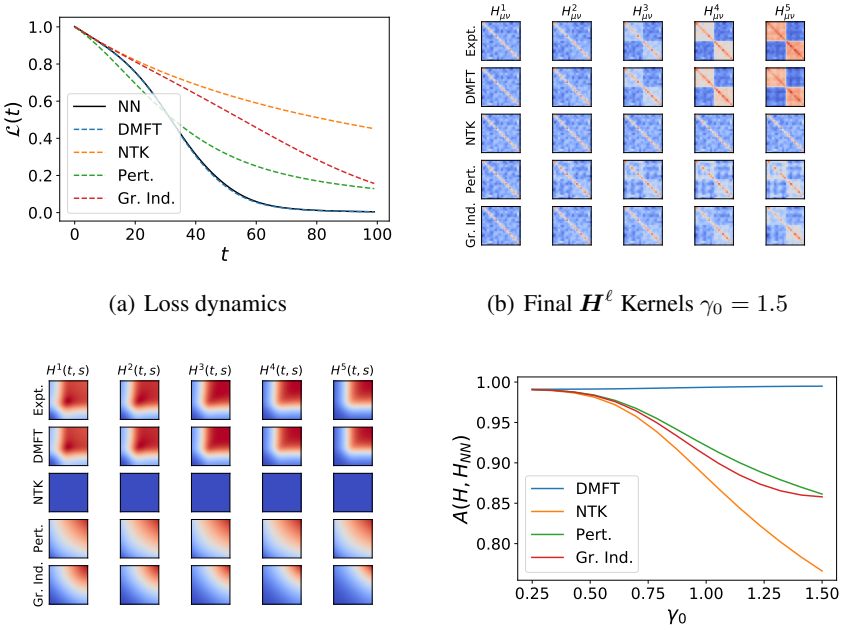
Using the timescales derived in the previous section, we find that the leading order correction to the kernels in infinite-width deep linear network have the form

$$K_{\mu\nu}^{NTK}(t,s) = (L+1)K_{\mu\nu}^x + \gamma_0^2 \frac{L(L+1)}{2} K_{\mu\nu}^x \sum_{\alpha\beta} K_{\alpha\beta}^x [v_{\alpha\beta}(t) + v_{\beta\alpha}(s) + v_\alpha(t) v_\beta(s)]$$
$$+ \gamma_0^2 \frac{L(L+1)}{2} \left[\sum_{\alpha\beta} K_{\mu\alpha}^x K_{\nu\beta}^x [v_{\alpha\beta}(t) + v_{\beta\alpha}(s)] + \sum_{\alpha\beta} K_{\mu\alpha}^x K_{\nu\beta}^x v_\alpha(t) v_\beta(s) \right] + \mathcal{O}(\gamma_0^4). \quad (12)$$

We see that the relative change in the NTK $|\boldsymbol{K}^{NTK} - \boldsymbol{K}^{NTK}(0)|/|\boldsymbol{K}^{NTK}(0)| \sim \mathcal{O}(\gamma_0^2 L) = \mathcal{O}(\gamma^2 L/N)$, so that large depth L networks exhibit more significant kernel evolution, which agrees with other perturbative studies [35, 27, 25] as well as the non-perturbative results in Figure 2. However at large γ_0 and large L, this theory begins to break down as we show in Figure 4.

6 Feature Learning Dynamics is Preserved at Fixed γ_0

Our DMFT suggests that for networks sufficiently wide for their kernels to concentrate, the dynamics of loss and kernels should be invariant under the rescaling $N \to RN, \gamma \to \gamma/\sqrt{R}$, which keeps γ_0 fixed. To evaluate how well this idea holds in a realistic deep learning problem, we trained CNNs of varying channel counts N on two-class CIFAR classification [80]. We tracked the dynamics of the loss and the last layer Φ^L kernel. The results are provided in Figure 5. We see that dynamics are largely independent of rescaling as predicted. Further, as expected, larger γ_0 leads to larger changes in kernel norm and faster alignment to the target function y, as was also found in [81]. Consequently, the higher γ_0 networks train more rapidly. The trend is consistent for width $N = 250$ and $N = 500$. More details about the experiment can be found in Appendix C.2.



(a) Loss dynamics

(b) Final $\boldsymbol{H}^\ell$ Kernels $\gamma_0 = 1.5$

(c) $\boldsymbol{H}^\ell$ Kernel Dynamics $\gamma_0 = 1.5$

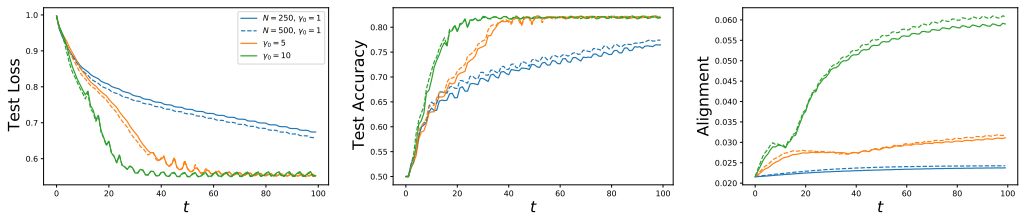
(d) Theory $\boldsymbol{H}^\ell$ vs NN with $N = 1000$

Figure 4: Comparison of DMFT to various approximation schemes in a $L = 5$ hidden layer, width $N = 1000$ linear network with $\gamma_0 = 1.0$ and $P = 100$. (a) The loss for the various approximations do not track the true trajectory induced by gradient descent in the large γ_0 regime. (b)-(c) The feature kernels $H^\ell_{\mu\alpha}(t, s)$ across each of the $L = 5$ hidden layers for each of the theories is compared to a width 1000 neural network. Again, we plot the sample-traced dynamics $\sum_{\mu\mu} H^\ell_{\mu\mu}(t, s)$. (d) The alignment of $\boldsymbol{H}^\ell$ compared to the finite NN $A(\boldsymbol{H}^\ell, \boldsymbol{H}^\ell_{NN})$ averaged across $\ell \in \{1, ..., 5\}$ for varying γ. The predictions of all of these theories coincide in the $\gamma_0 = 0$ limit but begin to deviate in the feature learning regime. Only the non-perturbative DMFT is accurate over a wide range of γ_0.

7 Discussion

We provided a unifying DMFT derivation of feature dynamics in infinite networks trained with gradient based optimization. Our theory interpolates between lazy infinite-width behavior of a static NTK in $\gamma_0 \to 0$ and rich feature learning. At $\gamma_0 = 1$, our DMFT construction agrees with the stochastic process derived previously with the Tensor Programs framework [1]. Our saddle point equations give self-consistency conditions which relate the stochastic fields to the kernels. These equations are exactly solveable in deep linear networks and can be efficiently solved with a numerical method in the nonlinear case. Comparisons with other approximation schemes show that DMFT can be accurate at a much wider range of γ_0. We believe our framework could be a useful perspective for future theoretical analyses of feature learning and generalization in wide networks.

Though our DMFT is quite general in regards to the data and architecture, the technique is not entirely rigorous and relies on heuristic physics techniques. Our theory holds in the $T, P = \mathcal{O}_N(1)$ and may break down otherwise; other asymptotic regimes (such as $P/N, T/\log(N) = \mathcal{O}_N(1)$, etc) may exhibit phenomena relevant to deep learning practice [32, 82]. The computational requirements of our method, while smaller than the exponential time complexity for exact solution [1], are still significant for large PT. In Table 1, we compare the time taken for various theories to compute the feature kernels throughout T steps of gradient descent. For a width N network, computation of each forward pass on all P data points takes $\mathcal{O}(PN^2)$ computations. The static NTK requires computation of $\mathcal{O}(P^2)$ entries in the kernel which do not need to be recomputed. However, the DMFT requires matrix multiplications on $PT \times PT$ matrices giving a $\mathcal{O}(P^3T^3)$ time scaling. Future work could aim to improve the computational overhead of the algorithm, by considering data averaged theories [64] or one pass SGD [1]. Alternative projected versions of gradient descent have also enabled much



(a) Test MSE (b) Classification Error (c) $A(\Phi^L, yy^\top)$ Dynamics

Figure 5: The dynamics of a depth 5 ($L = 4$ hidden) CNNs trained on first two classes of CIFAR (boat vs plane) exhibit consistency for different channel counts $N \in \{250, 500\}$ for fixed $\gamma_0 = \gamma/\sqrt{N}$. (a) We plot the test loss (MSE) and (b) test classification error. Networks with higher γ_0 train more rapidly. Time is measured in every 100 update steps. (c) The dynamics of the last layer feature kernel Φ^L, shown as alignment to the target function. As predicted by the DMFT, higher γ_0 corresponds to more active kernel evolution, evidenced by larger change in the alignment.

better computational scaling in evaluation of the theoretical predictions [46], allowing evaluation on full CIFAR-10.

Requirements	Width-N NN	Static NTK	Perturbative	Full DMFT
Memory for Kernels	$\mathcal{O}(N^2)$	$\mathcal{O}(P^2)$	$\mathcal{O}(P^4T)$	$\mathcal{O}(P^2T^2)$
Time for Kernels	$\mathcal{O}(PN^2T)$	$\mathcal{O}(P^2)$	$\mathcal{O}(P^4T)$	$\mathcal{O}(P^3T^3)$
Time for Final Outputs	$\mathcal{O}(PN^2T)$	$\mathcal{O}(P^3)$	$\mathcal{O}(P^4)$	$\mathcal{O}(P^3T^3)$

Table 1: Computational requirements to compute kernel dynamics and trained network predictions on P points in a depth N neural network on a grid of T time points trained with P data points for various theories. DMFT is faster and less memory intensive than a width N network only if $N \gg PT$. It is more computationally efficient to compute full DMFT kernels than leading order perturbation theory when $T \ll \sqrt{P}$. The expensive scaling with both samples and time are the cost of a full-batch non-perturbative theory of gradient based feature learning dynamics.

Acknowledgments and Disclosure of Funding

This work was supported by NSF grant DMS-2134157 and an award from the Harvard Data Science Initiative Competitive Research Fund. BB acknowledges additional support from the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (award #1764269) and the Harvard Q-Bio Initiative.

BB thanks Jacob Zavatone-Veth, Alex Atanasov, Abdulkadir Canatar, and Ben Ruben for comments on this manuscript as well as Greg Yang, Boris Hanin, Yasaman Bahri, and Jascha Sohl-Dickstein for useful discussions.

References

[1] 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.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[4] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[5] Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

[6] Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

[7] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 8571–8580. Curran Associates, Inc., 2018.

[8] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

[9] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[11] B. Bordelon, A. Canatar, and C. Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. *International Conference of Machine Learning*, 2020.

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

[13] Omry Cohen, Or Malka, and Zohar Ringel. Learning curves for overparametrized deep neural networks: A field theory perspective. *Physical Review Research*, 3(2):023034, 2021.

[14] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. *Advances in Neural Information Processing Systems*, 33:15568–15578, 2020.

[15] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Learning curves of generic features maps for realistic datasets with a teacher-student model. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[16] James B Simon, Madeline Dickens, and Michael R DeWeese. Neural tangent kernel eigenvalues accurately predict generalization. *arXiv preprint arXiv:2110.03922*, 2021.

[17] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

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

[19] 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.

[20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[21] 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*, pages 770–778, 2016.

[22] 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.

[23] Laurence Aitchison. Why bigger is not always better: on finite and infinite neural networks. In *International Conference on Machine Learning*, pages 156–164. PMLR, 2020.

[24] Sho Yaida. Non-gaussian processes and neural networks at finite widths. In *Mathematical and Scientific Machine Learning*, pages 165–192. PMLR, 2020.

[25] Jacob Zavatone-Veth, Abdulkadir Canatar, Ben Ruben, and Cengiz Pehlevan. Asymptotics of representation learning in finite bayesian neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[26] Gadi Naveh, Oded Ben David, Haim Sompolinsky, and Zohar Ringel. Predicting the outputs of finite deep neural networks trained with noisy gradients. *Physical Review E*, 104(6):064301, 2021.

[27] Daniel A Roberts, Sho Yaida, and Boris Hanin. The principles of deep learning theory. *arXiv preprint arXiv:2106.10165*, 2021.

[28] Boris Hanin. Correlation functions in random fully connected neural networks at finite width. *arXiv preprint arXiv:2204.01058*, 2022.

[29] Kai Segadlo, Bastian Epping, Alexander van Meegen, David Dahmen, Michael Krämer, and Moritz Helias. Unified field theory for deep and recurrent neural networks, 2021.

[30] Gadi Naveh and Zohar Ringel. A self consistent theory of gaussian processes captures feature learning effects in finite cnns. *Advances in Neural Information Processing Systems*, 34, 2021.

[31] Inbar Seroussi and Zohar Ringel. Separation of scales and a thermodynamic description of feature learning in some cnns. *arXiv preprint arXiv:2112.15383*, 2021.

[32] Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Physical Review X*, 11(3):031059, 2021.

[33] Jacob A Zavatone-Veth and Cengiz Pehlevan. Depth induces scale-averaging in overparameterized linear bayesian neural networks. *55th Asilomar Conference on Signals, Systems, and Computers*, 2021.

[34] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In *International conference on machine learning*, pages 4542–4551. PMLR, 2020.

[35] Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from feynman diagrams. *arXiv preprint arXiv:1909.11304*, 2019.

[36] Anders Andreassen and Ethan Dyer. Asymptotics of wide convolutional neural networks. *arXiv preprint arXiv:2008.08675*, 2020.

[37] Jacob A Zavatone-Veth, William L Tong, and Cengiz Pehlevan. Contrasting random and learned features in deep bayesian linear regression. *arXiv preprint arXiv:2203.00573*, 2022.

[38] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

[39] Dyego Araújo, Roberto I Oliveira, and Daniel Yukimura. A mean-field limit for certain deep neural networks. *arXiv preprint arXiv:1906.00193*, 2019.

[40] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

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

[42] 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.

[43] Phan-Minh Nguyen. Mean field limit of the learning dynamics of multilayer neural networks. *arXiv preprint arXiv:1902.02880*, 2019.

[44] Cong Fang, Jason Lee, Pengkun Yang, and Tong Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on learning theory*, pages 1887–1936. PMLR, 2021.

[45] Greg Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34, 2021.

[46] Greg Yang, Michael Santacroce, and Edward J Hu. Efficient computation of deep nonlinear infinite-width neural networks that learn features. In *International Conference on Learning Representations*, 2022.

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

[48] C De Dominicis. Dynamics as a substitute for replicas in systems with quenched random impurities. *Physical Review B*, 18(9):4913, 1978.

[49] Haim Sompolinsky and Annette Zippelius. Dynamic theory of the spin-glass phase. *Physical Review Letters*, 47(5):359, 1981.

[50] Haim Sompolinsky and Annette Zippelius. Relaxational dynamics of the edwards-anderson model and the mean-field theory of spin-glasses. *Physical Review B*, 25(11):6860, 1982.

[51] G Ben Arous and Alice Guionnet. Large deviations for langevin spin glass dynamics. *Probability Theory and Related Fields*, 102(4):455–509, 1995.

[52] G Ben Arous and Alice Guionnet. Symmetric langevin spin glass dynamics. *The Annals of Probability*, 25(3):1367–1422, 1997.

[53] Gérard Ben Arous, Amir Dembo, and Alice Guionnet. Cugliandolo-kurchan equations for dynamics of spin-glasses. *Probability theory and related fields*, 136(4):619–660, 2006.

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

[55] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural networks. *Physical review letters*, 61(3):259, 1988.

[56] Moritz Helias and David Dahmen. *Statistical Field Theory for Neural Networks*. Springer International Publishing, 2020.

[57] Lutz Molgedey, J Schuchhardt, and Heinz G Schuster. Suppressing chaos in neural networks by noise. *Physical review letters*, 69(26):3717, 1992.

[58] M Samuelides and Bruno Cessac. Random recurrent neural networks dynamics. *The European Physical Journal Special Topics*, 142(1):89–122, 2007.

[59] Kanaka Rajan, LF Abbott, and Haim Sompolinsky. Stimulus-dependent suppression of chaos in recurrent neural networks. *Physical review e*, 82(1):011903, 2010.

[60] Stefano Sarao Mannelli, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborova. Passed & spurious: Descent algorithms and local minima in spiked matrix-tensor models. In *international conference on machine learning*, pages 4333–4342. PMLR, 2019.

[61] Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Marvels and pitfalls of the langevin algorithm in noisy high-dimensional inference. *Physical Review X*, 10(1):011057, 2020.

[62] Francesca Mignacco, Pierfrancesco Urbani, and Lenka Zdeborová. Stochasticity helps to navigate rough landscapes: comparing gradient-descent-based algorithms in the phase retrieval problem. *Machine Learning: Science and Technology*, 2(3):035029, 2021.

[63] 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.

[64] 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.

[65] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

[66] Francesca Mignacco and Pierfrancesco Urbani. The effective noise of stochastic gradient descent. *arXiv preprint arXiv:2112.10852*, 2021.

[67] Alessandro Manacorda, Grégory Schehr, and Francesco Zamponi. Numerical solution of the dynamical mean field theory of infinite-dimensional equilibrium liquids. *The Journal of chemical physics*, 152(16):164506, 2020.

[68] Kenji Fukumizu. Dynamics of batch learning in multilayer neural networks. In *International Conference on Artificial Neural Networks*, pages 189–194. Springer, 1998.

[69] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[70] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019.

[71] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

[72] Arthur Jacot, François Ged, Franck Gabriel, Berfin Şimşek, and Clément Hongler. Deep linear networks dynamics: Low-rank biases induced by initialization scale and l2 regularization. *arXiv preprint arXiv:2106.15933*, 2021.

[73] Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022.

[74] Aitor Lewkowycz and Guy Gur-Ari. On the training dynamics of deep networks with l_2 regularization. *Advances in Neural Information Processing Systems*, 33:4790–4799, 2020.

[75] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in neural information processing systems*, 29, 2016.

[76] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *International Conference of Learning Representations*, 2017.

[77] Greg Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. *Advances in neural information processing systems*, 30, 2017.

[78] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

[79] Greg Yang and Etai Littwin. Tensor programs iib: Architectural universality of neural tangent kernel training dynamics. In *International Conference on Machine Learning*, pages 11762–11772. PMLR, 2021.

[80] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[81] Haozhe Shan and Blake Bordelon. A theory of neural tangent kernel alignment and its influence on training, 2021.

[82] Stéphane d'Ascoli, Maria Refinetti, and Giulio Biroli. Optimal learning rate schedules in high-dimensional non-convex optimization problems, 2022.

[83] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[84] Juha Honkonen. Ito and stratonovich calculuses in stochastic field theory. *arXiv preprint arXiv:1102.1581*, 2011.

[85] Crispin W Gardiner et al. *Handbook of stochastic methods*, volume 3. springer Berlin, 1985.

[86] Carl M Bender and Steven Orszag. *Advanced mathematical methods for scientists and engineers I: Asymptotic methods and perturbation theory*, volume 1. Springer Science & Business Media, 1999.

[87] John Hubbard. Calculation of partition functions. *Physical Review Letters*, 3(2):77, 1959.

[88] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, volume 6, pages 583–603. University of California Press, 1972.

[89] Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. *arXiv preprint arXiv:1810.05148*, 2018.

[90] Greg Yang. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. *Advances in Neural Information Processing Systems*, 32, 2019.

[91] Adam X. Yang, Maxime Robeyns, Edward Milsom, Nandi Schoots, and Laurence Aitchison. A theory of representation learning in deep neural networks gives a deep generalisation of kernel methods, 2021.

[92] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

[93] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[94] Gabriel Goh. Why momentum really works. *Distill*, 2017.

[95] Michael Muehlebach and Michael I Jordan. Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives. *Journal of Machine Learning Research*, 22(73):1–50, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] As described in the abstract and introduction, we provide a dynamical field theory of deep networks based on kernel evolution.

(b) Did you describe the limitations of your work? [Yes] We have an explicit limitations as the last paragraph of the paper in Section 7.

(c) Did you discuss any potential negative societal impacts of your work? [N/A] This work is theoretical and is very unlikely to present negative social impacts.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] We describe that our theory holds for NN architectures in the infinite-width $N \to \infty$ limit.

(b) Did you include complete proofs of all theoretical results? [Yes] All claims made in the main text are supported by derivations in the Appendix.

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code to reproduce experimental results is provided in the supplementary material.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We provide details of all experiments in C.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We provided errorbars in the alignment scores of DMFT as a function of width N in Figure 1. All other runs were over a single wide network, where performance is predicted to concentrate over initialization.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We mention our GPU usage in C.2.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We cited the creators of Jax, Neural Tangents, and CIFAR-10.

(b) Did you mention the license of the assets? [N/A] These are all open source provided they are appropriately credited in academic research.

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

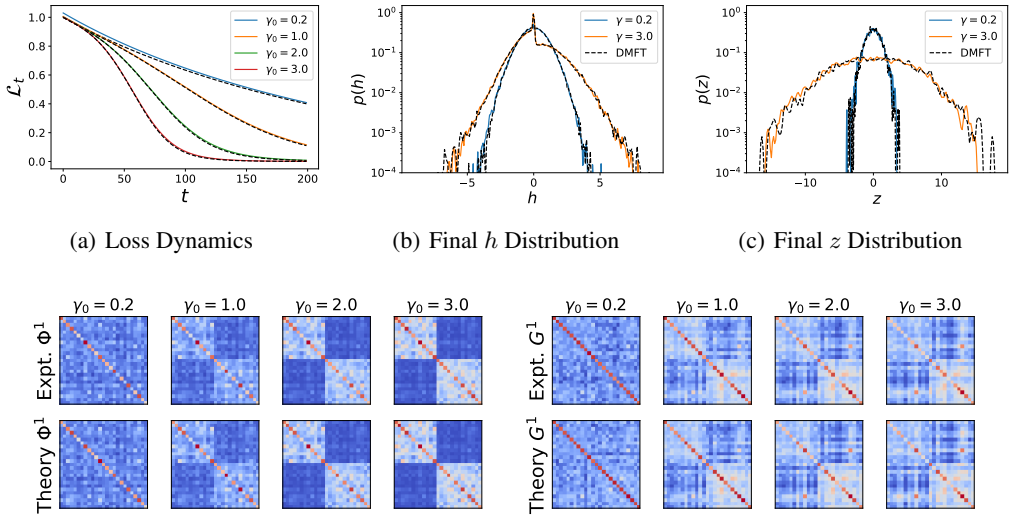
(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

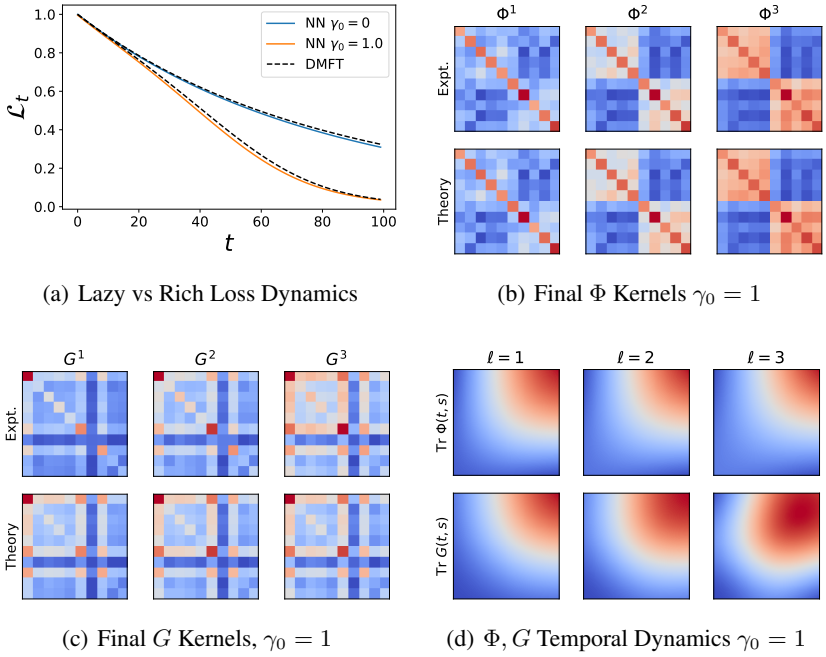
A Additional Figures



(a) Loss Dynamics (b) Final h Distribution (c) Final z Distribution

(d) Final Φ^1 Kernels (e) Final G^1 Kernels

Figure 6: Self-consistent DFT reproduces two layer ($L = 1$ hidden layer, width $N = 2000$) ReLU NN's preactivation density, loss dynamics and learned kernel. (a) The loss is obtained by taking saddle point results for Φ, G and calculating the NTK's dynamics. The $\gamma_0 \to 0$ limit is governed by a static NTK, while the $\gamma_0 > 0$ network exhibits kernel evolution and accelerated training. (b) We plot the preactivation h distribution for neurons in the hidden layer of the trained NN against the theoretical densities defined by $\mathcal{Z}[\Phi, G]$. For small γ_0, the final distribution is approximately Gaussian, but becomes non-Gaussian, asymmetric, and heavy tailed for large γ_0. The DMFT estimate of the distribution is noisy due to finite sampling error. (c) The pre-gradient distribution $p(z)$ in the trained network has larger final variance for large γ_0. (d)-(e) The final Φ, G are accurately predicted by the field theory and exhibit a block structure that increases with γ_0 due to feature learning.



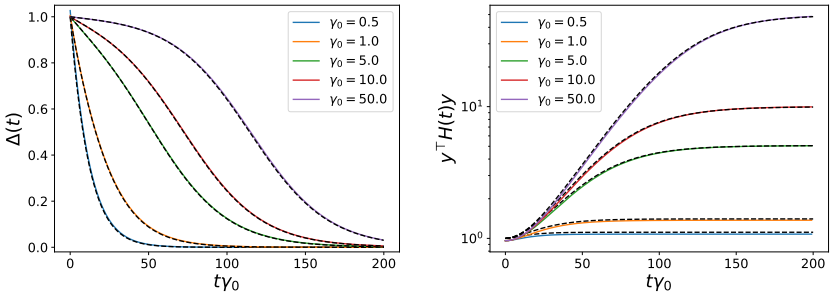
(a) Lazy vs Rich Loss Dynamics

(b) Final Φ Kernels $\gamma_0 = 1$

(c) Final G Kernels, $\gamma_0 = 1$

(d) Φ, G Temporal Dynamics $\gamma_0 = 1$

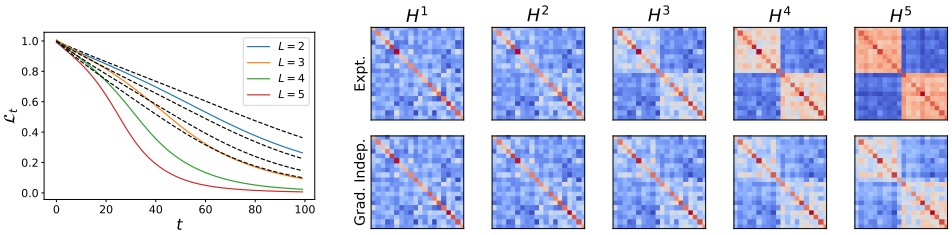
Figure 7: Self-consistent DFT reproduces loss dynamics, and kernels through time in a $L = 3$ tanh network. (a) The loss when training on synthetic data is obtained by taking saddle point results for Φ, G and calculating the NTK's dynamics. The $\gamma_0 \to 0$ limit is governed by a static NTK, while the $\gamma_0 > 0$ network exhibits kernel evolution and accelerated training. Solid lines are a $N = 2000$ NN and dashed lines are from solving DMFT equations. (b)-(c) The final learned kernels Φ (b) and G (c) are accurately predicted by the field theory and exhibits block structure due to clustering by class identity. (d) The temporal components of Φ, G reveals nontrivial dynamical structure.



(a) Two Layer Error Dynamics

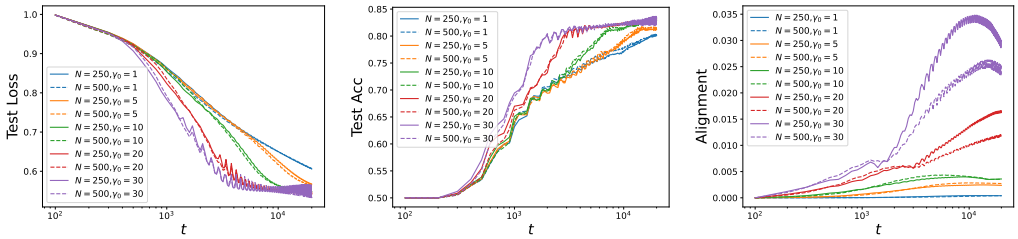
(b) Projection on Target

Figure 8: The error and kernel dynamics obtained by solving a one dimensional ODE system for a depth 2 linear network. (a) The $\Delta(t)$ error dynamics from F.1.1 allows one to solve for $\boldsymbol{H}(t)$ by solving a one dimensional ODE at each value of γ_0. The learning curves interpolate between exponential convergence at small γ_0 and logistic sigmoidal trajectories at large γ_0. (b) The projection of the kernel $\boldsymbol{H}(t)$ along the task relevant subspace $\boldsymbol{y} \in \mathbb{R}^P$.



(a) Grad. Independence DMFT (b) Grad. Independence Predicted Feature Kernels

Figure 9: Gradient independence fails to characterize feature learning dynamics in networks with $L > 1$ and large γ_0. (a) Loss curves for deep linear networks predicted under gradient independence ansatz for $\gamma_0 = 1.5$. (b) The predicted and experimental feature kernels $\boldsymbol{H}^\ell$ for the $L = 5$ hidden layer network demonstrate that gradient independence underestimates the size of kernel adaptation.



(a) Test MSE Loss (b) Test Classification Accuracy (c) $A(\Phi^L, yy^\top)$ Dynamics

Figure 10: Repeating the experiment of Figure 5 with depth 7 ($L = 6$ hidden layer) CNN trained on two class CIFAR over a wide range of γ_0 with $N \in \{250, 500\}$. We find consistent agreement of loss and prediction dynamics across widths but finite size effects become more significant when computing feature kernels of deeper layers. We note that, while higher γ_0 is associated with faster convergence, the final test accuracy for this model is roughly insensitive to choice of γ_0.

B Algorithmic Implementation

The alternating sample-and-solve procedure we developed and describe below for nonlinear networks is based on numerical recipes used in the dynamical mean field simulations in computational physics [67]. The basic principle is to leverage the fact that, conditional on kernels, we can easily draw

samples $\{u_\mu^\ell(t), r_\mu^\ell(t)\}$ from their appropriate GPs. From these sampled fields, we can identify the kernel order parameters by simple estimation of the appropriate moments.

Algorithm 1: Alternating Monte-Carlo Solution to Saddle Point Equations

Data: $\boldsymbol{K}^x, \boldsymbol{y}$, Initial Guesses $\{\boldsymbol{\Phi}^\ell, \boldsymbol{G}^\ell\}_{\ell=1}^L$, $\{\boldsymbol{A}^\ell, \boldsymbol{B}^\ell\}_{\ell=1}^{L-1}$, Sample count $\mathcal{S}$, Update Speed β
Result: Final Kernels $\{\boldsymbol{\Phi}^\ell, \boldsymbol{G}^\ell\}_{\ell=1}^L$, $\{\boldsymbol{A}^\ell, \boldsymbol{B}^\ell\}_{\ell=1}^{L-1}$, Network predictions through training $f_\mu(t)$

$\boldsymbol{\Phi}^0 = \boldsymbol{K}^x \otimes \boldsymbol{1}\boldsymbol{1}^\top$, $\boldsymbol{G}^{L+1} = \boldsymbol{1}\boldsymbol{1}^\top$;
while *Kernels Not Converged* **do**
  From $\{\boldsymbol{\Phi}^\ell, \boldsymbol{G}^\ell\}$ compute $\boldsymbol{K}^{NTK}(t,t)$ and solve $\frac{d}{dt} f_\mu(t) = \sum_\alpha \Delta_\alpha(t) K_{\mu\alpha}^{NTK}(t,t)$;
  $\ell = 1$;
  while $\ell < L+1$ **do**
    Draw $\mathcal{S}$ samples $\{u_{\mu,n}^\ell(t)\}_{n=1}^{\mathcal{S}} \sim \mathcal{GP}(0, \boldsymbol{\Phi}^{\ell-1})$, $\{r_{\mu,n}^\ell(t)\}_{n=1}^{\mathcal{S}} \sim \mathcal{GP}(0, \boldsymbol{G}^{\ell+1})$;
    Solve equation (10) for each sample to get $\{h_{\mu,n}^\ell(t), z_{\mu,n}^\ell(t)\}_{n=1}^{\mathcal{S}}$;
    Compute new $\boldsymbol{\Phi}^\ell, \boldsymbol{G}^\ell$ estimates:
    $\tilde{\Phi}_{\mu\alpha}^\ell(t,s) = \frac{1}{\mathcal{S}} \sum_{n\in[\mathcal{S}]} \phi(h_{\mu,n}^\ell(t))\phi(h_{\alpha,n}^\ell(s))$, $\tilde{G}_{\mu\alpha}^\ell(t,s) = \frac{1}{\mathcal{S}} \sum_{n\in[\mathcal{S}]} g_{\mu,n}^\ell(t) g_{\alpha,n}^\ell(s)$;
    Solve for Jacobians on each sample $\frac{\partial \phi(\boldsymbol{h}_n^\ell)}{\partial \boldsymbol{r}_n^{\ell\top}}$, $\frac{\partial \boldsymbol{g}_n^\ell}{\partial \boldsymbol{u}_n^{\ell\top}}$;
    Compute new $\boldsymbol{A}^\ell, \boldsymbol{B}^{\ell-1}$ estimates:
    $\tilde{\boldsymbol{A}}^\ell = \frac{1}{\mathcal{S}} \sum_{n\in[\mathcal{S}]} \frac{\partial \phi(\boldsymbol{h}_n^\ell)}{\partial \boldsymbol{r}_n^{\ell\top}}$, $\tilde{\boldsymbol{B}}^{\ell-1} = \frac{1}{\mathcal{S}} \sum_{n\in[\mathcal{S}]} \frac{\partial \boldsymbol{g}_n^\ell}{\partial \boldsymbol{u}_n^{\ell\top}}$;
    $\ell \leftarrow \ell + 1$;
  end
  $\ell = 1$;
  while $\ell < L+1$ **do**
    Update feature kernels: $\boldsymbol{\Phi}^\ell \leftarrow (1-\beta)\boldsymbol{\Phi}^\ell + \beta\tilde{\boldsymbol{\Phi}}^\ell$, $\boldsymbol{G}^\ell \leftarrow (1-\beta)\boldsymbol{G}^\ell + \beta\tilde{\boldsymbol{G}}^\ell$;
    if $\ell < L$ **then**
      Update $\boldsymbol{A}^\ell \leftarrow (1-\beta)\boldsymbol{A}^\ell + \beta\tilde{\boldsymbol{A}}^\ell$, $\boldsymbol{B}^\ell \leftarrow (1-\beta)\boldsymbol{B}^\ell + \beta\tilde{\boldsymbol{B}}^\ell$
    end
    $\ell \leftarrow \ell + 1$
  end
end
return $\{\boldsymbol{\Phi}^\ell, \boldsymbol{G}^\ell\}_{\ell=1}^L, \{\boldsymbol{A}^\ell, \boldsymbol{B}^\ell\}_{\ell=1}^{L-1}, \{f_\mu(t)\}_{\mu=1}^P$

The parameter β controls recency weighting of the samples obtained at each iteration. If $\beta = 1$, then the rank of the kernel estimates is limited to the number of samples $\mathcal{S}$ used in a single iteration, but with $\beta < 1$ smaller sample sizes $\mathcal{S}$ can be used to still obtain accurate results. We used $\beta = 0.6$ in our deep network experiments. Convergence is usually achieved in around ~ 15 steps for a depth 4 ($L = 3$ hidden layer) network such as the one in Figure 1 and 7.

C Experimental Details

All NN training was performed with Jax gradient descent optimizer [83] with fixed learning rate.

C.1 MLP Experiments

For the MLP experiments, we performed full batch gradient descent. Networks were initialized with Gaussian weights with unit standard deviation $W_{ij}^\ell \sim \mathcal{N}(0,1)$. The learning rate was chosen as $\eta_0\gamma^2 = \eta_0\gamma_0^2 N$ for a network of width N. The hidden features $\boldsymbol{h}_\mu^\ell(t) \in \mathbb{R}^N$ were stored throughout training and used to compute the kernels $\Phi_{\mu\alpha}^\ell(t,s) = \frac{1}{N}\phi(\boldsymbol{h}_\mu^\ell(t)) \cdot \phi(\boldsymbol{h}_\alpha^\ell(s))$. These experiments can be reproduced with provided jupyter notebooks.

C.2 CNN Experiments on CIFAR-10

We define a depth L CNN model with ReLU activations and stride 1, which is implemented as a pytree of parameters in JAX [83]. We apply global average pooling in the final layer before a dense readout layer. The code to initialize and evaluate the model is provided below.

```python
from jax import random, lax
import jax.numpy as jnp

#L: number of hidden layers, N: width
def initialize_cnn(L, N, seed=0):
    key = random.PRNGKey(seed)
    params = [] # creates list of L+1 weights
    params += [ random.normal(key, (3,3,3,N)) ] # HWIO
    for l in range(L-1):
        key,_ = random.split(key)
        params += [random.normal(key, (3,3,N,N))]
    params += [ random.normal(key, (N,))]
    return params

dn=lax.conv_dimension_numbers((1,3,3,3),(3,3,3,1),('NHWC', 'HWIO','NHWC')) # defines which axis used for convolution
nonlin_fn = lambda h: (h>0.0) * h # ReLU activation
def cnn(params,X):
    L = len(params)-1 # number of hidden layers
    N = params[0].shape[-1] # width
    h = lax.conv_general_dilated(X, params[0],(1,1),'SAME', (1,1),(1,1),dn) # h1
    phi = nonlin_fn(h) # phi(h1)
    for i in range(1,L-1):
        h = 1/jnp.sqrt(N) * lax.conv_general_dilated(phi,
                params[i], (1,1),'SAME',
                (1,1),(1,1),dn) # recurrence for h
        phi = nonlin_fn(h) # phi(h)
    phi = phi.mean(axis = (1,2)) # global average pooling
    w = params[-1]
    f = 1/N * phi @ w # Mean-field parameterization
    return f
```

After constructing a CNN model, we train using MSE loss with base learning rate $\eta_0 = 2.0 \times 10^{-4}$, batch size 250. The learning rate passed to the optimizer is thus $\eta = \eta_0 \gamma^2 = \eta_0 \gamma_0^2 N$. We optimize the loss function which is scaled appropriately as $\ell(\gamma_0^{-1} f, y)$. Throughout training, we compute the last layer's embedding $\phi(\boldsymbol{h}^L)$ on the test set to calculate the alignment $A(\boldsymbol{\Phi}^L, \boldsymbol{y}\boldsymbol{y}^\top)$. Training was performed on 4 NVIDIA GPUs. Training a $L = 3$ network of width 500 takes roughly 1 hour.

D Derivation of Self-Consistent Dynamical Field Theory

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 [47], of which a useful review for random recurent networks can be found here [54]. Similar computations can be found in recent works which consider typical behavior in high dimensional classification on random data [63, 64].

D.1 Deep Network Field Definitions and Scaling

As discussed in the main text, we consider the following wide network architecture parameterzied by trainable weights $\boldsymbol{\theta} = \text{Vec}\{\boldsymbol{W}^0, \boldsymbol{W}^1, ...\boldsymbol{w}^L\}$, giving network output f_μ defined as

$$
\begin{aligned}
f_\mu &= \frac{1}{\gamma} h_\mu^{L+1} \ , \ h_\mu^{L+1} = \frac{1}{\sqrt{N}} \boldsymbol{w}_L \cdot \phi(\boldsymbol{h}_\mu^L) \\
\boldsymbol{h}_\mu^{\ell+1} &= \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell \phi(\boldsymbol{h}_\mu^\ell) \ , \ \boldsymbol{h}_\mu^1 = \frac{1}{\sqrt{D}} \boldsymbol{W}^0 \boldsymbol{x}_\mu
\end{aligned} \tag{13}
$$

Using gradient flow with learning rate η on cost $\mathcal{L} = \sum_\mu \ell(f_\mu, y_\mu)$ for loss function, we introduce functions $\Delta_\mu = -\frac{\partial \mathcal{L}}{\partial f_\mu}$ and η for learning rate, gradient flow induces the following dynamics

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\eta}{\gamma} \sum_\mu \Delta_\mu \frac{\partial h_\mu^{L+1}}{\partial \boldsymbol{\theta}} \ , \ \frac{\partial f_\mu}{\partial t} = \frac{\eta}{\gamma^2} \sum_\alpha \Delta_\alpha K_{\mu\alpha}^{NTK} \ , \ K_{\mu\alpha}^{NTK} = \frac{\partial h_\mu^{L+1}}{\partial \boldsymbol{\theta}} \cdot \frac{\partial h_\alpha^{L+1}}{\partial \boldsymbol{\theta}} \tag{14}$$

Since K_{NTK} is $O_\gamma(1)$ at initialization, it is clear that to have $O_\gamma(1)$ evolution of the network output at initialization we need $\eta = \gamma^2$. With this scaling, we have the following

$$\frac{d\boldsymbol{\theta}}{dt} = \gamma \sum_\mu \Delta_\mu \frac{\partial h_\mu^{L+1}}{\partial \boldsymbol{\theta}} \ , \ \frac{\partial f_\mu}{\partial t} = \sum_\alpha \Delta_\alpha K_{\mu\alpha}^{NTK} \tag{15}$$

Now, to build a valid field theory, we want to express everything in terms of features $\boldsymbol{h}_\mu^\ell$ rather than parameters $\boldsymbol{\theta}$ and we will define the following gradient features $\boldsymbol{g}_\mu^\ell = \sqrt{N} \frac{\partial h_\mu^{L+1}}{\partial \boldsymbol{h}_\mu^\ell}$ which admit the recursion and base case

$$\begin{aligned} \boldsymbol{g}_\mu^\ell &= \sqrt{N} \frac{\partial h_\mu^{L+1}}{\partial \boldsymbol{h}_\mu^\ell} = \left(\frac{\partial \boldsymbol{h}_\mu^{\ell+1}}{\partial \boldsymbol{h}_\mu^\ell} \right)^\top \left(\sqrt{N} \frac{\partial h_\mu^{L+1}}{\partial \boldsymbol{h}_\mu^{\ell+1}} \right) = \dot{\phi}(\boldsymbol{h}_\mu^\ell) \odot \boldsymbol{z}_\mu^\ell \ , \ \boldsymbol{z}_\mu^\ell = \frac{1}{\sqrt{N}} \boldsymbol{W}^{\ell\top} \boldsymbol{g}_\mu^{\ell+1} \\ \boldsymbol{g}_\mu^L &= \dot{\phi}(\boldsymbol{h}_\mu^L) \odot \boldsymbol{w}^L \end{aligned} \tag{16}$$

We define the *pre-gradient* field $\boldsymbol{z}_\mu^\ell = \frac{1}{\sqrt{N}} \boldsymbol{W}^{\ell\top} \boldsymbol{g}_\mu^{\ell+1}$ so that $\boldsymbol{g}_\mu^\ell = \dot{\phi}(\boldsymbol{h}_\mu^\ell) \odot \boldsymbol{z}_\mu^\ell(t)$. From these quantities, we can derive the gradients with respect to parameters

$$\frac{\partial h_\mu^{L+1}}{\partial \boldsymbol{W}^\ell} = \sum_{i=1}^N \frac{\partial h_\mu^{L+1}}{\partial h_{\mu,i}^{\ell+1}} \frac{\partial h_{\mu,i}^{\ell+1}}{\partial \boldsymbol{W}^\ell} = \frac{1}{N} \boldsymbol{g}_\mu^{\ell+1} \phi(\boldsymbol{h}_\mu^\ell)^\top \tag{17}$$

which allows us to compute the NTK in terms of these features

$$K_{\mu\alpha}^{NTK} = \frac{1}{N} \phi(\boldsymbol{h}_\mu^L) \cdot \phi(\boldsymbol{h}_\alpha^L) + \sum_{\ell=1}^{L-1} \left(\frac{\boldsymbol{g}_\mu^{\ell+1} \cdot \boldsymbol{g}_\alpha^{\ell+1}}{N} \right) \left(\frac{\phi(\boldsymbol{h}_\mu^\ell) \cdot \phi(\boldsymbol{h}_\alpha^\ell)}{N} \right) + \frac{\boldsymbol{g}_\mu^1 \cdot \boldsymbol{g}_\alpha^1}{N} K_{\mu\alpha}^x \tag{18}$$

where $K_{\mu\alpha}^x = \frac{\boldsymbol{x}_\mu \cdot \boldsymbol{x}_\alpha}{D}$ is the input Grammian. We see that the NTK can be built out of the following primitive kernels

$$\Phi_{\mu\nu}^\ell = \frac{1}{N} \phi(\boldsymbol{h}_\mu^\ell) \cdot \phi(\boldsymbol{h}_\nu^\ell) \ , \ G_{\mu\nu}^\ell = \frac{1}{N} \boldsymbol{g}_\mu^\ell \cdot \boldsymbol{g}_\nu^\ell \tag{19}$$

We utilize the parameter space dynamics to express $\boldsymbol{W}^\ell$ in terms of the $\{\boldsymbol{g}, \boldsymbol{h}\}$ fields

$$\boldsymbol{W}^\ell(t) = \boldsymbol{W}^\ell(0) + \frac{\gamma}{N} \int_0^t ds \sum_\mu \Delta_\alpha(s) \boldsymbol{g}_\mu^{\ell+1}(s) \phi(\boldsymbol{h}_\mu^\ell(s))^\top \tag{20}$$

Using the field recurrences $\boldsymbol{h}_\mu^{\ell+1}(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(t) \phi(\boldsymbol{h}_\mu^\ell(t))$ we can derive the following recursive dynamics for the features

$$\begin{aligned} \boldsymbol{h}_\mu^{\ell+1}(t) &= \boldsymbol{\chi}_\mu^{\ell+1}(t) + \frac{\gamma}{\sqrt{N}} \int_0^t ds \sum_\nu \Delta_\nu \boldsymbol{g}_\nu^{\ell+1}(t) \Phi_{\nu\mu}^\ell(s,t) \\ \boldsymbol{z}_\mu^\ell(t) &= \boldsymbol{\xi}_\mu^\ell(t) + \frac{\gamma}{\sqrt{N}} \int_0^t ds \sum_\nu \Delta_\nu(s) \phi(\boldsymbol{h}_\nu^\ell(s)) G_{\nu\mu}^{\ell+1}(s,t) \ , \ \boldsymbol{g}_\mu^\ell(t) = \dot{\phi}(\boldsymbol{h}_\mu^\ell(t)) \odot \boldsymbol{z}_\mu^\ell(t) \\ \frac{\partial f_\mu}{\partial t} &= \sum_\alpha \Delta_\alpha(t) \left[\Phi_{\mu\alpha}^L(t,t) + \sum_{\ell=1}^{L-1} G_{\mu\alpha}^{\ell+1}(t,t) \Phi_{\mu\alpha}^\ell(t,t) + G_{\mu\alpha}^1(t,t) K_{\mu\alpha}^x \right] \end{aligned} \tag{21}$$

where we introduced the following random fields $\boldsymbol{\chi}_\mu^\ell(t), \boldsymbol{\xi}_\mu^\ell(t)$ which involve the random initial conditions

$$\boldsymbol{\chi}_\mu^\ell(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0) \phi(\boldsymbol{h}_\mu^\ell(t)) \ , \ \boldsymbol{\xi}_\mu^\ell(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0)^\top \boldsymbol{g}_\mu^{\ell+1}(t) \tag{22}$$

We observe that the dynamics of the hidden features is controlled by the factor $\frac{\gamma}{\sqrt{N}}$. If $\gamma = O_N(1)$ then we recover static NTK in the limit as $N \to \infty$. However, if $\gamma = O_N(\sqrt{N})$ then we obtain $O_N(1)$ evolution of our features and we reach a new rich regime. We choose the scaling $\gamma = \gamma_0 \sqrt{N}$ for our field theory so that $\gamma_0 > 0$ will give a feature learning network.

D.2 Warmup: DMFT for One Hidden Layer NN

In this section, we provide a warmup problem of a $L = 1$ hidden layer network which allows us to illustrate the mechanics of the MSRDJ formalism. A more detailed computation can be found 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 $\boldsymbol{\chi}_\mu = \frac{1}{\sqrt{D}}\boldsymbol{W}^0(0)\boldsymbol{x}_\mu$ and $\boldsymbol{\xi} = \boldsymbol{w}^1(0)$. Unlike the deeper $L \geq 2$ 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

$$\begin{aligned}\boldsymbol{h}_\mu(t) &= \boldsymbol{\chi}_\mu + \gamma_0 \int_0^t ds \sum_\alpha [\boldsymbol{z}(s) \odot \dot{\phi}(\boldsymbol{h}_\alpha(s))] K^x_{\mu\alpha} \Delta_\alpha(s) \\ \boldsymbol{z}(t) &= \boldsymbol{\xi} + \gamma_0 \int_0^t ds \sum_\alpha \phi(\boldsymbol{h}_\alpha(s)) \Delta_\alpha(s). \end{aligned} \tag{23}$$

where the network predictions evolve as $\frac{\partial}{\partial t} f_\mu(t) = \sum_\alpha [\Phi_{\mu\alpha}(t,t) + G_{\mu\alpha}(t,t) K^x_{\mu\alpha}] \Delta_\alpha(t)$ for kernels $\Phi_{\mu\alpha}(t,t) = \frac{1}{N}\phi(\boldsymbol{h}_\mu(t)) \cdot \phi(\boldsymbol{h}_\alpha(t))$ and $G_{\mu\alpha}(t,t) = \frac{1}{N}\boldsymbol{g}_\mu(t) \cdot \boldsymbol{g}_\alpha(t)$. 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

$$Z[\{\boldsymbol{j}_\mu\}_{\mu \in [P]}, \boldsymbol{v}] = \left\langle \exp\left(\sum_\mu \boldsymbol{j}_\mu \cdot \boldsymbol{\chi}_\mu + \boldsymbol{\xi} \cdot \boldsymbol{v} \right) \right\rangle_{\boldsymbol{\theta}_0}. \tag{24}$$

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

$$\begin{aligned} 1 &= \int d\boldsymbol{\chi}_\mu \delta\left(\boldsymbol{\chi}_\mu - \frac{1}{\sqrt{N}}\boldsymbol{W}^0(0)\boldsymbol{x}_\mu\right) = \int \frac{d\boldsymbol{\chi}_\mu d\hat{\boldsymbol{\chi}}_\mu}{(2\pi)^N} \exp\left(i\hat{\boldsymbol{\chi}}_\mu \cdot \left(\boldsymbol{\chi}_\mu - \frac{1}{\sqrt{D}}\boldsymbol{W}^0(0)\boldsymbol{x}_\mu\right)\right) \\ 1 &= \int d\boldsymbol{\xi}\, \delta\left(\boldsymbol{\xi} - \boldsymbol{w}^1(0)\right) = \int \frac{d\boldsymbol{\xi} d\hat{\boldsymbol{\xi}}}{(2\pi)^N} \exp\left(i\hat{\boldsymbol{\xi}} \cdot \left(\boldsymbol{\xi} - \boldsymbol{w}^1(0)\right)\right). \end{aligned} \tag{25}$$

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 mulitplying by these factors of unity and performing the Gaussian integrals, we obtain

$$Z = \int \prod_\mu \frac{d\boldsymbol{\chi} d\hat{\boldsymbol{\chi}}}{(2\pi)^N} \frac{d\boldsymbol{\xi} d\hat{\boldsymbol{\xi}}}{(2\pi)^N} \exp\left(-\frac{1}{2}\sum_{\mu\alpha} \hat{\boldsymbol{\chi}}_\mu \cdot \hat{\boldsymbol{\chi}}_\alpha K^x_{\mu\alpha} + \sum_\mu \boldsymbol{\chi}_\mu \cdot (i\hat{\boldsymbol{\chi}}_\mu + \boldsymbol{j}_\mu) - \frac{1}{2}|\hat{\boldsymbol{\xi}}|^2 + \boldsymbol{\xi} \cdot (i\hat{\boldsymbol{\xi}} + \boldsymbol{v})\right) \tag{26}$$

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

$$\begin{aligned} 1 &= N \int d\Phi_{\mu\alpha}(t,s)\, \delta\left(N\Phi_{\mu\alpha}(t,s) - \phi(\boldsymbol{h}_\mu(t)) \cdot \phi(\boldsymbol{h}_\alpha(s))\right) \\ &= \int \frac{d\Phi_{\mu\alpha}(t,s) d\hat{\Phi}_{\mu\alpha}(t,s)}{2\pi i N^{-1}} \exp\left(N\hat{\Phi}_{\mu\alpha}(t,s)\left(N\Phi_{\mu\alpha}(t,s) - \phi(\boldsymbol{h}_\mu(t)) \cdot \phi(\boldsymbol{h}_\alpha(s))\right)\right) \\ 1 &= N \int dG_{\mu\alpha}(t,s)\, \delta\left(NG_{\mu\alpha}(t,s) - \boldsymbol{g}_\mu(t) \cdot \boldsymbol{g}_\alpha(s)\right) \\ &= \int \frac{dG_{\mu\alpha}(t,s) d\hat{G}_{\mu\alpha}(t,s)}{2\pi i N^{-1}} \exp\left(N\hat{G}_{\mu\alpha}(t,s)\left(NG_{\mu\alpha}(t,s) - \boldsymbol{g}_\mu(t) \cdot \boldsymbol{g}_\alpha(s)\right)\right), \end{aligned} \tag{27}$$

where the fields $\boldsymbol{h}_\mu(t), \boldsymbol{g}_\mu(t)$ are regarded as functions of $\{\boldsymbol{\chi}_\mu\}_\mu, \boldsymbol{\xi}$ (see Equation (23)) and the $\hat{\Phi}, \hat{G}$ integrals run over the imaginary axis $(-i\infty, i\infty)$. After this step, we can write

$$Z \propto \int \prod_{\mu\alpha t s} d\Phi_{\mu\alpha}(t,s) d\hat{\Phi}_{\mu\alpha}(t,s) dG_{\mu\alpha}(t,s) d\hat{G}_{\mu\alpha}(t,s) \exp\left(NS[\Phi, \hat{\Phi}, G, \hat{G}]\right) \tag{28}$$

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

$$S[\Phi, \hat{\Phi}, G, \hat{G}] = \sum_{\mu\alpha} \int dtds[\Phi_{\mu\alpha}(t,s)\hat{\Phi}_{\mu\alpha}(t,s) + G_{\mu\alpha}(t,s)\hat{G}_{\mu\alpha}(t,s)] + \frac{1}{N}\sum_{i=1}^{N} \ln \mathcal{Z}[j_i, v_i]. \tag{29}$$

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

$$\begin{aligned}\mathcal{Z}[j,v] = \int \prod_{\mu} \frac{d\chi_\mu d\hat{\chi}_\mu}{2\pi} \frac{d\xi d\hat{\xi}}{2\pi} \exp\left(-\frac{1}{2}\sum_{\mu\alpha} \hat{\chi}_\mu \hat{\chi}_\alpha K^x_{\mu\alpha} + (j_\mu + i\hat{\chi}_\mu)\chi_\mu - \frac{1}{2}\hat{\xi}^2 + (v + i\hat{\xi})\xi\right) \\ \times \exp\left(-\int_0^\infty dt \int_0^\infty ds \sum_{\mu\alpha}[\hat{\Phi}_{\mu\alpha}(t,s)\phi(h_\mu(t))\phi(h_\alpha(s)) + \hat{G}_{\mu\alpha}(t,s) g_\mu(t) g_\alpha(s)]\right)\end{aligned} \tag{30}$$

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 $Z \propto \int d\Phi d\hat{\Phi} dG d\hat{G} \exp\left(NS[\Phi, \hat{\Phi}, G, \hat{G}]\right)$, 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.

$$\begin{aligned}\frac{\delta S}{\delta \Phi_{\mu\alpha}(t,s)} = \hat{\Phi}_{\mu\alpha}(t,s) = 0\ ,\ \frac{\delta S}{\delta \hat{\Phi}_{\mu\alpha}(t,s)} = \Phi_{\mu\alpha}(t,s) - \frac{1}{N}\sum_{i=1}^{N} \langle \phi(h_\mu(t))\phi(h_\alpha(s))\rangle_i = 0 \\ \frac{\delta S}{\delta G_{\mu\alpha}(t,s)} = \hat{G}_{\mu\alpha}(t,s) = 0\ ,\ \frac{\delta S}{\delta \hat{G}_{\mu\alpha}(t,s)} = G_{\mu\alpha}(t,s) - \frac{1}{N}\sum_{i=1}^{N} \langle g_\mu(t) g_\alpha(s)\rangle_i = 0\end{aligned} \tag{31}$$

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

$$\left\langle O(\chi, \hat{\chi}, \xi, \hat{\xi})\right\rangle_i = \frac{1}{\mathcal{Z}[j_i, v_i]} \int \prod_\mu d\chi_\mu d\hat{\chi}_\mu d\xi d\hat{\xi} \exp\left(-\mathcal{H}[\{\chi_\mu, \hat{\chi}_\mu\}, \xi, \hat{\xi}, j_i, v_i]\right) O(\chi, \hat{\chi}, \xi, \hat{\xi}) \tag{32}$$

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 $\boldsymbol{j}, \boldsymbol{v} \to 0$.

D.2.1 Final $L = 1$ DMFT equations

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

$$\begin{aligned}&\{\chi_\mu\}_{\mu\in[P]} \sim \mathcal{N}(0, \boldsymbol{K}^x)\ ,\ \xi \sim \mathcal{N}(0,1) \\ &h_\mu(t) = \chi_\mu + \gamma_0 \int_0^t ds \sum_\alpha [z(s)\dot{\phi}(h_\alpha(s))] K^x_{\mu\alpha} \Delta_\alpha(s)\ ,\ z(t) = \xi + \gamma_0 \int_0^t ds \sum_\alpha \phi(h_\alpha(s))\Delta_\alpha(s) \\ &\Phi_{\mu\alpha}(t,s) = \langle \phi(h_\mu(t))\phi(h_\alpha(s))\rangle\ ,\ G_{\mu\alpha}(t,s) = \langle g_\mu(t) g_\alpha(s)\rangle = \left\langle z(t)z(s)\dot{\phi}(h_\mu(t))\dot{\phi}(h_\alpha(s))\right\rangle \\ &\frac{\partial f_\mu}{\partial t} = \sum_\alpha [\Phi_{\mu\alpha}(t,t) + G_{\mu\alpha}(t,t) K^x_{\mu\alpha}]\Delta_\alpha(t)\end{aligned} \tag{33}$$

We see that for $L = 1$ 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.

D.3 Path Integral Formulation for Deep Networks

As discussed in the main text, we study the distribution over fields by computing the moment generating functional for the stochastic processes $\{\boldsymbol{\chi}^\ell, \boldsymbol{\xi}^\ell\}_{\ell=1}^L$

$$Z[\{\boldsymbol{j}^\ell, \boldsymbol{v}^\ell\}] = \left\langle \exp\left(\sum_{\ell,\mu} \int_0^\infty dt \left[\boldsymbol{j}_\mu^\ell(t) \cdot \boldsymbol{\chi}_\mu^\ell(t) + \boldsymbol{v}_\mu^\ell(t) \cdot \boldsymbol{\xi}_\mu^\ell(t)\right]\right)\right\rangle_{\boldsymbol{\theta}_0 = \text{Vec}\{\boldsymbol{W}^0(0), \ldots \boldsymbol{w}^L(0)\}} \tag{34}$$

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

$$\left\langle \chi_{\mu_1}^{\ell_1}(t_1) \ldots \chi_{\mu_n}^{\ell_n}(t_n) \xi_{\mu_1}^{\ell_1}(t_1) \ldots \xi_{\mu_m}^{\ell_m}(t_m)\right\rangle = \frac{\delta}{\delta j_{\mu_1}^{\ell_1}(t_1)} \cdots \frac{\delta}{\delta j_{\mu_n}^{\ell_n}(t_n)} \frac{\delta}{\delta v_{\mu_1}^{\ell_1}(t_1)} \cdots \frac{\delta}{\delta v_{\mu_m}^{\ell_m}(t_m)} Z[\{\boldsymbol{j}^\ell, \boldsymbol{v}^\ell\}]|_{\boldsymbol{j}=\boldsymbol{v}=0}. \tag{35}$$

To perform the average over the initial parameters, we enforce the definition of the fields $\boldsymbol{\chi}^{\ell+1}(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0) \phi(\boldsymbol{h}_\mu^\ell(t))$, $\boldsymbol{\xi}_\mu^\ell(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0)^\top \boldsymbol{g}_\mu^{\ell+1}(t)$, by inserting the following terms in the definition of $Z[\{\boldsymbol{j}, \boldsymbol{v}\}]$ so we may more easily perform the average over weights $\boldsymbol{\theta}_0$. We enforce these definitions with an integral representation of the Dirac-Delta function $1 = \int_{\mathbb{R}} dx\ \delta(x) = \frac{1}{2\pi} \int_{\mathbb{R}} dx \int_{\mathbb{R}} d\hat{x} \exp(ix\hat{x})$. We note that we are implicitly working in the Ito scheme, where factors of Jacobian determinants are equal to one [54, 84, 85] (we note that $\boldsymbol{h}_\mu^\ell(t)$ does not causally depend on $\boldsymbol{\chi}_\mu^{\ell+1}(t)$ and $\boldsymbol{g}_\mu^\ell(t)$ does not causally depend on $\boldsymbol{\xi}^\ell(t)$). Applying this to fields $\boldsymbol{\chi}, \boldsymbol{\xi}$, we have

$$\begin{aligned}
1 &= \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \frac{d\boldsymbol{\chi}_\mu^1(t) d\hat{\boldsymbol{\chi}}_\mu^1(t)}{(2\pi)^N} \exp\left(i\hat{\boldsymbol{\chi}}_\mu^1(t) \cdot \left[\boldsymbol{\chi}_\mu^1(t) - \frac{1}{\sqrt{D}} \boldsymbol{W}^\ell(0) \boldsymbol{x}_\mu\right]\right) \\
1 &= \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \frac{d\boldsymbol{\chi}_\mu^{\ell+1}(t) d\hat{\boldsymbol{\chi}}_\mu^{\ell+1}(t)}{(2\pi)^N} \exp\left(i\hat{\boldsymbol{\chi}}_\mu^{\ell+1}(t) \cdot \left[\boldsymbol{\chi}_\mu^{\ell+1}(t) - \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0) \phi(\boldsymbol{h}_\mu^\ell(t))\right]\right)\ ,\ \ell \in \{1, ..., L-1\} \\
1 &= \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \frac{d\boldsymbol{\xi}_\mu^L(t) d\hat{\boldsymbol{\xi}}_\mu^L(t)}{(2\pi)^N} \exp\left(i\hat{\boldsymbol{\xi}}_\mu^L(t) \cdot \left[\boldsymbol{\xi}_\mu^L(t) - \boldsymbol{w}^L(0)\right]\right) \\
1 &= \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \frac{d\boldsymbol{\xi}_\mu^\ell(t) d\hat{\boldsymbol{\xi}}_\mu^\ell(t)}{(2\pi)^N} \exp\left(i\hat{\boldsymbol{\xi}}_\mu^\ell(t) \cdot \left[\boldsymbol{\xi}_\mu^\ell(t) - \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0)^\top \boldsymbol{g}_\mu^\ell(t)\right]\right)\ ,\ \ell \in \{1, ..., L-1\}
\end{aligned} \tag{36}$$

where $\{h^\ell, g^\ell\}$ are understood to be stochastic processes which are causally determined by the $\{\chi^\ell, \xi^\ell\}$ fields, in the sense that $h^\ell(t)$ only depends on $\chi^\ell(s)$ for $s < t$. We thus have an expression of the form

$$\begin{aligned}
Z[\{\boldsymbol{j}^\ell, \boldsymbol{v}^\ell\}] = \int \prod_{\ell\mu t} \frac{d\boldsymbol{\chi}_\mu^{\ell+1}(t) d\hat{\boldsymbol{\chi}}_\mu^{\ell+1}(t)}{(2\pi)^N} \prod_{\ell\mu t} \frac{d\boldsymbol{\xi}_\mu^\ell(t) d\hat{\boldsymbol{\xi}}_\mu^\ell(t)}{(2\pi)^N} \exp\left(\sum_{\ell,\mu} \int_0^\infty dt \left[\boldsymbol{j}_\mu^\ell(t) \cdot \boldsymbol{\chi}_\mu^\ell(t) + \boldsymbol{v}_\mu^\ell(t) \cdot \boldsymbol{\xi}_\mu^\ell(t)\right]\right) \\
\times \prod_{\ell=1}^{L-1} \left\langle \exp\left(-\frac{i}{\sqrt{N}} \sum_\mu \int_0^\infty dt \left[\hat{\boldsymbol{\chi}}_\mu^{\ell+1}(t)^\top \boldsymbol{W}^\ell(0) \phi(\boldsymbol{h}_\mu^\ell(t)) + \boldsymbol{g}_\mu^{\ell+1}(t)^\top \boldsymbol{W}^\ell(0) \hat{\boldsymbol{\xi}}_\mu^\ell(t)\right]\right)\right\rangle_{\boldsymbol{W}^\ell(0)} \\
\times \left\langle \exp\left(-\frac{i}{\sqrt{D}} \sum_\mu \int_0^\infty dt\ \hat{\boldsymbol{\chi}}_\mu^1(t)^\top \boldsymbol{W}^0(0) \boldsymbol{x}_\mu\right)\right\rangle_{\boldsymbol{W}^0(0)} \\
\times \left\langle \exp\left(-i \sum_\mu \int_0^\infty \hat{\boldsymbol{\xi}}_\mu^L(t) \cdot \boldsymbol{w}^L(0)\right)\right\rangle_{\boldsymbol{w}^L(0)} \\
\times \prod_{\ell=1}^L \exp\left(i \sum_\mu \int_0^\infty dt\ \left[\hat{\boldsymbol{\chi}}_\mu^\ell(t) \cdot \boldsymbol{\chi}_\mu^\ell(t) + \hat{\boldsymbol{\xi}}_\mu^\ell(t) \cdot \boldsymbol{\xi}_\mu^\ell(t)\right]\right)
\end{aligned} \tag{37}$$

Since $\boldsymbol{W}^\ell(0)$ are all Gaussian random variables, these averages can be performed quite easily yielding

$$
\begin{aligned}
&\left\langle \exp\left(-\frac{i}{\sqrt{D}}\sum_{\mu}\int_0^\infty dt \hat{\boldsymbol{\chi}}^1_\mu(t)^\top \boldsymbol{W}^0(0)\boldsymbol{x}_\mu\right)\right\rangle_{\boldsymbol{W}^0(0)} = \exp\left(-\frac{1}{2}\int_0^\infty\int_0^\infty dtds \sum_{\mu\alpha}\hat{\boldsymbol{\chi}}^1_\mu(t)\cdot\hat{\boldsymbol{\chi}}^1_\alpha(s)K^x_{\mu\alpha}\right)\\
&\left\langle \exp\left(-i\sum_{\mu}\int_0^\infty \hat{\boldsymbol{\xi}}^L_\mu(t)\cdot\boldsymbol{w}^L(0)\right)\right\rangle_{\boldsymbol{w}^L(0)} = \exp\left(-\frac{1}{2}\sum_{\mu\alpha}\int_0^\infty\int_0^\infty dtds\, \hat{\boldsymbol{\xi}}^L_\mu(t)\cdot\hat{\boldsymbol{\xi}}^L_\alpha(s)\right)\\
&\left\langle \exp\left(-\frac{i}{\sqrt{N}}\sum_{\mu}\int_0^\infty dt\left[\hat{\boldsymbol{\chi}}^{\ell+1}_\mu(t)^\top\boldsymbol{W}^\ell(0)\phi(\boldsymbol{h}^\ell_\mu(t)) + \boldsymbol{g}^{\ell+1}_\mu(t)^\top\boldsymbol{W}^\ell(0)\hat{\boldsymbol{\xi}}^\ell_\mu(t)\right]\right)\right\rangle_{\boldsymbol{W}^\ell(0)}\\
&= \exp\left(-\frac{1}{2N}\sum_{\mu\alpha}\int_0^\infty\int_0^\infty dtds\left[\hat{\boldsymbol{\chi}}^{\ell+1}_\mu(t)\cdot\hat{\boldsymbol{\chi}}^{\ell+1}_\mu(t)\phi(\boldsymbol{h}^\ell_\mu(t))\cdot\phi(\boldsymbol{h}^\ell_\alpha(s)) + \hat{\boldsymbol{\xi}}^\ell_\mu(t)\cdot\hat{\boldsymbol{\xi}}^\ell_\alpha(s)\boldsymbol{g}^{\ell+1}_\mu(t)\cdot\boldsymbol{g}^{\ell+1}_\alpha(s)\right]\right)\\
&\times\exp\left(-\frac{1}{N}\sum_{\mu\alpha}\int_0^\infty\int_0^\infty dtds\, \hat{\boldsymbol{\chi}}^{\ell+1}_\mu(t)\cdot\boldsymbol{g}^{\ell+1}_\alpha(s)\;\phi(\boldsymbol{h}^\ell_\mu(t))\cdot\hat{\boldsymbol{\xi}}^\ell_\alpha(s)\right)
\end{aligned}
\tag{38}
$$

D.4 Order Parameters and Action Definition

We define the following order parameters which we will show concentrate in the $N\to\infty$ limit

$$
\Phi^\ell_{\mu,\alpha}(t,s) = \frac{1}{N}\phi(\boldsymbol{h}^\ell_\mu(t))\cdot\phi(\boldsymbol{h}^\ell_\alpha(s))\ ,\ G^\ell_{\mu\alpha}(t,s) = \frac{1}{N}\boldsymbol{g}^\ell_\mu(t)\cdot\boldsymbol{g}^\ell_\alpha(s)\ ,\ A^\ell_{\mu\alpha}(t,s) = -\frac{i}{N}\phi(\boldsymbol{h}^\ell_\mu(t))\cdot\hat{\boldsymbol{\xi}}^\ell_\alpha(s).
\tag{39}
$$

The NTK only depends on $\{\Phi^\ell, G^\ell\}$ so from these order parameters, we can compute the function evolution. The parameter $\boldsymbol{A}^\ell$ arises from the coupling of the fields across a single layer's initial weight matrix $\boldsymbol{W}^\ell(0)$. We can again enforce these definitions with integral representations of the Dirac-delta function. For each pair of samples μ,α and each pair of times t,s, we multiply by

$$
\begin{aligned}
1 &= \int\int\frac{d\Phi^\ell_{\mu\alpha}(t,s)d\hat{\Phi}^\ell_{\mu\alpha}(t,s)}{2\pi iN^{-1}}\exp\left(N\Phi^\ell_{\mu\alpha}(t,s)\hat{\Phi}^\ell_{\mu\alpha}(t,s) - \hat{\Phi}^\ell_{\mu\alpha}(t,s)\phi(\boldsymbol{h}^\ell_\mu(t))\cdot\phi(\boldsymbol{h}^\ell_\alpha(s))\right)\ \ell\in\{1,...,L\}\\
1 &= \int\int\frac{dG_{\mu\alpha}(t,s)d\hat{G}_{\mu\alpha}(t,s)}{2\pi iN^{-1}}\exp\left(NG^\ell_{\mu\alpha}(t,s)\hat{G}^\ell_{\mu\alpha}(t,s) - \hat{G}^\ell_{\mu\alpha}(t,s)\boldsymbol{g}^\ell_\mu(t)\cdot\boldsymbol{g}^\ell_\alpha(s)\right)\ ,\ \ell\in\{1,...,L\}\\
1 &= \int\int\frac{dA^\ell_{\mu\alpha}(t,s)dB^\ell_{\mu\alpha}(t,s)}{2\pi iN^{-1}}\exp\left(-NA^\ell_{\mu\alpha}(t,s)B^\ell_{\mu\alpha}(t,s) - iB^\ell_{\mu\alpha}(t,s)\phi(\boldsymbol{h}^\ell_\mu(t))\cdot\hat{\boldsymbol{\xi}}^\ell_\alpha(s))\right)\ ,\ \ell\in\{1,...,L-1\}
\end{aligned}
\tag{40}
$$

After introducing these order parameters into the definition of the partition function, we have a factorization of the integrals over each of the N sites in each hidden layer. This gives the following partition function

$$
\begin{aligned}
Z &= \int\prod_{\ell,\mu\alpha,ts}\frac{d\Phi^\ell_{\mu\alpha}(t,s)d\hat{\Phi}^\ell_{\mu\alpha}(t,s)}{2\pi iN^{-1}}\frac{dG_{\mu\alpha}(t,s)d\hat{G}_{\mu\alpha}(t,s)}{2\pi iN^{-1}}\frac{dA^\ell_{\mu\alpha}(t,s)dB^\ell_{\mu\alpha}(t,s)}{2\pi iN^{-1}}\exp\left(NS[\{\Phi,\hat{\Phi},G,\hat{G},A,B\}]\right)\\
S[\{\Phi,\hat{\Phi},G,\hat{G},A,B\}] &= \sum_{\ell\mu\alpha}\int_0^\infty\int_0^\infty dtds\left[\Phi^\ell_{\mu\alpha}(t,s)\hat{\Phi}^\ell_{\mu\alpha}(t,s) + G^\ell_{\mu\alpha}(t,s)\hat{G}^\ell_{\mu\alpha}(t,s) - A^\ell_{\mu\alpha}(t,s)B^\ell_{\mu\alpha}(t,s)\right]\\
&\quad + \ln\mathcal{Z}[\{\Phi,\hat{\Phi},G,\hat{G},A,B,j,v\}]
\end{aligned}
\tag{41}
$$

We thus see that the action S consists of inner-products between order parameters $\{\Phi, G, A\}$ and their duals $\{\hat{\Phi}, \hat{G}, B\}$ as well as a single site MGF $\mathcal{Z}[\{\Phi, \hat{\Phi}, G, \hat{G}, A, B, j, v\}]$, which is defined as

$$\begin{aligned}
\mathcal{Z} = &\int \prod_{\ell\mu t} \frac{d\hat{\chi}^\ell_\mu(t) d\chi^\ell_\mu(t)}{2\pi} \frac{d\hat{\xi}^\ell_\mu(t) d\xi^\ell_\mu(t)}{2\pi} \exp\left(\sum_{\ell\mu} \int_0^\infty dt \left[\left(j^\ell_\mu(t) + i\hat{\chi}^\ell_\mu(t) \right) \chi^\ell_\mu(t) + \left(v^\ell_\mu(t) + i\hat{\xi}^\ell_\mu(t) \right) \xi^\ell_\mu(t) \right] \right) \\
&\times \exp\left(-\frac{1}{2} \sum_{\mu\alpha} \int_0^\infty dt \int_0^\infty ds \hat{\chi}^1_\mu(t) \hat{\chi}^1_\alpha(s) K^x_{\mu\alpha} - \frac{1}{2} \sum_{\mu\alpha} \int_0^\infty dt \int_0^\infty ds \hat{\xi}^L_\mu(t) \hat{\xi}^L_\alpha(s) \right) \\
&\times \exp\left(-\frac{1}{2} \sum_{\ell=1}^{L-1} \sum_{\mu\alpha} \int_0^\infty dt \int_0^\infty ds \left[\hat{\chi}^{\ell+1}_\mu(t) \hat{\chi}^{\ell+1}_\alpha(s) \Phi^\ell_{\mu\alpha}(t,s) + \hat{\xi}^\ell_\mu(t) \hat{\xi}^\ell_\alpha(s) G^{\ell+1}_{\mu\alpha}(t,s) \right] \right) \\
&\times \exp\left(-\sum_{\ell=1}^{L} \sum_{\mu\alpha} \int_0^\infty dt \int_0^\infty ds \left[\phi(h^\ell_\mu(t)) \phi(h^\ell_\alpha(s)) \hat{\Phi}^\ell_{\mu\alpha}(t,s) + g^\ell_\mu(t) g^\ell_\alpha(s) \hat{G}^\ell_{\mu\alpha}(t,s) \right] \right) \\
&\times \exp\left(-i \sum_{\ell=1}^{L} \sum_{\mu\alpha} \int_0^\infty dt \int_0^\infty ds \left[\phi(h^\ell_\mu(t)) \hat{\xi}^\ell_\alpha(s) B^\ell_{\mu\alpha}(t,s) + \hat{\chi}^{\ell+1}_\mu(t) g^{\ell+1}_\alpha(s) A^\ell_{\mu\alpha}(t,s) \right] \right)
\end{aligned} \tag{42}$$

D.5 Saddle Point Equations

Since the integrand in the moment generating function Z takes the form $e^{NS[\{\Phi,\hat{\Phi},G,\hat{G},A,B\}]}$, the $N \to \infty$ limit can be obtained from saddle point integration, also known as the method of steepest descent [86]. This consists in finding order parameters $\{\Phi, \hat{\Phi}, G, \hat{G}, A, B\}$ which render the action S locally stationary. Concretely, this leads to the following saddle point equations.

$$\begin{aligned}
\frac{\delta S}{\delta \hat{\Phi}^\ell_{\mu\alpha}(t,s)} &= \Phi^\ell_{\mu\alpha}(t,s) + \frac{1}{\mathcal{Z}} \frac{\delta \mathcal{Z}}{\delta \hat{\Phi}^\ell_{\mu\alpha}(t,s)} = \Phi^\ell_{\mu\alpha}(t,s) - \left\langle \phi(h^\ell_\mu(t)) \phi(h^\ell_\alpha(s)) \right\rangle = 0 \\
\frac{\delta S}{\delta \Phi^\ell_{\mu\alpha}(t,s)} &= \hat{\Phi}^\ell_{\mu\alpha}(t,s) + \frac{1}{\mathcal{Z}} \frac{\delta \mathcal{Z}}{\delta \Phi^\ell_{\mu\alpha}(t,s)} = \hat{\Phi}^\ell_{\mu\alpha}(t,s) - \frac{1}{2} \left\langle \hat{\chi}^{\ell+1}_\mu(t) \hat{\chi}^{\ell+1}_\alpha(s) \right\rangle = 0 \\
\frac{\delta S}{\delta \hat{G}^\ell_{\mu\alpha}(t,s)} &= G^\ell_{\mu\alpha}(t,s) + \frac{1}{\mathcal{Z}} \frac{\delta \mathcal{Z}}{\delta \hat{G}^\ell_{\mu\alpha}(t,s)} = G^\ell_{\mu\alpha}(t,s) - \left\langle g^\ell_\mu(t) g^\ell_\alpha(s) \right\rangle = 0 \\
\frac{\delta S}{\delta G^\ell_{\mu\alpha}(t,s)} &= \hat{G}^\ell_{\mu\alpha}(t,s) + \frac{1}{\mathcal{Z}} \frac{\delta \mathcal{Z}}{\delta G^\ell_{\mu\alpha}(t,s)} = \hat{G}^\ell_{\mu\alpha}(t,s) - \frac{1}{2} \left\langle \hat{g}^\ell_\mu(t) \hat{g}^\ell_\alpha(s) \right\rangle = 0 \\
\frac{\delta S}{\delta A^\ell_{\mu\alpha}(t,s)} &= -B^\ell_{\mu\alpha}(t,s) + \frac{1}{\mathcal{Z}} \frac{\delta \mathcal{Z}}{\delta A^\ell_{\mu\alpha}(t,s)} = -B^\ell_{\mu\alpha}(t,s) - i \left\langle \hat{\chi}^{\ell+1}_\mu(t) g^{\ell+1}_\alpha(s) \right\rangle = 0 \\
\frac{\delta S}{\delta B^\ell_{\mu\alpha}(t,s)} &= -A^\ell_{\mu\alpha}(t,s) + \frac{1}{\mathcal{Z}} \frac{\delta \mathcal{Z}}{\delta B^\ell_{\mu\alpha}(t,s)} = -A^\ell_{\mu\alpha}(t,s) - i \left\langle \phi(h^\ell_\mu(t)) \hat{\xi}^\ell_\alpha(s) \right\rangle = 0
\end{aligned} \tag{43}$$

We use the notation $\langle\rangle$ to denote an average over the self-consistent distribution on fields induced by the single-site moment generating function $\mathcal{Z}$ at the saddle point. Concretely if $\mathcal{Z} = \int d\chi d\xi d\hat{\chi} d\hat{\xi} \exp\left(-\mathcal{H}[\chi, \xi, \hat{\chi}, \hat{\xi}] \right)$ then the single-site self-consistent average of observable $O([\chi, \xi, \hat{\chi}, \hat{\xi}])$ is defined as

$$\left\langle O([\chi, \xi, \hat{\chi}, \hat{\xi}]) \right\rangle = \frac{1}{\mathcal{Z}} \int d\chi d\xi d\hat{\chi} d\hat{\xi}\, O([\chi, \xi, \hat{\chi}, \hat{\xi}]) \exp\left(-\mathcal{H}[\chi, \xi, \hat{\chi}, \hat{\xi}] \right) \tag{44}$$

To calculate the averages of the dual variables such as $\left\langle \hat{\chi}^{\ell+1} \hat{\chi}^{\ell+1} \right\rangle$, it will be convenient to work with vector and matrix notation. We let $\boldsymbol{\chi}^\ell = \text{Vec}\{\chi^\ell_\mu(t)\}_{\mu\in[P], t\in\mathbb{R}_+}$ represent the vectorization of the stochastic process over different samples and times and define the dot product between two of these vectors as $\boldsymbol{a} \cdot \boldsymbol{b} = \sum_{\mu=1}^P \int_0^\infty dt\, a_\mu(t) b_\mu(t)$. We also apply this procedure on the kernels so that $\boldsymbol{\Phi} = \text{Mat}\{\Phi_{\mu\alpha}(t,s)\}_{\mu\alpha\in[P], t,s\in\mathbb{R}_+}$. Matrix vector products take the form $[\boldsymbol{A}\boldsymbol{b}]_{\mu,t} = \int_0^\infty ds \sum_\alpha A_{\mu\alpha}(t,s) b_\alpha(s)$. We can obtain the behavior of $\left\langle \hat{\boldsymbol{\chi}}^{\ell+1}_\mu \hat{\boldsymbol{\chi}}^{\ell+1\top}_\mu \right\rangle$ in terms of primal fields $\{\chi, \xi, h, z\}$ by insertion of a dummy source $\boldsymbol{u}$ into the effective partition function.

$$
\begin{aligned}
\left\langle \hat{\boldsymbol{\chi}}^{\ell+1}\hat{\boldsymbol{\chi}}^{\ell+1}\right\rangle &= -\frac{\partial^2}{\partial \boldsymbol{u}\partial \boldsymbol{u}^\top}\left\langle \exp\left(i\boldsymbol{u}\cdot\hat{\boldsymbol{\chi}}^{\ell+1}\right)\right\rangle|_{\boldsymbol{u}=\boldsymbol{0}} \\
&= -\frac{1}{\mathcal{Z}}\frac{\partial^2}{\partial \boldsymbol{u}\partial \boldsymbol{u}^\top}\int d\boldsymbol{\chi}^{\ell+1}...\exp\left(-\frac{1}{2}\left(\boldsymbol{\chi}^{\ell+1}+\boldsymbol{u}-\boldsymbol{A}^\ell\boldsymbol{g}^{\ell+1}\right)^\top[\boldsymbol{\Phi}^\ell]^{-1}\left(\boldsymbol{\chi}^{\ell+1}+\boldsymbol{u}-\boldsymbol{A}^\ell\boldsymbol{g}^{\ell+1}\right)-...\right) \\
&= [\boldsymbol{\Phi}^\ell]^{-1}-[\boldsymbol{\Phi}^\ell]^{-1}\left\langle\left(\boldsymbol{\chi}^{\ell+1}-\boldsymbol{A}^\ell\boldsymbol{g}^{\ell+1}\right)\left(\boldsymbol{\chi}^{\ell+1}-\boldsymbol{A}^\ell\boldsymbol{g}^{\ell+1}\right)^\top\right\rangle\left[\boldsymbol{\Phi}^\ell\right]^{-1} \qquad (45)
\end{aligned}
$$

Similarly, we can obtain the equation for $\left\langle\hat{\boldsymbol{\xi}}^\ell\hat{\boldsymbol{\xi}}^{\ell\top}\right\rangle$ by inserting a dummy source $\boldsymbol{r}$ and differentiating near zero source

$$
\begin{aligned}
\left\langle\hat{\boldsymbol{\xi}}^\ell\hat{\boldsymbol{\xi}}^\ell\right\rangle &= -\frac{\partial^2}{\partial \boldsymbol{r}\partial \boldsymbol{r}^\top}\left\langle\exp\left(i\hat{\boldsymbol{r}}\cdot\hat{\boldsymbol{\xi}}^\ell\right)\right\rangle|_{\boldsymbol{r}=\boldsymbol{0}} \\
&= [\boldsymbol{G}^{\ell+1}]^{-1}-[\boldsymbol{G}^{\ell+1}]^{-1}\left\langle(\boldsymbol{\xi}^\ell-\boldsymbol{B}^{\ell\top}\boldsymbol{\phi}^\ell)(\boldsymbol{\xi}^\ell-\boldsymbol{B}^{\ell\top}\boldsymbol{\phi}^\ell)^\top\right\rangle[\boldsymbol{G}^{\ell+1}]^{-1} \qquad (46)
\end{aligned}
$$

As we will demonstrate in the next subsection, these correlators must vanish. Lastly, we can calculate the remaining correlators in terms of primal variables

$$
\begin{aligned}
-i\left\langle\hat{\boldsymbol{\chi}}^{\ell+1}\boldsymbol{g}^{\ell+1,\top}\right\rangle &= \frac{\partial}{\partial\boldsymbol{u}}\left\langle\exp\left(-i\hat{\boldsymbol{u}}\cdot\hat{\boldsymbol{\chi}}^{\ell+1}\right)\boldsymbol{g}^{\ell+1\top}\right\rangle = [\boldsymbol{\Phi}^\ell]^{-1}\left\langle(\boldsymbol{\chi}^{\ell+1}-\boldsymbol{A}^\ell\boldsymbol{g}^{\ell+1})\boldsymbol{g}^{\ell+1\top}\right\rangle \\
-i\left\langle\phi(\boldsymbol{h}^\ell)\hat{\boldsymbol{\xi}}^{\ell\top}\right\rangle &= \frac{\partial}{\partial\boldsymbol{r}^\top}\left\langle\phi(\boldsymbol{h})\exp\left(-i\boldsymbol{r}\cdot\hat{\boldsymbol{\xi}}^\ell\right)\right\rangle = \left\langle\boldsymbol{\phi}(\boldsymbol{h}^\ell)(\boldsymbol{\xi}^\ell-\boldsymbol{B}^{\ell\top}\boldsymbol{\phi}(\boldsymbol{h}^\ell))\right\rangle[\boldsymbol{G}^{\ell+1}]^{-1} \qquad (47)
\end{aligned}
$$

D.6 Single Site Stochastic Process: Hubbard Trick

To get a better sense of this distribution, we can now simplify the quadratic forms appearing in $\mathcal{Z}$ using the Hubbard trick [87], which merely relates a Gaussian function to its Fourier transform.

$$
\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}\right) = \int_{\mathbb{R}^d}\frac{d\boldsymbol{u}}{(2\pi)^{d/2}\sqrt{\det\boldsymbol{A}}}\exp\left(-\frac{1}{2}\boldsymbol{u}^\top\boldsymbol{A}^{-1}\boldsymbol{u}-i\boldsymbol{u}\cdot\boldsymbol{x}\right) = \left\langle\exp\left(-i\boldsymbol{u}\cdot\boldsymbol{x}\right)\right\rangle_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{A})} \qquad (48)
$$

Applying this to the quadratic forms in the single-site MGF $\mathcal{Z}$, we get

$$
\begin{aligned}
\exp\left(-\frac{1}{2}\sum_{\mu\alpha}\int_0^\infty dt\int_0^\infty ds\;\hat{\chi}^1_\mu(t)\hat{\chi}^1_\alpha(s)K^x_{\mu\alpha}\right) &= \left\langle\exp\left(-i\sum_\mu\int_0^\infty dt\;u^1_\mu(t)\hat{\chi}^{\ell+1}_\mu(t)\right)\right\rangle_{\{u^1\}\sim\mathcal{GP}(0,\boldsymbol{K}^x\otimes\boldsymbol{1}\boldsymbol{1}^\top)} \\
\exp\left(-\frac{1}{2}\sum_{\mu\alpha}\int_0^\infty dt\int_0^\infty ds\;\hat{\chi}^{\ell+1}_\mu(t)\hat{\chi}^{\ell+1}_\alpha(s)\Phi^\ell_{\mu\alpha}(t,s)\right) &= \left\langle\exp\left(-i\sum_\mu\int_0^\infty dt\;u^{\ell+1}_\mu(t)\hat{\chi}^{\ell+1}_\mu(t)\right)\right\rangle_{\{u^\ell\}\sim\mathcal{GP}(0,\boldsymbol{\Phi}^\ell)} \\
\exp\left(-\frac{1}{2}\sum_{\mu\alpha}\int_0^\infty dt\int_0^\infty ds\;\hat{\xi}^\ell_\mu(t)\hat{\xi}^\ell_\alpha(s)G^{\ell+1}_{\mu\alpha}(t,s)\right) &= \left\langle\exp\left(-i\sum_\mu\int_0^\infty dt\;r^\ell_\mu(t)\hat{\xi}^\ell_\mu(t)\right)\right\rangle_{\{r^\ell\}\sim\mathcal{GP}(0,\boldsymbol{G}^{\ell+1})} \\
\exp\left(-\frac{1}{2}\sum_{\mu\alpha}\int_0^\infty dt\int_0^\infty ds\;\hat{\xi}^L_\mu(t)\hat{\xi}^L_\alpha(s)\right) &= \left\langle\exp\left(-i\sum_\mu\int_0^\infty dt\;r^L_\mu(t)\hat{\xi}^\ell_\mu(t)\right)\right\rangle_{\{r^L\}\sim\mathcal{GP}(0,\boldsymbol{1}\boldsymbol{1}^\top)}
\end{aligned} \qquad (49)
$$

Next, we integrate over all $\hat{\chi}^\ell, \hat{\xi}^\ell$ variables which yield Dirac-delta functions

$$
\begin{aligned}
\int\prod_{\mu t}\frac{d\hat{\chi}^\ell_\mu(t)}{2\pi}\exp\left(i\hat{\boldsymbol{\chi}}^\ell\cdot\left[\boldsymbol{\chi}^\ell-\boldsymbol{u}^\ell-\boldsymbol{A}^{\ell-1}\boldsymbol{g}^\ell\right]\right) &= \delta\left(\boldsymbol{\chi}^\ell-\boldsymbol{u}^\ell-\boldsymbol{A}^{\ell-1}\boldsymbol{g}^\ell\right) \\
\int\prod_{\mu t}\frac{d\hat{\xi}^\ell_\mu(t)}{2\pi}\exp\left(i\hat{\boldsymbol{\xi}}^\ell\cdot\left[\boldsymbol{\xi}^\ell-\boldsymbol{r}^\ell-\boldsymbol{B}^{\ell\top}\phi(\boldsymbol{h}^\ell)\right]\right) &= \delta\left(\boldsymbol{\xi}^\ell-\boldsymbol{r}^\ell-\boldsymbol{B}^{\ell\top}\phi(\boldsymbol{h}^\ell)\right) \qquad (50)
\end{aligned}
$$

To remedy the notational asymmetry, we redefine $\boldsymbol{B}^\ell$ as its transpose $\boldsymbol{B}^\ell \to \boldsymbol{B}^{\ell\top}$. The presence of these delta-functions in the MGF $\mathcal{Z}$ indicate the constraints $\boldsymbol{u}^\ell = \boldsymbol{\chi}^\ell - \boldsymbol{A}^{\ell-1}\boldsymbol{g}^\ell$ and $\boldsymbol{r}^\ell =$

$\boldsymbol{\xi}^\ell - \boldsymbol{B}^\ell \phi(\boldsymbol{h}^\ell)$. We can thus return to the $\hat{\Phi}$ and $\hat{G}$ saddle point equations and verify that these order parameters vanish

$$\hat{\boldsymbol{\Phi}}^\ell = -\frac{1}{2}\left\langle \hat{\boldsymbol{\chi}}^{\ell+1}\hat{\boldsymbol{\chi}}^{\ell+1\top}\right\rangle = \frac{1}{2}[\boldsymbol{\Phi}^\ell]^{-1}\left\langle \left(\boldsymbol{\chi}^{\ell+1} - \boldsymbol{A}^\ell \boldsymbol{g}^{\ell+1}\right)\left(\boldsymbol{\chi}^{\ell+1} - \boldsymbol{A}^\ell \boldsymbol{g}^{\ell+1}\right)^\top\right\rangle \left[\boldsymbol{\Phi}^\ell\right]^{-1} - \frac{1}{2}[\boldsymbol{\Phi}^\ell]^{-1}$$
$$= \frac{1}{2}[\boldsymbol{\Phi}^\ell]^{-1}\left\langle \boldsymbol{u}^{\ell+1}\boldsymbol{u}^{\ell+1\top}\right\rangle \left[\boldsymbol{\Phi}^\ell\right]^{-1} - \frac{1}{2}[\boldsymbol{\Phi}^\ell]^{-1} = 0, \tag{51}$$

since $\left\langle \boldsymbol{u}^{\ell+1}\boldsymbol{u}^{\ell+1\top}\right\rangle = \boldsymbol{\Phi}^\ell$. Following an identical argument, $\hat{\boldsymbol{G}}^\ell = 0$. After this simplification, the single site MGF takes the form

$$\mathcal{Z}[\{\boldsymbol{j}^\ell, \boldsymbol{v}^\ell\}] = \left\langle \int \prod_\ell d\boldsymbol{\chi}^\ell d\boldsymbol{\xi}^\ell \delta\left(\boldsymbol{\chi}^\ell - \boldsymbol{u}^\ell - \boldsymbol{A}^{\ell-1}\boldsymbol{g}^\ell\right)\delta\left(\boldsymbol{\xi}^\ell - \boldsymbol{r}^\ell - \boldsymbol{B}^\ell\phi(\boldsymbol{h}^\ell)\right)\exp\left(i\boldsymbol{j}^\ell\cdot\boldsymbol{\chi}^\ell + i\boldsymbol{v}^\ell\cdot\boldsymbol{\xi}^\ell\right)\right\rangle_{\{\boldsymbol{u}^\ell,\boldsymbol{r}^\ell\}} \tag{52}$$

The interpretation is thus that $\boldsymbol{u}^\ell, \boldsymbol{r}^\ell$ are sampled independently from their respective Gaussian processes and the fields $\boldsymbol{\chi}^\ell$ and $\boldsymbol{\xi}^\ell$ are determined in terms of $\boldsymbol{u}^\ell, \boldsymbol{r}^\ell, \boldsymbol{h}^\ell, \boldsymbol{g}^\ell$. This means that we can apply Stein's Lemma (integration by parts) [88] to simplify the last two saddle point equations

$$\boldsymbol{A}^\ell = \left\langle \phi(\boldsymbol{h}^\ell)\boldsymbol{r}^{\ell\top}\right\rangle [\boldsymbol{G}^{\ell+1}]^{-1} = \left\langle \frac{\partial\phi(\boldsymbol{h}^\ell)}{\partial \boldsymbol{r}^{\ell\top}}\right\rangle \ , \ \boldsymbol{B}^\ell = \left\langle \boldsymbol{g}^{\ell+1}\boldsymbol{u}^{\ell+1}\right\rangle [\boldsymbol{\Phi}^\ell]^{-1} = \left\langle \frac{\partial \boldsymbol{g}^{\ell+1}}{\partial \boldsymbol{u}^{\ell+1\top}}\right\rangle \tag{53}$$

D.7 Final DMFT Equations

We can now close this stochastic process in terms of preactivations h^ℓ and pre-gradients z^ℓ. To match the formulas provided in the main text, we rescale $A^\ell \to A^\ell/\gamma_0 = \mathcal{O}_{\gamma_0}(1)$ and $B^\ell \to B^\ell/\gamma_0 = \mathcal{O}_{\gamma_0}(1)$, which makes it clear that the non-Gaussian corrections to the $h^\ell_\mu(t), z^\ell_\mu(t)$ fields are $\mathcal{O}(\gamma_0)$. After this rescaling, we have the following complete DMFT equations.

$$\begin{aligned}
h^\ell_\mu(t) &= \chi^\ell_\mu(t) + \gamma_0 \int_0^t ds \sum_\alpha \Delta_\alpha(s)\Phi^{\ell-1}_{\mu\alpha}(t,s) z_\alpha(s)\dot{\phi}(h^\ell_\alpha(s)) \\
&= u^\ell_\mu(t) + \gamma_0 \int_0^t ds \sum_\alpha \left[A^{\ell-1}_{\mu\alpha}(t,s) + \Delta_\alpha(s)\Phi^{\ell-1}_{\mu\alpha}(t,s)\right]\dot{\phi}(h^\ell_\alpha(s)) z^\ell_\alpha(s) \\
z^\ell_\mu(t) &= \xi^\ell_\mu(t) + \gamma_0 \int_0^t ds \sum_\alpha \Delta_\alpha(s) G^{\ell+1}_{\mu\alpha}(t,s)\phi(h^\ell_\alpha(s)) \\
&= r^\ell_\mu(t) + \gamma_0 \int_0^t \sum_\alpha \left[B^\ell_{\mu\alpha}(t,s) + \Delta_\alpha(s) G^{\ell+1}_{\mu\alpha}(t,s)\right]\phi(h^\ell_\alpha(s)) \\
\Phi^\ell_{\mu\alpha}(t,s) &= \left\langle \phi(h^\ell_\mu(t))\phi(h^\ell_\alpha(s))\right\rangle \ , \ G^\ell_{\mu\alpha}(t,s) = \left\langle g^\ell_\mu(t) g^\ell_\alpha(s)\right\rangle \\
A^\ell_{\mu\alpha}(t,s) &= \gamma_0^{-1}\left\langle \frac{\delta\phi(h^\ell_\mu(t))}{\delta r^\ell_\alpha(s)}\right\rangle \ , \ B^\ell_{\mu\alpha}(t,s) = \gamma_0^{-1}\left\langle \frac{\delta g^{\ell+1}_\mu(t)}{\delta u^{\ell+1}_\alpha(s)}\right\rangle
\end{aligned}$$

The base cases in the above equations are that $A^0 = B^L = 0$ and $\Phi^0_{\mu\alpha}(t,s) = K^x_{\mu\alpha}$ and $G^{L+1}_{\mu\alpha}(t,s) = 1$. From the above self-consistent equations, one obtains the NTK dynamics and consequently the output predictions of the network with $\frac{\partial f_\mu}{\partial t} = \sum_\alpha \Delta_\alpha(t)\left[\sum_\ell G^{\ell+1}_{\mu\alpha}(t,t)\Phi^\ell_{\mu\alpha}(t,t)\right]$.

D.8 Varying Network Widths and Initialization Scales

In this section, we relax the assumption of network widths being equal while taking all widths to infinity at a fixed ratio. This will allow us to analyze the influence of bottlenecks on the dynamics. We let $N^\ell = a_\ell N$ represent the width of layer ℓ. Without loss of generality, we can choose that $N^L = N$ and proceed by defining order parameters in the usual way

$$\Phi^\ell_{\mu\alpha}(t,s) = \frac{1}{N^\ell}\phi(\boldsymbol{h}^\ell_\mu(t))\cdot\phi(\boldsymbol{h}^\ell_\alpha(s)) \ , \ G^\ell_{\mu\alpha}(t,s) = \frac{1}{N^\ell}\boldsymbol{g}^\ell_\mu(t)\cdot\boldsymbol{g}^\ell_\alpha(s) \tag{55}$$

Since $N^L = N$, the variable $\boldsymbol{g}^L = \sqrt{N^L}\frac{\partial h^{L+1}}{\partial \boldsymbol{h}^L} = \boldsymbol{w}^L \odot \dot{\phi}(\boldsymbol{h}^L) = \mathcal{O}_{N,\gamma}(1)$ as desired. We extend this definition to each layer as before $\boldsymbol{g}^\ell = \sqrt{N^\ell}\frac{\partial h^{L+1}}{\partial \boldsymbol{h}^\ell}$ which again satisfies the recursion

$$\boldsymbol{g}^\ell_\mu(t) = \boldsymbol{z}^\ell_\mu(t) \odot \dot{\phi}(\boldsymbol{h}^\ell_\mu(t)) \ , \ \boldsymbol{z}^\ell_\mu(t) = \frac{1}{\sqrt{N^{\ell+1}}}\boldsymbol{W}^\ell(t)^\top \boldsymbol{g}^{\ell+1}_\mu(t) \tag{56}$$

Now, we need to calculate the dynamics on weights $\boldsymbol{W}^\ell$

$$\begin{aligned}\frac{d}{dt}\boldsymbol{W}^\ell &= \gamma^2 \sum_\mu \Delta_\mu \frac{\partial f_\mu}{\partial \boldsymbol{W}^\ell} = \gamma^2 \sum_\mu \Delta_\mu \frac{\partial f_\mu}{\partial \boldsymbol{h}^{\ell+1}_\mu} \cdot \frac{\partial \boldsymbol{h}^{\ell+1}_\mu}{\partial \boldsymbol{W}^\ell} \\ &= \frac{\gamma}{\sqrt{N^\ell}\sqrt{N^{\ell+1}}} \sum_\mu \Delta_\mu \boldsymbol{g}^{\ell+1}_\mu \phi(\boldsymbol{h}^\ell_\mu)^\top \end{aligned} \tag{57}$$

Using our definition of the kernels and the $\boldsymbol{h}, \boldsymbol{z}$ fields

$$\begin{aligned}\boldsymbol{h}^\ell_\mu(t) &= \boldsymbol{\chi}^\ell_\mu(t) + \frac{\gamma}{\sqrt{N^\ell}} \sum_\alpha \int_0^t ds \ \Delta_\alpha(s) \boldsymbol{g}^\ell_\alpha(s) \Phi^{\ell-1}_{\mu\alpha}(t,s) \\ \boldsymbol{z}^\ell_\mu(t) &= \boldsymbol{\xi}^\ell_\mu(t) + \frac{\gamma}{\sqrt{N^\ell}} \sum_\alpha \int_0^t ds \ \Delta_\alpha(s) \phi(\boldsymbol{h}^\ell_\alpha(s)) G^{\ell+1}_{\mu\alpha}(t,s) \end{aligned} \tag{58}$$

We also find the usual formula for the NTK

$$K^{NTK}_{\mu\alpha} = \gamma^2 \sum_\ell \text{Tr}\left[\frac{\partial f_\mu}{\partial \boldsymbol{W}^\ell}\right]^\top \frac{\partial f_\alpha}{\partial \boldsymbol{W}^\ell} = \Phi^L_{\mu\alpha} + \sum_{\ell=1}^{L-1} G^{\ell+1}_{\mu\alpha}\Phi^\ell_{\mu\alpha} + G^1_{\mu\alpha}K^x_{\mu\alpha} \tag{59}$$

Now, as before, we need to consider the distribution of $\boldsymbol{\chi}, \boldsymbol{\xi}$ fields. We assume $W^\ell_{ij}(0) \sim \mathcal{N}(0, \sigma^2_\ell)$. This requires computing integrals like

$$\begin{aligned}&\left\langle \exp\left(i \sum_\mu \int_0^\infty dt \left[\hat{\boldsymbol{\chi}}^{\ell+1}_\mu(t)^\top \boldsymbol{W}^\ell(0)\phi(\boldsymbol{h}^\ell_\mu(t))/\sqrt{N^\ell} + \boldsymbol{g}^{\ell+1}_\mu(t)^\top \boldsymbol{W}^\ell(0)\hat{\boldsymbol{\xi}}^\ell_\mu(t)/\sqrt{N^{\ell+1}} \right] \right) \right\rangle_{\boldsymbol{W}^\ell(0)} \\ &= \exp\left(-\frac{\sigma^2_\ell}{2} \sum_{\mu\alpha} \int_0^\infty dt \int_0^\infty ds \left[\hat{\boldsymbol{\chi}}^{\ell+1}_\mu(t) \cdot \hat{\boldsymbol{\chi}}^{\ell+1}_\mu(t) \Phi^\ell_{\mu\alpha}(t,s) + \hat{\boldsymbol{\xi}}^\ell_\mu(t) \cdot \hat{\boldsymbol{\xi}}^\ell_\mu(t) G^{\ell+1}_{\mu\alpha}(t,s) \right] \right) \\ &\times \exp\left(-i\sigma^2_\ell \sqrt{\frac{a_\ell}{a_{\ell+1}}} \sum_{\mu\alpha} \int_0^\infty dt \int_0^\infty ds A^\ell_{\mu\alpha}(t,s) \boldsymbol{\chi}^{\ell+1}_\mu(t) \cdot \boldsymbol{g}^{\ell+1}_\alpha(s) \right) \end{aligned} \tag{60}$$

where $\boldsymbol{A}^\ell_{\mu\alpha}(t,s) = -\frac{i}{N^\ell}\boldsymbol{\phi}(\boldsymbol{h}^\ell_\mu(t)) \cdot \hat{\boldsymbol{\xi}}^\ell_\alpha(s)$. The action thus takes the form

$$S = \sum_\ell a_\ell \text{Tr}\left[\hat{\boldsymbol{\Phi}}^{\ell\top}\boldsymbol{\Phi}^\ell + \boldsymbol{G}^{\ell\top}\hat{\boldsymbol{G}}^\ell - \boldsymbol{A}^{\ell\top}\boldsymbol{B}^\ell \right] + \sum_\ell a_\ell \ln \mathcal{Z}_\ell \tag{61}$$

where the zero-source MGF for layer ℓ has the form

$$\begin{aligned}\mathcal{Z}_\ell = \int \prod_{\mu t} \frac{d\chi^\ell_\mu(t) d\hat{\chi}^\ell_\mu(t)}{2\pi} \frac{d\xi^\ell_\mu(t) d\hat{\xi}^\ell_\mu(t)}{2\pi} &\exp\left(-\phi(\boldsymbol{h}^\ell)^\top \hat{\boldsymbol{\Phi}}^\ell \phi(\boldsymbol{h}^\ell) - \boldsymbol{g}^{\ell\top}\hat{\boldsymbol{G}}^\ell \boldsymbol{g}^\ell + i\boldsymbol{\chi}^\ell \cdot \hat{\boldsymbol{\chi}}^\ell + i\boldsymbol{\xi}^\ell \cdot \hat{\boldsymbol{\xi}}^\ell \right) \\ &\exp\left(-\frac{\sigma^2_{\ell-1}}{2} \hat{\boldsymbol{\chi}}^\ell \boldsymbol{\Phi}^{\ell-1} \hat{\boldsymbol{\chi}}^\ell - \frac{\sigma^2_\ell}{2} \hat{\boldsymbol{\xi}}^\ell \boldsymbol{G}^{\ell+1} \hat{\boldsymbol{\xi}}^\ell \right) \\ &\exp\left(-i\sigma^2_{\ell-1}\sqrt{\frac{a_{\ell-1}}{a_\ell}} \hat{\boldsymbol{\chi}}^\ell \boldsymbol{A}^{\ell-1}\boldsymbol{g}^\ell - i\phi(\boldsymbol{h}^\ell)^\top \boldsymbol{B}^\ell \hat{\boldsymbol{\xi}}^\ell \right) \end{aligned} \tag{62}$$

The saddle point equations give

$$\begin{aligned}&\boldsymbol{\Phi}^\ell = \left\langle \phi(\boldsymbol{h}^\ell)\phi(\boldsymbol{h}^\ell)^\top \right\rangle \ , \ \boldsymbol{G}^\ell = \left\langle \boldsymbol{g}^\ell \boldsymbol{g}^{\ell\top} \right\rangle \\ &\boldsymbol{A}^\ell = -i\left\langle \phi(\boldsymbol{h}^\ell)\hat{\boldsymbol{\xi}}^{\ell\top} \right\rangle = \left\langle \frac{\partial \phi(\boldsymbol{h}^\ell)}{\partial \boldsymbol{r}^{\ell\top}} \right\rangle \\ &a_\ell \boldsymbol{B}^\ell = -ia_{\ell+1}\sigma^2_\ell \sqrt{\frac{a_\ell}{a_{\ell+1}}} \left\langle \hat{\boldsymbol{\chi}}^{\ell+1} \boldsymbol{g}^{\ell+1,\top} \right\rangle \implies \boldsymbol{B}^\ell = \sigma^2_\ell \sqrt{\frac{a_{\ell+1}}{a_\ell}} \left\langle \frac{\partial \boldsymbol{g}^{\ell+1\top}}{\partial \boldsymbol{u}^{\ell+1}} \right\rangle \end{aligned} \tag{63}$$

where $\boldsymbol{u}^\ell \sim \mathcal{GP}(0, \sigma_{\ell-1}^2 \boldsymbol{\Phi}^{\ell-1}), \boldsymbol{r}^\ell \sim \mathcal{GP}(0, \sigma_\ell^2 \boldsymbol{G}^{\ell+1})$. We redefine $\boldsymbol{B}^\ell \to \frac{1}{\sigma_\ell^2}\sqrt{\frac{a_\ell}{a_{\ell+1}}}\boldsymbol{B}^\ell$. To take the $N \to \infty$ limit of the field dynamics, again use $\gamma_0 = \gamma/\sqrt{N} = O_N(1)$. The field equations take the form

$$
\begin{aligned}
h_\mu^\ell(t) &= u_\mu^\ell(t) + \int_0^\infty \sum_{\alpha=1}^P \left[\sigma_{\ell-1}^2 \sqrt{\frac{a_{\ell-1}}{a_\ell}} A_{\mu\alpha}^{\ell-1}(t,s) + \frac{\gamma_0}{\sqrt{a_\ell}}\Theta(t-s)\Phi_{\mu\alpha}^{\ell-1}(t,s)\right] \dot\phi(h_\alpha^\ell(s)) z_\alpha^\ell(s) \\
z_\mu^\ell(t) &= r_\mu^\ell(t) + \int_0^\infty \sum_{\alpha=1}^P \left[\sigma_\ell^2 \sqrt{\frac{a_{\ell+1}}{a_\ell}} B_{\mu\alpha}^{\ell}(t,s) + \frac{\gamma_0}{\sqrt{a_\ell}}\Theta(t-s)G_{\mu\alpha}^{\ell+1}(t,s)\right] \phi(h_\alpha^\ell(s))
\end{aligned} \tag{64}
$$

We thus find that the evolution of the scalar fields in a given layer is set by the parameter $\gamma_0/\sqrt{a_\ell}$, indicating that relatively wider layers evolve less and contribute less of a change to the overall NTK. This definition for $\boldsymbol{A}^\ell, \boldsymbol{B}^\ell$ is non-ideal to extract intuition about bottlenecks since $\boldsymbol{A}^{\ell-1} \sim \mathcal{O}\left(\frac{\gamma_0}{\sqrt{a_{\ell-1}}}\right)$ and $\boldsymbol{B}^\ell \sim \mathcal{O}\left(\frac{\gamma_0}{\sqrt{a_{\ell+1}}}\right)$. To remedy this, we redefine $\tilde{\boldsymbol{A}}^\ell = \frac{\sqrt{a_\ell}}{\gamma_0}\boldsymbol{A}^\ell, \tilde{\boldsymbol{B}}^\ell = \frac{\sqrt{a_{\ell+1}}}{\gamma_0}\boldsymbol{B}^\ell$. With this choice, we have

$$
\begin{aligned}
h_\mu^\ell(t) &= u_\mu^\ell(t) + \frac{\gamma_0}{\sqrt{a_\ell}} \int_0^\infty \sum_{\alpha=1}^P \left[\sigma_{\ell-1}^2 \tilde{A}_{\mu\alpha}^{\ell-1}(t,s) + \Theta(t-s)\Phi_{\mu\alpha}^{\ell-1}(t,s)\right] \dot\phi(h_\alpha^\ell(s)) z_\alpha^\ell(s) \\
z_\mu^\ell(t) &= r_\mu^\ell(t) + \frac{\gamma_0}{\sqrt{a_\ell}} \int_0^\infty \sum_{\alpha=1}^P \left[\sigma_\ell^2 \tilde{B}_{\mu\alpha}^{\ell}(t,s) + \Theta(t-s)G_{\mu\alpha}^{\ell+1}(t,s)\right] \phi(h_\alpha^\ell(s))
\end{aligned} \tag{65}
$$

where $\tilde{A}^{\ell-1}, \tilde{B}^\ell$ do not have a leading order scaling with $a_{\ell-1}$ or $a_{\ell+1}$ respectively. Under this change of variables, it is now apparent that a very wide layer ℓ, where $\frac{\gamma_0}{\sqrt{a_\ell}} \ll 1$ is small, the fields h^ℓ, z^ℓ become well approximated by the Gaussian processes u^ℓ, r^ℓ, albeit with evolving covariances $\boldsymbol{\Phi}^{\ell-1}, \boldsymbol{G}^{\ell+1}$ respectively. In a realistic CNN architecture where the number of channels increases across layers, this result would predict that more feature learning and deviations from Gaussianity to occur in the early layers and the later layers to be well approximated as Gaussian fields u^ℓ, r^ℓ with temporally evolving covariances for $\ell \sim L$. We leave evaluation of this prediction to future work.

E Two Layer Networks

In a two layer network, there are no $\boldsymbol{A}$ or $\boldsymbol{B}$ order parameters, so the fields χ^1 and ξ^1 are always independent. Further, χ^1 and ξ^1 are both constant throughout training dynamics. Thus we can obtain differential rather than integral equations for the stochastic fields h^1, z^1 which are

$$
\begin{aligned}
&\frac{\partial}{\partial t} h_\mu^1(t) = \gamma_0 \sum_{\alpha=1}^P \Delta_\alpha(t) K_{\mu\alpha}^x \dot\phi(h_\alpha^1(t)) z^1(t)\ ,\ \frac{\partial}{\partial t} z^1(t) = \gamma_0 \sum_{\alpha=1}^P \Delta_\alpha(t)\phi(h_\alpha^1(t)) \\
&\Phi_{\mu\alpha}^1(t) = \left\langle \phi(h_\mu^1(t))\phi(h_\alpha^1(t))\right\rangle\ ,\ G_{\mu\alpha}^1(t) = \left\langle z(t)^2 \dot\phi(h_\mu^1(t))\dot\phi(h_\alpha^1(t))\right\rangle \\
&\frac{\partial}{\partial t}\Delta_\mu(t) = -\sum_{\alpha=1}^P \left[G_{\mu\alpha}^1(t) K_{\mu\alpha}^x + \Phi_{\mu\alpha}^1(t)\right]\Delta_\alpha(t)
\end{aligned} \tag{66}
$$

where the average is taken over the random initial conditions $\boldsymbol{h}^1(0) \sim \mathcal{N}(0, \boldsymbol{K}^x)$ and $\boldsymbol{z}^1(0) \sim \mathcal{N}(0, \mathbf{1}\mathbf{1}^\top)$. An example of the two layer theory for a ReLU network can be found in Appendix Figure 6. In this two layer setting, a drift PDE can be obtained for the joint density of preactivations

and feedback fields $p(\boldsymbol{h}, z; t)$

$$\frac{\partial}{\partial t} p(\boldsymbol{h}, z, t) = -p(\boldsymbol{h}, z, t) z(t) \sum_{\mu} \Delta_\mu(t) K^x_{\mu\mu} \ddot{\phi}(h_\mu(t))$$
$$- \gamma_0 \sum_{\mu\alpha} K^x_{\mu\alpha} \Delta_\alpha \dot{\phi}(h_\alpha(t)) z(t) \frac{\partial p(\boldsymbol{h}, \boldsymbol{z}, t)}{\partial h_\mu} - \gamma_0 \sum_{\mu\alpha} \Delta_\alpha \phi(h_\alpha) \frac{\partial p(\boldsymbol{h}, \boldsymbol{z}, t)}{\partial z_\mu}$$
$$\frac{\partial}{\partial t} \Delta_\mu(t) = -\sum_{\alpha=1}^{P} \left[G^1_{\mu\alpha}(t) K^x_{\mu\alpha} + \Phi^1_{\mu\alpha}(t) \right] \Delta_\alpha(t)$$
$$\Phi^1_{\mu\alpha}(t) = \left\langle \phi(h^1_\mu(t)) \phi(h^1_\alpha(t)) \right\rangle \ , \ G^1_{\mu\alpha}(t) = \left\langle z^1(t)^2 \dot{\phi}(h^1_\mu(t)) \dot{\phi}(h^1_\alpha(t)) \right\rangle , \tag{67}$$

which is a zero-diffusion feature space version of the PDE derived in the original two layer mean field limit of neural networks [22, 42, 43].

F Deep Linear Networks

In the deep linear case, the $g^\ell_\mu(t)$ fields are independent of sample index μ. We introduce the kernel $H^\ell_{\mu\alpha}(t, s) = \left\langle h^\ell_\mu(t) h^\ell_\alpha(s) \right\rangle$. The field equations are

$$h^\ell_\mu(t) = u^\ell_\mu(t) + \gamma_0 \int_0^\infty \sum_{\alpha=1}^{P} \left[A^{\ell-1}_{\mu\alpha}(t, s) + \Theta(t-s) H^{\ell-1}_{\mu\alpha}(t, s) \right] \Delta_\alpha(s) g^\ell(s)$$
$$g^\ell(t) = r^\ell(t) + \gamma_0 \int_0^\infty \sum_{\alpha=1}^{P} [B^\ell_\alpha(t, s) + \gamma_0 \Theta(t-s) G^{\ell+1}(t, s)] \Delta_\alpha(s) h^\ell_\alpha(s) \tag{68}$$

Or in vector notation $\boldsymbol{h}^\ell = \boldsymbol{u}^\ell + \gamma_0 \boldsymbol{C}^\ell \boldsymbol{g}^\ell$ and $\boldsymbol{g}^\ell = \boldsymbol{r}^\ell + \gamma_0 \boldsymbol{D}^\ell \boldsymbol{h}^\ell$ where

$$C^\ell_\mu(t, s) = \sum_{\alpha=1}^{P} [A^{\ell-1}_{\mu\alpha}(t, s) + \Theta(t-s) H^{\ell-1}_{\mu\alpha}(t, s)] \Delta_\alpha(s) \ , \ D^\ell_\mu(t, s) = [B^\ell_\mu(t, s) + \Theta(t-s) G^{\ell+1}(t, s)] \Delta_\mu(s) \tag{69}$$

Using the formulas which define the fields, we have

$$\boldsymbol{h}^\ell = \boldsymbol{u}^\ell + \gamma_0 \boldsymbol{C}^\ell \boldsymbol{r}^\ell + \gamma_0^2 \boldsymbol{C}^\ell \boldsymbol{D}^\ell \boldsymbol{h}^\ell \implies \boldsymbol{h}^\ell = (\mathbf{I} - \gamma_0^2 \boldsymbol{C}^\ell \boldsymbol{D}^\ell)^{-1} [\boldsymbol{u}^\ell + \gamma_0 \boldsymbol{C}^\ell \boldsymbol{r}^\ell]$$
$$\boldsymbol{g}^\ell = \boldsymbol{r}^\ell + \gamma_0 \boldsymbol{D}^\ell \boldsymbol{u}^\ell + \gamma_0^2 \boldsymbol{D}^\ell \boldsymbol{C}^\ell \boldsymbol{g}^\ell \implies \boldsymbol{g}^\ell = (\mathbf{I} - \gamma_0^2 \boldsymbol{D}^\ell \boldsymbol{C}^\ell)^{-1} [\boldsymbol{r}^\ell + \gamma_0 \boldsymbol{D}^\ell \boldsymbol{u}^\ell] \tag{70}$$

The saddle point equations can thus be written as

$$\boldsymbol{H}^\ell = \left\langle \boldsymbol{h}^\ell \boldsymbol{h}^{\ell\top} \right\rangle = (\mathbf{I} - \gamma_0^2 \boldsymbol{C}^\ell \boldsymbol{D}^\ell)^{-1} [\boldsymbol{H}^{\ell-1} + \gamma_0^2 \boldsymbol{C}^\ell \boldsymbol{G}^{\ell+1} \boldsymbol{C}^{\ell\top}] \left[(\mathbf{I} - \gamma_0^2 \boldsymbol{C}^\ell \boldsymbol{D}^\ell)^\top \right]^{-1}$$
$$\boldsymbol{G}^\ell = \left\langle \boldsymbol{g}^\ell \boldsymbol{g}^{\ell\top} \right\rangle = \left(\mathbf{I} - \gamma_0^2 \boldsymbol{D}^\ell \boldsymbol{C}^\ell \right)^{-1} \left[\boldsymbol{G}^{\ell+1} + \gamma_0^2 \boldsymbol{D}^\ell \boldsymbol{H}^{\ell-1} \boldsymbol{D}^{\ell\top} \right] \left[\left(\mathbf{I} - \gamma_0^2 \boldsymbol{D}^\ell \boldsymbol{C}^\ell \right)^\top \right]^{-1}$$
$$\boldsymbol{A}^\ell = (\mathbf{I} - \gamma_0^2 \boldsymbol{C}^\ell \boldsymbol{D}^\ell)^{-1} \boldsymbol{C}^\ell \ , \ \boldsymbol{B}^{\ell-1} = (\mathbf{I} - \gamma_0^2 \boldsymbol{D}^\ell \boldsymbol{C}^\ell)^{-1} \boldsymbol{D}^\ell \tag{71}$$

We solve these equations by repeatedly updating $\boldsymbol{H}^\ell, \boldsymbol{G}^\ell$, using Equation (71) and the current estimate of $\boldsymbol{C}^\ell, \boldsymbol{D}^\ell$. We then use the new $\boldsymbol{H}^\ell, \boldsymbol{G}^\ell$ to recompute $\boldsymbol{K}^{NTK}$ and $\boldsymbol{\Delta}(t)$, calculating $\boldsymbol{C}^\ell, \boldsymbol{D}^\ell$ and then recomputing $\boldsymbol{H}^\ell, \boldsymbol{G}^\ell$. This procedure usually converges in $\sim 5 - 10$ steps.

F.1 Two Layer Linear Network

As we saw in Appendix E, the field dynamics simplify considerably in the two layer case, allowing description of all fields in terms of differential equations. In a two layer linear network, we let $\boldsymbol{h}(t) \in \mathbb{R}^P$ represent the hidden activation field and $g(t) \in \mathbb{R}$ represent the gradient

$$\frac{\partial}{\partial t} \boldsymbol{h}(t) = \gamma_0 g(t) \boldsymbol{K}^x \boldsymbol{\Delta}(t) \ , \ \frac{\partial}{\partial t} g(t) = \gamma_0 \boldsymbol{\Delta}(t) \cdot \boldsymbol{h}(t) \tag{72}$$

The kernels $\boldsymbol{H}(t) = \left\langle \boldsymbol{h}(t)\boldsymbol{h}(t)^\top \right\rangle$ and $G(t) = \left\langle g(t)^2 \right\rangle$ thus evolve as

$$\frac{\partial}{\partial t}\boldsymbol{H}(t) = \gamma_0 \boldsymbol{K}^x \boldsymbol{\Delta} \left\langle g(t)\boldsymbol{h}(t)^\top \right\rangle + \gamma_0 \left\langle g(t)\boldsymbol{h}(t) \right\rangle \boldsymbol{\Delta}^\top \boldsymbol{K}^x$$
$$\frac{\partial}{\partial t}G(t) = 2\gamma_0 \left\langle g(t)\boldsymbol{h}(t) \right\rangle \cdot \boldsymbol{\Delta}(t) \tag{73}$$

It is easy to verify that the network predictions on the P training points are $\boldsymbol{f}(t) = \boldsymbol{y} - \boldsymbol{\Delta}(t) = \frac{1}{\gamma_0}\left\langle g(t)\boldsymbol{h}(t) \right\rangle \in \mathbb{R}^P$. Thus the dynamics of $\boldsymbol{H}(t), G(t)$ and $\boldsymbol{\Delta}(t)$ close

$$\frac{\partial}{\partial t}\boldsymbol{H}(t) = \gamma_0^2 \boldsymbol{K}^x \boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\Delta})^\top + \gamma_0^2 (\boldsymbol{y} - \boldsymbol{\Delta})\boldsymbol{\Delta}^\top \boldsymbol{K}^x$$
$$\frac{\partial}{\partial t}G(t) = 2\gamma_0^2 (\boldsymbol{y} - \boldsymbol{\Delta}) \cdot \boldsymbol{\Delta}(t)$$
$$\frac{\partial}{\partial t}\boldsymbol{\Delta}(t) = -[\boldsymbol{H}(t) + G(t)\boldsymbol{K}^x]\boldsymbol{\Delta}(t) \tag{74}$$

where the initial conditions are $\boldsymbol{H}(0) = \boldsymbol{I}$, $G(0) = 1$ and $\boldsymbol{\Delta}(0) = \boldsymbol{y}$. These equations hold for any choice of data $\boldsymbol{K}^x, \boldsymbol{y}$.

F.1.1 Whitened Data in Two Layer Linear

For input data which is whitened where $\boldsymbol{K}^x = \mathbf{I}$, then the dynamics can be simplified even further, recovering the sigmoidal curves very similar to those obtained under a special initialization [68, 69, 71, 73]. In this case we note that the error signal always evolves in the $\boldsymbol{y}$ direction, $\boldsymbol{\Delta}(t) = \Delta(t)\frac{\boldsymbol{y}}{|\boldsymbol{y}|}$, and that $\boldsymbol{H}$ only evolves in a rank one direction $\boldsymbol{y}\boldsymbol{y}^\top$ direction as well. Let $\frac{1}{|\boldsymbol{y}|^2}\boldsymbol{y}^\top \boldsymbol{H}(t)\boldsymbol{y} = H_y(t)$. Let $y = |\boldsymbol{y}|$ represent the norm of the target vector, then the relevant scalar dynamics are

$$\frac{\partial}{\partial t}H_y(t) = 2\gamma_0^2 \Delta(t)(y - \Delta(t)) \ , \ \frac{\partial}{\partial t}G(t) = 2\gamma_0^2 \Delta(t)(y - \Delta(t))$$
$$\frac{\partial}{\partial t}\Delta(t) = -[H_y(t) + G(t)]\Delta(t) \tag{75}$$

Now note that, at initialization $H_y(0) = G(0) = 1$ and that $\frac{\partial}{\partial t}H_y(t) = \frac{\partial}{\partial t}G(t)$. Thus, we have an automatic balancing condition $H_y(t) = G(t)$ for all $t \in \mathbb{R}_+$ and the dynamics reduce to two variables

$$\frac{\partial}{\partial t}H_y(t) = 2\gamma_0^2 \Delta(t)(y - \Delta(t)) \ , \ \frac{\partial}{\partial t}\Delta(t) = -2H_y(t)\Delta(t) \tag{76}$$

We note that this system obeys a conservation law which constrains $(H_y, y - \Delta)$ to a hyperbola

$$\frac{1}{2}\frac{\partial}{\partial t}\left[H_y^2 - \gamma_0^2 (y - \Delta(t))^2\right] = 2\gamma_0^2 H_y \Delta(y - \Delta) - 2\gamma_0^2 H_y \Delta(y - \Delta) = 0 \tag{77}$$

This conservation law implies that $H_y(0)^2 = 1 = \lim_{t\to\infty} H_y(t)^2 - \gamma_0^2 y^2$ or that the final kernel has the form $\lim_{t\to\infty} \boldsymbol{H}(t) = \frac{1}{y^2}[\sqrt{1 + \gamma_0^2 y^2} - 1]\boldsymbol{y}\boldsymbol{y}^\top + \mathbf{I}$. The result that the final kernel becomes a rank one spike in the direction of the target function was also obtained in finite width networks in the limit of small initialization [73] and also from a normative toy model of feature learning [81]. We can use the conservation law above $1 = H_y(t)^2 - \gamma_0^2(\Delta(t) - y)^2$ to simplify the dynamics to a one dimensional system

$$\frac{\partial}{\partial t}\Delta(t) = -2\sqrt{1 + \gamma_0^2(\Delta(t) - y)^2}\ \Delta(t) \implies \frac{\partial}{\partial t}f = 2\sqrt{1 + \gamma_0^2 f^2}(y - f) \tag{78}$$

where $f = y - \Delta$. We see that increasing γ_0 provides strict acceleration in the learning dynamics, illustrating the training benefits of feature evolution. Since this system is separable, we can solve for the time it takes for the network output norm to reach output level f

$$2t = \int_0^f \frac{ds}{(y - s)\sqrt{1 + \gamma_0^2 s^2}} = \frac{1}{\sqrt{1 + \gamma_0^2 y^2}} \tanh^{-1}\left(\frac{1 + \gamma_0^2 y f}{\sqrt{1 + \gamma_0^2 y^2}\sqrt{1 + \gamma_0^2 f^2}}\right)$$
$$- \frac{1}{\sqrt{1 + \gamma_0^2 y^2}} \tanh^{-1}\left(\frac{1}{\sqrt{1 + \gamma_0^2 y^2}}\right) \tag{79}$$

The NTK limit can be obtained by taking $\gamma_0 \to 0$ which gives

$$\frac{\partial}{\partial t}\Delta(t) \sim -2\Delta(t) \implies \Delta(t) \sim e^{-2t} \tag{80}$$

which recovers the usual convergence rate of a linear model. The right hand side of Equation (79) has a perturbation series in γ_0^2 which converges in the disk $\gamma_0 < \frac{1}{y}$. The other limit of interest is the $\gamma_0 \to \infty$ limit where

$$\frac{d}{dt}\Delta(t) \sim -2\gamma_0(y - \Delta(t))\Delta(t) \tag{81}$$

which recovers the logistic growth observed in the initialization scheme of prior works [68, 69]. The timescale τ required to learn is only $\tau \sim \frac{1}{\gamma_0} \ll 1$, which is much smaller than the $O_{\gamma_0}(1)$ time to learn predicted from the small γ_0 expansion. We note that the above leading order asymptotic behavior at large γ_0 considers the DMFT initial condition $\Delta(0) = y$ as an unstable fixed point. For realistic learning curves, one would need to stipulate some alternative initial condition such as $\Delta = y - \epsilon$ for some small $\epsilon > 0$ in order to have nontrivial leading order dynamics.

F.2 Deep Linear Whitened Data

In this section, we examine the role of depth when linear networks are trained on whitened data. As in the two layer case, all hidden kernels $\boldsymbol{H}^\ell(t,s)$ need only be tracked in the one dimensional task relevant subspace along the vector $\boldsymbol{y}$. We let $\Delta(t) = \frac{1}{y}\boldsymbol{y} \cdot \boldsymbol{\Delta}(t)$ and let $h_y(t) = \frac{1}{y}\boldsymbol{h}^\ell(t) \cdot \boldsymbol{y}$. We have

$$\begin{aligned} h_y^\ell(t) &= u_y^\ell(t) + \gamma_0 \int_0^\infty ds\, C^\ell(t,s) g^\ell(s)\ ,\ C^\ell(t,s) = A_y^{\ell-1}(t,s) + \Theta(t-s) H_y^{\ell-1}(t,s)\Delta(s) \\ g^\ell(t) &= r^\ell(t) + \gamma_0 \int_0^\infty ds\, D^\ell(t,s) h_y^\ell(s)\ ,\ D^\ell(t,s) = B_y^{\ell-1}(t,s) + \Theta(t-s) G^{\ell+1}(t,s)\Delta(s) \end{aligned} \tag{82}$$

Lastly we have the simple evolution equation for the scalar error $\Delta(t)$

$$\frac{\partial \Delta(t)}{\partial t} = -\sum_{\ell=0}^{L} G^{\ell+1}(t,t) H_y^\ell(t,t)\Delta(t) \implies \Delta(t) = \exp\left(-\int_0^t ds \sum_{\ell=0}^{L} G^{\ell+1}(s,s) H_y^\ell(s,s)\right) y \tag{83}$$

Vectorizing we find the following equations for the time $\times$ time matrix order parameters $\boldsymbol{h}^\ell = \boldsymbol{u}^\ell + \gamma_0 \boldsymbol{C}^\ell \boldsymbol{g}^\ell$, $\boldsymbol{g}^\ell = \boldsymbol{r}^\ell + \gamma_0 \boldsymbol{D}^\ell \boldsymbol{h}^\ell$, we can solve for the response functions $\boldsymbol{A}^\ell = \left(\mathbf{I} - \gamma_0^2 \boldsymbol{C}^\ell \boldsymbol{D}^\ell\right)^{-1} \boldsymbol{C}^\ell$ and $\boldsymbol{B}^\ell = \left(\mathbf{I} - \gamma_0^2 \boldsymbol{D}^\ell \boldsymbol{C}^\ell\right)^{-1} \boldsymbol{D}^\ell$. This formulation has the advantage that it no longer has any sample-size dependence: arbitrary sample sizes can be considered with no computational cost.

G Convolutional Networks with Infinite Channels

The DMFT described in this work can be extended to CNNs with infinitely many channels, much in the same way that infinite CNNs have a well defined kernel limit [89, 90]. We let $W_{ij,\mathfrak{a}}^\ell$ represent the value of the filter at spatial displacement $\mathfrak{a}$ from the center of the filter, which maps relates activity at channel j of layer ℓ to channel i of layer $\ell+1$. The fields $h_{\mu,i,\mathfrak{a}}^\ell$ are defined recursively as

$$h_{\mu,i,\mathfrak{a}}^{\ell+1} = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \sum_{\mathfrak{b} \in \mathcal{S}^\ell} W_{ij,\mathfrak{b}}^\ell \phi(h_{\mu,j,\mathfrak{a}+\mathfrak{b}}^\ell) \tag{84}$$

where $\mathcal{S}^\ell$ is the spatial receptive field at layer ℓ. For example, a $(2k+1) \times (2k+1)$ convolution will have $\mathcal{S}^\ell = \{(i,j) \in \mathbb{Z}^2 : -k \le i \le k, -k \le j \le k\}$. The output function is obtained from the last layer is defined as $f_\mu = \frac{1}{\gamma_0 N} \sum_{i=1}^N w_{i,\mathfrak{a}}^L \phi(h_{\mu,i,\mathfrak{a}}^L)$. The gradient fields have the same definition as before $\boldsymbol{g}_{\mu,\mathfrak{a}}^\ell = \gamma_0 N \frac{\partial f_\mu}{\partial \boldsymbol{h}_{\mu,\mathfrak{a}}^\ell}$, which as before enjoy the following recursion

$$\boldsymbol{g}_{\mu,\mathfrak{a}}^\ell = \sum_{\mathfrak{b} \in \mathbb{S}^\ell} \dot{\phi}(\boldsymbol{h}_{\mu,\mathfrak{b}+\mathfrak{a}}^\ell) \odot \left[\frac{1}{\sqrt{N}} \boldsymbol{W}_{\mathfrak{b}}^{\ell\top} \boldsymbol{g}_{\mu,\mathfrak{a}+\mathfrak{b}}^{\ell+1}\right] \tag{85}$$

The feature space description of the forward and backward pass relations is

$$\boldsymbol{h}^{\ell+1}_{\mu,\mathfrak{a}}(t) = \boldsymbol{\chi}^{\ell+1}_{\mu,\mathfrak{a}}(t) + \gamma_0 \int_0^t ds \sum_{\alpha\mathfrak{b}} \Delta_\alpha(s) \Phi^\ell_{\mu\alpha,\mathfrak{a},\mathfrak{b}}(t,s) \boldsymbol{g}^{\ell+1}_{\alpha,\mathfrak{b}-\mathfrak{a}}(s)$$
$$\boldsymbol{z}^\ell_{\mu,\mathfrak{a}}(t) = \boldsymbol{\xi}^\ell_{\mu\mathfrak{a}}(t) + \gamma_0 \int_0^t ds \sum_{\alpha\mathfrak{b}} \Delta_\alpha(s) G^{\ell+1}_{\mu\alpha,\mathfrak{a},\mathfrak{b}}(t,s) \phi(\boldsymbol{h}^\ell_{\alpha,\mathfrak{b}-\mathfrak{a}}) \tag{86}$$

where $\boldsymbol{\chi}^{\ell+1}_{\mu,\mathfrak{a}}(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0) \phi(\boldsymbol{h}^\ell_{\mu\mathfrak{a}}(t))$. The order parameters for this network architecture are

$$\Phi^\ell_{\mu\alpha,\mathfrak{a}\mathfrak{b}}(t,s) = \frac{1}{N} \phi(\boldsymbol{h}^\ell_{\mu\mathfrak{a}}(t)) \cdot \phi(\boldsymbol{h}^\ell_{\alpha\mathfrak{b}}(s)) \ , \ G^\ell_{\mu\alpha,\mathfrak{a}\mathfrak{b}}(t,s) = \frac{1}{N} \boldsymbol{g}^\ell_{\mu\mathfrak{a}}(t) \cdot \boldsymbol{g}^\ell_{\alpha\mathfrak{b}}(s) \tag{87}$$

These two order parameters per layer collectively define the neural tangent kernel. Following the computation in D, we obtain the following field theory

$$\{u^\ell_{\mu\mathfrak{a}}(t)\} \sim \mathcal{GP}(0, \boldsymbol{\Phi}^{\ell-1}) \ , \ \{r^\ell_{\mu\mathfrak{a}}(t)\} \sim \mathcal{GP}(0, \boldsymbol{G}^{\ell+1})$$
$$h^\ell_{\mu\mathfrak{a}}(t) = u^\ell_{\mu\mathfrak{a}}(t) + \gamma_0 \int_0^\infty ds \sum_{\alpha,\mathfrak{b}} [A^{\ell-1}_{\mu\alpha,\mathfrak{a}\mathfrak{b}}(t,s) + \Theta(t-s) \Phi^{\ell-1}_{\mu\alpha,\mathfrak{a}\mathfrak{b}} \Delta_\alpha(s)] \dot{\phi}(h^\ell_{\alpha\mathfrak{b}}(s)) z^\ell_{\alpha\mathfrak{b}}(s)$$
$$z^\ell_{\mu\mathfrak{a}}(t) = r^\ell_{\mu\mathfrak{a}}(t) + \gamma_0 \int_0^\infty ds \sum_{\alpha,\mathfrak{b}} [B^\ell_{\mu\alpha,\mathfrak{a}\mathfrak{b}}(t,s) + \Theta(t-s) G^{\ell+1}_{\mu\alpha,\mathfrak{a}\mathfrak{b}} \Delta_\alpha(s)] \phi(h^\ell_{\alpha\mathfrak{b}}(s))$$
$$\Phi^\ell_{\mu\alpha,\mathfrak{a}\mathfrak{b}}(t,s) = \left\langle \phi(h^\ell_{\mu\mathfrak{a}}(t)) \cdot \phi(h^\ell_{\alpha\mathfrak{b}}(s)) \right\rangle \ , \ G^\ell_{\mu\alpha,\mathfrak{a}\mathfrak{b}}(t,s) = \left\langle g^\ell_{\mu\mathfrak{a}}(t) \cdot g^\ell_{\alpha\mathfrak{b}}(s) \right\rangle$$
$$A^\ell_{\mu\alpha,\mathfrak{a}\mathfrak{b}}(t,s) = \frac{1}{\gamma_0} \left\langle \frac{\delta \phi(h^\ell_{\mu\mathfrak{a}}(t))}{\delta r^\ell_{\alpha\mathfrak{b}}(s)} \right\rangle \ , \ B^\ell_{\mu\alpha,\mathfrak{a}\mathfrak{b}}(t,s) = \frac{1}{\gamma_0} \left\langle \frac{\delta g^{\ell+1}_{\mu\mathfrak{a}}(t)}{\delta u^{\ell+1}_{\alpha\mathfrak{b}}(s)} \right\rangle \tag{88}$$

We see that this field theory essentially multiples the number of sample indices by the number of spatial indices $P \to P|\mathcal{S}|$. Thus the time complexity of evaluation of this theory scales very poorly as $\mathcal{O}(P^3|\mathcal{S}|^3 T^3)$, rendering DMFT solutions very computationally intensive.

H Trainable Bias Parameter

If we include a bias $\boldsymbol{b}^\ell(t) \in \mathbb{R}^N$ in our trainable model, so that

$$\boldsymbol{h}^{\ell+1}_\mu(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(t) \phi(\boldsymbol{h}^\ell_\mu(t)) + \boldsymbol{b}^\ell(t) \tag{89}$$

then the dynamics on $\boldsymbol{b}^\ell(t)$ induced by gradient flow is

$$\frac{d}{dt} \boldsymbol{b}^\ell(t) = \gamma^2 \sum_\alpha \Delta_\alpha(t) \frac{\partial f_\alpha}{\partial b^\ell} = \frac{\gamma}{\sqrt{N}} \sum_\alpha \Delta_\alpha(t) \boldsymbol{g}^{\ell+1}_\alpha(t) = \gamma_0 \sum_\alpha \Delta_\alpha(t) \boldsymbol{g}^\ell_\alpha(t) \tag{90}$$

Assuming that $b^\ell_i(0) \sim \mathcal{N}(0,1)$, the dynamics of the DMFT becomes

$$\{u^\ell\} \sim \mathcal{GP}(0, \boldsymbol{\Phi}^{\ell-1} + \mathbf{1}\mathbf{1}^\top) \ , \ \{r^\ell\} \sim \mathcal{GP}(0, \boldsymbol{G}^{\ell+1})$$
$$h^\ell_\mu(t) = u^\ell_\mu(t) + \gamma_0 \int_0^\infty ds \sum_\alpha [A^{\ell-1}_{\mu\alpha}(t,s) + \Theta(t-s) \Delta_\alpha(s) \Phi^{\ell-1}_{\mu\alpha}(t,s)] g^\ell_\alpha(s) + \gamma_0 \int_0^t ds \sum_\alpha \Delta_\alpha(s) g^\ell_\alpha(s)$$
$$z^\ell_\mu(t) = r^\ell_\mu(t) + \gamma_0 \int_0^\infty ds \sum_\alpha [B^\ell_{\mu\alpha}(t,s) + \Theta(t-s) \Delta_\alpha(s) G^{\ell+1}_{\mu\alpha}(t,s)] \phi(h^\ell_\alpha(s)) \tag{91}$$

I Multiple Output Channels

We now consider network outputs on $C = \mathcal{O}_N(1)$ classes. The prediction for a data point $\mu \in [P]$ at time $t \in \mathbb{R}_+$ is $\boldsymbol{f}_\mu(t) \in \mathbb{R}^C$. As before, we define the error signal as $\boldsymbol{\Delta}_\mu = -\frac{\partial}{\partial \boldsymbol{f}_\mu} \ell(\boldsymbol{f}_\mu, \boldsymbol{y}_\mu) \in \mathbb{R}^C$. For any pair of data points μ, α the NTK is a $C \times C$ matrix $\boldsymbol{K}^{NTK}_{\mu\alpha} \in \mathbb{R}^{C \times C}$ with entries

$K^{NTK}_{\mu\alpha,cc'} = \frac{\partial f_c(\boldsymbol{x}_\mu)}{\partial \boldsymbol{\theta}} \cdot \frac{\partial f_{c'}(\boldsymbol{x}_\alpha)}{\partial \boldsymbol{\theta}}$. From these matrices, we can compute the evolution of the predictions in the network.

$$\frac{d}{dt}\boldsymbol{f}_\mu = \sum_{\alpha=1}^{P} \boldsymbol{K}^{NTK}_{\mu\alpha} \boldsymbol{\Delta}_\alpha \tag{92}$$

In this case, we have matrices for the backprop features $\boldsymbol{g}^\ell = \gamma\sqrt{N}\frac{\partial \boldsymbol{f}^\top}{\partial \boldsymbol{h}^\ell} \in \mathbb{R}^{N\times C}$. These satisfy the usual recursion

$$\boldsymbol{g}^\ell = \gamma\sqrt{N}\frac{\partial \boldsymbol{f}^\top}{\partial \boldsymbol{h}^\ell} = \gamma\sqrt{N}\left(\frac{\partial \boldsymbol{h}^{\ell+1}}{\partial \boldsymbol{h}^\ell}\right)^\top \frac{\partial \boldsymbol{f}^\top}{\partial \boldsymbol{h}^{\ell+1}} = \left[\dot{\phi}(\boldsymbol{h}^\ell)\mathbf{1}^\top\right] \odot \left[\frac{1}{\sqrt{N}}\boldsymbol{W}^{\ell\top}\boldsymbol{g}^{\ell+1}\right] \tag{93}$$

We can now compute the NTK for samples μ, α

$$\begin{aligned}\boldsymbol{K}^{NTK}_{\mu\alpha} &= \sum_\ell \frac{\partial \boldsymbol{f}(\boldsymbol{x}_\mu)}{\partial \boldsymbol{W}^\ell} \cdot \frac{\partial \boldsymbol{f}(\boldsymbol{x}_\alpha)}{\partial \boldsymbol{W}^\ell} \\ &= \Phi^L_{\mu\alpha}\,\mathbf{I} + \sum_{\ell=1}^{L-1} \boldsymbol{G}^{\ell+1}_{\mu\alpha}\Phi^\ell_{\mu\alpha} + \boldsymbol{G}^1_{\mu\alpha}K^x_{\mu\alpha}\end{aligned} \tag{94}$$

where $\boldsymbol{G}^\ell_{\mu\alpha} = \frac{1}{N}\boldsymbol{g}^{\ell\top}_\mu \boldsymbol{g}^\ell_\alpha \in \mathbb{R}^{C\times C}$ and $\Phi^\ell_{\mu\alpha} = \frac{1}{N}\phi(\boldsymbol{h}^\ell_\mu)\cdot\phi(\boldsymbol{h}^\ell_\alpha) \in \mathbb{R}$. Next we introduce kernels $\boldsymbol{A}^\ell_{\mu\alpha}(t,s) \in \mathbb{R}^C$ and $\boldsymbol{B}^\ell_{\mu\alpha}(t,s) \in \mathbb{R}^C$ which are defined in the usual way. The corresponding field theory has the form

$$\begin{aligned}h^\ell_\mu(t) &= \chi^\ell_\mu(t) + \gamma_0 \int_0^\infty ds \sum_{\alpha=1}^{P} \left[\boldsymbol{A}^{\ell-1}_{\mu\alpha}(t,s) + \Theta(t-s)\boldsymbol{\Delta}_\alpha(s)\Phi^{\ell-1}_{\mu\alpha}(t,s)\right]\cdot \boldsymbol{g}^\ell_\alpha(s) \in \mathbb{R} \\ \boldsymbol{z}^\ell_\mu(t) &= \boldsymbol{\xi}^\ell_\mu(t) + \gamma_0 \int_0^\infty ds \sum_{\alpha=1}^{P} \left[\boldsymbol{B}^\ell_{\mu\alpha}(t,s) + \Theta(t-s)\boldsymbol{G}^{\ell+1}_{\mu\alpha}\boldsymbol{\Delta}_\alpha(s)\right]\phi(h^\ell_\mu(t)) \in \mathbb{R}^C \\ \boldsymbol{g}^\ell_\mu(t) &= \dot{\phi}(h^\ell_\mu(t))\boldsymbol{z}^\ell_\mu(t) \in \mathbb{R}^C\end{aligned} \tag{95}$$

From these fields, the saddle point equations define the kernels as

$$\begin{aligned}&\Phi^\ell_{\mu\alpha}(t,s) = \left\langle \phi(h^\ell_\mu(t))\phi(h^\ell_\alpha(s))\right\rangle \in \mathbb{R}\ ,\ \boldsymbol{G}^\ell_{\mu\alpha}(t,s) = \left\langle \boldsymbol{g}^\ell_\mu(t)\boldsymbol{g}^\ell_\alpha(s)^\top\right\rangle \in \mathbb{R}^{C\times C} \\ &\boldsymbol{A}^\ell_{\mu\alpha}(t,s) = \frac{1}{\gamma_0}\left\langle \frac{\delta\phi(h^\ell_\mu(t))}{\delta \boldsymbol{r}^\ell_\alpha(s)}\right\rangle \in \mathbb{R}^C\ ,\ \boldsymbol{B}^\ell_{\mu\alpha}(t,s) = \frac{1}{\gamma_0}\left\langle \frac{\delta \boldsymbol{g}^\ell_\mu(t)}{\delta u^\ell_\alpha(s)}\right\rangle \in \mathbb{R}^C.\end{aligned} \tag{96}$$

This allows studying the multi-class structure of learned representations.

J Weight Decay in Deep Homogenous Networks

If we train with weight decay, $\frac{d}{dt}\boldsymbol{\theta} = -\gamma^2\nabla_{\boldsymbol{\theta}}\mathcal{L} - \lambda\boldsymbol{\theta}$, in a κ-degree homogenous network ($f(c\boldsymbol{\theta}) = c^\kappa f(\boldsymbol{\theta})$), then the prediction dynamics satisfy

$$\frac{d}{dt}f(\boldsymbol{x},t) = \sum_\alpha \Delta_\alpha(t)K^{NTK}_{\mu\alpha}(\boldsymbol{x},\boldsymbol{x}_\alpha,t) - \lambda\kappa f(\boldsymbol{x},t)\ ,$$

This holds by the following identity $\frac{\partial}{\partial c}f(c\boldsymbol{\theta}) = \frac{\partial}{\partial c}c^\kappa f(\boldsymbol{\theta})$, which when evaluated at $c = 1$ gives $\frac{\partial}{\partial\boldsymbol{\theta}}f(\boldsymbol{\theta})\cdot\boldsymbol{\theta} = \kappa f(\boldsymbol{\theta})$. This identity was utilized in a prior work which studied L2 regularization in the lazy regime [74]. For a L-hidden layer ReLU network $\phi(h) = \max(0,h)$, the degree is $\kappa = L+1$, while rectified power law nonlinearities $\phi(h) = \max(0,h)^q$ give degrees $\kappa = \frac{q^{L+1}-1}{q-1}$. We note that the fixed point of the function dynamics above gives a representer theorem with the final NTK

$$f(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x})^\top\left[\boldsymbol{K} + \lambda\kappa\boldsymbol{I}\right]^{-1}\boldsymbol{y} \tag{97}$$

where $[\boldsymbol{k}(x)]_\mu = \lim_{t\to\infty}K(\boldsymbol{x},\boldsymbol{x}_\mu,t)$ and $K_{\mu\alpha} = \lim_{t\to\infty}K(\boldsymbol{x}_\mu,\boldsymbol{x}_\alpha,t)$. The prior work of Lewkowycz et al [74] considered NTK parameterization $\gamma_0 = 0$. In this limit, the kernel (and consequently output function) decay to zero at large time, but if $\gamma_0 > 0$, then the network converges

to a nontrivial fixed point as $t \to \infty$. In the DMFT limit we can determine the final kernel by solving the following field dynamics

$$h_\mu^\ell(t) = e^{-\lambda t}\chi_\mu^\ell(t) + \gamma_0 \int_0^t ds\, e^{-\lambda(t-s)} \sum_{\alpha=1}^P \Delta_\alpha(s) g_\alpha^\ell(s) \Phi_{\mu\alpha}^{\ell-1}(t,s)$$
$$z_\mu^\ell(t) = e^{-\lambda t}\xi_\mu^\ell(t) + \gamma_0 \int_0^t ds\, e^{-\lambda(t-s)} \sum_{\alpha=1}^P \Delta_\alpha(s) \phi(h_\alpha^\ell(s)) G_{\mu\alpha}^{\ell+1}(t,s). \tag{98}$$

We see that the contribution from initial conditions is exponentially suppressed at large time t while the second term contributes most when the system has equilibrated. We provide an example of the weight decay DMFT showing its validity in a two layer ReLU network in Figure 3.

K Bayesian/Langevin Trained Mean Field Networks

Rather than studying exact gradient flow, many works have considered Langevin dynamics (gradient flow with white noise process on the weights) of neural network training [25, 32, 91, 30, 31]. This setting is of special theoretical interest since the distribution of parameters converges at long times to a Gibbs equilibrium distribution which has a Bayesian interpretation [4, 5, 91]. The relevant Langevin equation for our mean field gradient flow is

$$d\boldsymbol{\theta}(t) = -\gamma^2 \nabla L(\boldsymbol{\theta}(t))dt - \lambda\beta^{-1}\boldsymbol{\theta}(t)dt + \sqrt{2\beta^{-1}}d\boldsymbol{\epsilon}(t), \tag{99}$$

where λ is a ridge penalty which controls the scale of parameters, and $d\boldsymbol{\epsilon}(t)$ is a Brownian motion term which has covariance structure $\left\langle d\boldsymbol{\epsilon}(t) d\boldsymbol{\epsilon}(t')^\top \right\rangle = \delta(t-t')\mathbf{I}$. The parameter β, known as the inverse temperature controls the scale of the random Gaussian noise injected into this stochastic process. The dynamical treatment of the $\beta \to \infty$ limit will coincide with our usual DMFT while the $\beta \ll \infty$ will exhibit a nontrivial balance between the usual DMFT feature updates and the random Langevin noise. At late times, such a system will equilibrate to its Gibbs distribution.

K.1 Dynamical Analysis

In this section we analyze the dynamical mean field theory for these Langevin dynamics. First we note that the effect of regularization can be handled with a simple integrating factor

$$d[\boldsymbol{W}^\ell(t) e^{\frac{\lambda t}{\beta}}] = e^{\frac{\lambda}{\beta}t} \left[\frac{\gamma_0}{\sqrt{N}} \sum_\mu \Delta_\mu(t) \boldsymbol{g}_\mu^{\ell+1}(t) \phi(\boldsymbol{h}_\mu^\ell(t))^\top \right] dt + \sqrt{2\beta^{-1}} e^{\frac{\lambda t}{\beta}} d\boldsymbol{\epsilon}^\ell(t). \tag{100}$$

where $d\boldsymbol{\epsilon}(t) \in \mathbb{R}^{N\times N}$ is the Gaussian noise for layer ℓ at time t. It is straightforward to verify by Ito's lemma that, under mean field parameterization, the fluctuations in $f's$ dynamics due to Brownian motion are $\frac{\partial f}{\partial \boldsymbol{\theta}} \cdot d\boldsymbol{\epsilon}(t) \sim \mathcal{O}(N^{-1/2})$ and are thus negligible in the $N \to \infty$ limit. Thus the evolution of the network function takes the form

$$\frac{\partial f_\mu(t)}{\partial t} = \sum_\alpha \Delta_\alpha(t) K_{\mu\alpha}(t,t) - \lambda\beta^{-1}\boldsymbol{\theta}(t) \cdot \nabla_{\boldsymbol{\theta}} f_\mu(t) + \frac{1}{\beta}\mathrm{Tr}\nabla_{\boldsymbol{\theta}}^2 f_\mu(t)$$

We can express both of these parameter contractions in feature space provided we introduce the new features $r_{i,\mu}^\ell(t) = \frac{\partial g_{i,\mu}^\ell}{\partial h_{i,\mu}^\ell}$ which are necessary to compute Hessian terms like $\frac{\partial^2 f}{\partial W_{ij}^\ell \partial W_{ij}^\ell} = N^{-3/2}\frac{\partial}{\partial W_{ij}^\ell}[g_i^{\ell+1}\phi(h_j^\ell)] = N^{-2}\, r_i\, \phi(h_j^\ell)^2$ in each layer. This gives the following evolution

$$\frac{\partial f_\mu(t)}{\partial t} = \sum_\alpha \Delta_\alpha(t) K_{\mu\alpha}(t,t) - \lambda\beta^{-1}\sum_\ell \left\langle z_\mu^\ell(t)\phi(h_\mu^\ell(t)) \right\rangle + \beta^{-1}\sum_\ell \left\langle r_\mu^{\ell+1}(t) \right\rangle \left\langle \phi(h_\mu^\ell(t))^2 \right\rangle \tag{101}$$

As before, we compute the next layer field $\boldsymbol{h}^{\ell+1}$ in terms of $\boldsymbol{\chi}^{\ell+1}$ and $\boldsymbol{z}^\ell$ in terms of $\boldsymbol{\xi}^\ell$

$$\boldsymbol{h}_\mu^{\ell+1}(t) = e^{-\frac{\lambda}{\beta}t}\boldsymbol{\chi}_\mu^{\ell+1}(t) + \int_0^t e^{-\frac{\lambda}{\beta}(t-s)} \left[ds \frac{\gamma_0}{N} \sum_\alpha \Delta_\alpha(s) \boldsymbol{g}_\alpha^{\ell+1}(s) \phi(\boldsymbol{h}_\alpha^\ell(s))^\top + \sqrt{\frac{2}{\beta N}} d\boldsymbol{\epsilon}^\ell(s) \right] \phi(\boldsymbol{h}_\mu^\ell(t))$$
$$\boldsymbol{z}_\mu^{\ell+1}(t) = e^{-\frac{\lambda}{\beta}t}\boldsymbol{\xi}_\mu^{\ell+1}(t) + \int_0^t e^{-\frac{\lambda}{\beta}(t-s)} \left[ds \frac{\gamma_0}{N} \sum_\alpha \Delta_\alpha(s) \boldsymbol{g}_\alpha^{\ell+1}(s) \phi(\boldsymbol{h}_\alpha^\ell(s))^\top + \sqrt{\frac{2}{\beta N}} d\boldsymbol{\epsilon}^\ell(s) \right]^\top \boldsymbol{g}_\mu^{\ell+1}(t)$$

The dependence on the initial condition through $\boldsymbol{\chi}, \boldsymbol{\xi}$ is suppressed at long times due the regularization factor $e^{-\frac{\lambda}{\beta}t}$, while the Brownian motion and gradient updates will survive in the $t \to \infty$ limit. In addition to the usual $\{\boldsymbol{\chi}^\ell, \boldsymbol{\xi}^\ell\}$ fields which arise from the initial condition, we see that $\boldsymbol{h}^\ell(t), \boldsymbol{z}^\ell(t)$ also depend on the following fields which arise from the integrated Brownian motion

$$\begin{aligned}
\boldsymbol{\chi}_\mu^{\epsilon,\ell}(t) &= \sqrt{\frac{2}{\beta N}} \int_0^\infty ds\, e^{-\frac{\lambda}{\beta}(t-s)} \Theta(t-s) d\boldsymbol{\epsilon}^\ell(s) \phi(\boldsymbol{h}_\mu^\ell(t)) \\
\boldsymbol{\xi}_\mu^{\epsilon,\ell}(t) &= \sqrt{\frac{2}{\beta N}} \int_0^\infty ds\, e^{-\frac{\lambda}{\beta}(t-s)} \Theta(t-s) d\boldsymbol{\epsilon}^\ell(s)^\top \boldsymbol{g}_\mu^{\ell+1}(t)
\end{aligned} \tag{102}$$

Our aim is now to compute the moment generating function for the $\{\boldsymbol{\chi}, \boldsymbol{\xi}, \boldsymbol{\chi}^\epsilon, \boldsymbol{\xi}^\epsilon\}$ fields which causally determine $\{\boldsymbol{h}, \boldsymbol{z}\}$. This MGF has the form

$$Z = \left\langle \exp\left(\sum_{\ell\mu} \int_0^\infty \left[\boldsymbol{j}_\mu^\ell(t) \cdot \boldsymbol{\chi}_\mu^\ell(t) + \boldsymbol{v}_\mu^\ell(t) \cdot \boldsymbol{\xi}_\mu^\ell(t) + \boldsymbol{j}_\mu^{\epsilon,\ell}(t) \cdot \boldsymbol{\chi}_\mu^{\epsilon,\ell}(t) + \boldsymbol{v}_\mu^{\epsilon,\ell}(t) \cdot \boldsymbol{\xi}_\mu^{\epsilon\ell}(t) \right] \right) \right\rangle_{\boldsymbol{\theta}_0, \boldsymbol{\epsilon}(t)} \tag{103}$$

We insert Dirac-delta functions in the usual way to enforce the definitions of $\boldsymbol{\chi}, \boldsymbol{\xi}, \boldsymbol{\chi}^\epsilon, \boldsymbol{\xi}^\epsilon$ and then average over $\boldsymbol{\theta}_0, \boldsymbol{\epsilon}(t)$. These averages can be performed separately with the $\boldsymbol{\theta}_0$ average giving the identical terms as derived in previous sections. We focus on the average over Brownian disorder

$$\begin{aligned}
&\ln \left\langle \exp\left(i\sqrt{2}\beta^{-1/2} N^{-1/2} \sum_\mu \int_0^\infty dt\, \mathrm{Tr}\left[\hat{\boldsymbol{\chi}}_\mu^{\epsilon,\ell+1}(t) \phi(\boldsymbol{h}_\mu^\ell(t))^\top + \boldsymbol{g}_\mu^{\ell+1}(t) \hat{\boldsymbol{\xi}}_\mu^{\epsilon,\ell}(t)^\top \right] \int e^{-\frac{\lambda}{\beta}(t-s)} \Theta(t-s) d\boldsymbol{\epsilon}(s) \right) \right\rangle_{\boldsymbol{\epsilon}(t)} \\
&= -\frac{1}{\beta N} \int_0^\infty ds \left| \int dt\, \Theta(t-s) e^{-\frac{\lambda}{\beta}(t-s)} \sum_\mu \left[\hat{\boldsymbol{\chi}}_\mu^{\epsilon,\ell+1}(t) \phi(\boldsymbol{h}_\mu^\ell(t))^\top + \boldsymbol{g}_\mu^{\ell+1}(t) \hat{\boldsymbol{\xi}}_\mu^{\epsilon,\ell}(t)^\top \right] \right|^2 \\
&= -\frac{1}{\beta} \int_0^\infty ds \int_0^\infty dt \int_0^\infty dt' \Theta(t-s)\Theta(t'-s) e^{-\frac{\lambda}{\beta}(t-s+t'-s)} \\
&\times \sum_{\mu\alpha} \left[\hat{\boldsymbol{\chi}}_\mu^{\epsilon,\ell+1}(t) \cdot \hat{\boldsymbol{\chi}}_\alpha^{\epsilon,\ell+1}(t') \Phi_{\mu\alpha}^\ell(t,t') + \hat{\boldsymbol{\xi}}_\mu^{\epsilon,\ell}(t) \cdot \hat{\boldsymbol{\xi}}_\alpha^{\epsilon,\ell}(t') G_{\mu\alpha}^{\ell+1}(t,t') + 2i \hat{\boldsymbol{\chi}}_\mu^{\epsilon,\ell+1}(t) \cdot \boldsymbol{g}_\alpha^{\ell+1}(t') A_{\mu\alpha}^{\epsilon,\ell}(t,t') \right] \\
&= -\frac{1}{2\lambda} \int_0^\infty dt \int_0^\infty dt' \exp\left(-\frac{\lambda}{\beta}(t+t') \right) \left[e^{2\frac{\lambda}{\beta}\min\{t,t'\}} - 1 \right] \\
&\times \sum_{\mu\alpha} \left[\hat{\boldsymbol{\chi}}_\mu^{\epsilon,\ell+1}(t) \cdot \hat{\boldsymbol{\chi}}_\alpha^{\epsilon,\ell+1}(t') \Phi_{\mu\alpha}^\ell(t,t') + \hat{\boldsymbol{\xi}}_\mu^{\epsilon,\ell}(t) \cdot \hat{\boldsymbol{\xi}}_\alpha^{\epsilon,\ell}(t') G_{\mu\alpha}^{\ell+1}(t,t') + 2i \hat{\boldsymbol{\chi}}_\mu^{\epsilon,\ell+1}(t) \cdot \boldsymbol{g}_\alpha^{\ell+1}(t') A_{\mu\alpha}^{\epsilon,\ell}(t,t') \right]
\end{aligned} \tag{104}$$

where we introduced the order parameter $iA_{\mu\alpha}^{\epsilon,\ell}(t,t') = \frac{1}{N}\phi(\boldsymbol{h}_\mu^\ell(t)) \cdot \hat{\boldsymbol{\xi}}_\alpha^{\epsilon,\ell}(s)$. We will use the shorthand for the temporal prefactor in the above $C_{\lambda,\beta}(t,t') = \frac{1}{\lambda}\exp\left(-\frac{\lambda}{\beta}(t+t')\right)\left[e^{2\frac{\lambda}{\beta}\min\{t,t'\}} - 1\right] \sim_{t,t'\to\infty} \frac{1}{\lambda}\exp\left(-\frac{\lambda}{\beta}|t-t'|\right)$. We insert a Lagrange multiplier $B^{\epsilon,\ell}$ to enforce the definition of $A^{\epsilon,\ell}$. After

$$\begin{aligned}
Z \propto &\int d\Phi_{\mu\alpha}^\ell(t,s) d\hat{\Phi}_{\mu\alpha}^\ell(t,s) dG_{\mu\alpha}^\ell(t,s) d\hat{G}_{\mu\alpha}^\ell(t,s) dA_{\mu\alpha}^\ell(t,s) dB_{\mu\alpha}^\ell(t,s) dA_{\mu\alpha}^{\epsilon\ell}(t,s) dB_{\mu\alpha}^{\epsilon\ell}(t,s) \\
&\times \exp\left(NS[\Phi, \hat{\Phi}, G, \hat{G}, A, B, A^\epsilon, B^\epsilon] \right)
\end{aligned} \tag{105}$$

The order parameters can be determined by the saddle point equations. These equations for $\Phi, \hat{\Phi}, G, \hat{G}, A, B$ are the same as before. The new equations are

$$\begin{aligned}
\frac{\delta S}{\delta A_{\mu\alpha}^{\epsilon,\ell}(t,s)} &= -B_{\mu\alpha}^{\epsilon,\ell}(t,s) - iC_{\lambda,\beta}(t,s) \left\langle \hat{\chi}_\mu^{\epsilon,\ell+1}(t) g_\alpha^{\ell+1}(s) \right\rangle = 0 \\
\frac{\delta S}{\delta B_{\mu\alpha}^{\epsilon,\ell}(t,s)} &= -A_{\mu\alpha}^{\epsilon,\ell}(t,s) - iC_{\lambda,\beta}(t,s) \left\langle \phi(h_\mu^\ell(t)) \hat{\xi}_\alpha^{\epsilon,\ell}(s) \right\rangle = 0
\end{aligned} \tag{106}$$

Using the fact that $\boldsymbol{\Phi}^\ell, G^\ell$ concentrate, we can use the Hubbard trick to linearize the quadratic terms in $\hat{\chi}^\epsilon$ and $\hat{\xi}^\epsilon$.

$$\exp\left(-\frac{1}{2}\int_0^\infty dt\int_0^\infty ds\, C_{\lambda,\beta}(t,s)\sum_{\mu\alpha}\hat{\chi}_\mu^{\epsilon,\ell+1}(t)\hat{\chi}_\alpha^{\epsilon,\ell+1}(s)\Phi_{\mu\alpha}^\ell(t,s)\right)$$
$$=\left\langle \exp\left(-i\sum_\mu\int_0^\infty dt\, u_\mu^{\epsilon,\ell+1}(t)\hat{\chi}_\mu^{\epsilon,\ell+1}(t)\right)\right\rangle_{u_\mu^{\epsilon,\ell+1}(t)\sim\mathcal{GP}(0,C\odot\Phi^\ell)} \tag{107}$$
$$\exp\left(-\frac{1}{2}\int_0^\infty dt\int_0^\infty ds\, C_{\lambda,\beta}(t,s)\sum_{\mu\alpha}\hat{\xi}_\mu^{\epsilon,\ell}(t)\hat{\xi}_\alpha^{\epsilon,\ell}(s)G_{\mu\alpha}^{\ell+1}(t,s)\right)$$
$$=\left\langle \exp\left(-i\sum_\mu\int_0^\infty dt\, r_\mu^{\epsilon,\ell}(t)\hat{\xi}_\mu^{\epsilon,\ell}(t)\right)\right\rangle_{r_\mu^{\epsilon,\ell}(t)\sim\mathcal{GP}(0,C\odot G^{\ell+1})} \tag{108}$$

Using the vectorization notation, we find the interpretation that $\boldsymbol{\chi}^{\epsilon,\ell}$ and $\boldsymbol{\xi}^{\epsilon,\ell}$ decouple as

$$\boldsymbol{\chi}^{\epsilon,\ell+1}=\boldsymbol{u}^{\epsilon,\ell+1}+\boldsymbol{A}^{\epsilon,\ell+1}\boldsymbol{g}^{\ell+1}\ ,\ \boldsymbol{\xi}^{\epsilon,\ell}=\boldsymbol{r}^{\epsilon,\ell}+\boldsymbol{B}^{\epsilon,\ell\top}\phi(\boldsymbol{h}^\ell) \tag{109}$$
$$\boldsymbol{A}^{\epsilon,\ell+1}=\boldsymbol{C}_{\lambda,\beta}\odot\left\langle\frac{\partial\phi(\boldsymbol{h}^\ell)}{\partial\boldsymbol{r}^{\epsilon,\ell}}\right\rangle\ ,\ \boldsymbol{B}^{\epsilon\ell}=\boldsymbol{C}_{\lambda,\beta}\odot\left\langle\frac{\partial\boldsymbol{g}^{\ell+1}}{\partial\boldsymbol{u}^{\ell+1}}\right\rangle^\top \tag{110}$$

As before, we make the substitutions $\boldsymbol{B}\to\gamma_0^{-1}\boldsymbol{B}^\top$ and $\boldsymbol{A}\to\gamma_0^{-1}\boldsymbol{A}$ and arrive at the final DMFT equations

$$\begin{aligned}
&\{u_\mu^\ell(t)\}\sim\mathcal{GP}(0,\Phi^{\ell-1})\ ,\ \{r_\mu^\ell(t)\}\sim\mathcal{GP}(0,G^{\ell+1})\\
&\{u_\mu^{\epsilon,\ell}(t)\}\sim\mathcal{GP}(0,C_{\lambda,\beta}\odot\Phi^{\ell-1})\ ,\ \{r_\mu^{\epsilon,\ell}(t)\}\sim\mathcal{GP}(0,C_{\lambda,\beta}\odot G^{\ell+1})\\
h_\mu^\ell(t)&=e^{-\frac{\lambda}{\beta}t}\left[u_\mu^\ell(t)+\gamma_0\int_0^t ds\sum_\alpha A_{\mu\alpha}^{\ell-1}(t,s)g_\alpha^\ell(s)\right]\\
&\quad+u_\mu^{\epsilon,\ell}(t)+\gamma_0\int_0^t ds\ \sum_\alpha\left[A_{\mu\alpha}^{\epsilon,\ell-1}(t,s)+e^{-\frac{\lambda}{\beta}(t-s)}\Delta_\alpha(s)\Phi_{\mu\alpha}^{\ell-1}(t,s)\right]g_\alpha^\ell(s)\\
z_\mu^\ell(t)&=e^{-\frac{\lambda}{\beta}t}\left[r_\mu^\ell(t)+\gamma_0\int_0^t ds\sum_\alpha B_{\mu\alpha}^\ell(t,s)g_\alpha^\ell(s)\right]\\
&\quad+r_\mu^{\epsilon,\ell}(t)+\gamma_0\int_0^t ds\ \sum_\alpha\left[B_{\mu\alpha}^{\epsilon,\ell}(t,s)+e^{-\frac{\lambda}{\beta}(t-s)}\Delta_\alpha(s)G_{\mu\alpha}^{\ell+1}(t,s)\right]\phi(h_\alpha^\ell(s))
\end{aligned} \tag{111}$$

where the kernels are defined in the usual way. As expected, the contributions from the initial conditions χ^ℓ,ξ^ℓ are exponentially suppressed at late time whereas the contributions from the Brownian disorder $\chi^{\epsilon,\ell},\xi^{\epsilon,\ell}$ persist at late time.

K.2 Weak Feature Learning, Long Time Limit

In the weak feature learning $\gamma_0\to 0$ and long time $t\to\infty$ limit, the preactivation fields equilibrate to Gaussian processes $h_\mu^\ell(t)\sim u_\mu^{\epsilon,\ell}(t), z_\mu^\ell(t)\sim r_\alpha^{\epsilon,\ell}(t)$, which have respective covariances $H_{\mu\alpha}^\ell(t,s)=\left\langle h_\mu^\ell(t)h_\alpha^\ell(s)\right\rangle=C_{\lambda,\beta}(t,s)\Phi_{\mu\alpha}^{\ell-1}(t,s), Z_{\mu\alpha}^\ell(t,s)=\left\langle z_\mu^\ell(t)z_\alpha^\ell(s)\right\rangle=C_{\lambda,\beta}(t,s)G_{\mu\alpha}^{\ell+1}(t,s)$. In this long time limit, the feature kernels will be time translation invariant eg $\Phi_{\mu\alpha}^\ell(t,s)=\Phi_{\mu\alpha}^\ell(|t-s|)$. Letting $\tau=|t-s|$ and $C_{\lambda,\beta}(\tau)=\frac{1}{\lambda}\exp\left(-\frac{\lambda}{\beta}\tau\right)$, we have the following recurrence for H^ℓ,Φ^ℓ

$$H_{\mu\alpha}^1(\tau)=C_{\lambda,\beta}(\tau)K_{\mu\alpha}^x\ ,\ \Phi_{\mu\alpha}^1(\tau)=\langle\phi(h)\phi(h')\rangle_{h,h'\sim\mathcal{N}(0,\boldsymbol{H}^1)}\ ,\ \boldsymbol{H}^1=\begin{bmatrix}H_{\mu\mu}^1(0) & H_{\mu\alpha}^1(\tau)\\ H_{\mu\alpha}^1(\tau) & H_{\alpha\alpha}^1(0)\end{bmatrix}$$
$$H_{\mu\alpha}^{\ell+1}(\tau)=C_{\lambda,\beta}(\tau)\Phi_{\mu\alpha}^\ell(\tau)\ ,\ \Phi_{\mu\alpha}^{\ell+1}(t,s)=\langle\phi(h)\phi(h')\rangle_{h,h'\sim\mathcal{N}(0,\boldsymbol{H}^{\ell+1})}\ ,\ \boldsymbol{H}^{\ell+1}=\begin{bmatrix}H_{\mu\mu}^{\ell+1}(0) & H_{\mu\alpha}^{\ell+1}(\tau)\\ H_{\mu\alpha}^{\ell+1}(\tau) & H_{\alpha\alpha}^{\ell+1}(0)\end{bmatrix} \tag{112}$$

Similarly, we can obtain Z^ℓ and G^ℓ in a backward pass recursion

$$
\begin{aligned}
&Z^L_{\mu\alpha}(\tau) = C_{\lambda,\beta}(\tau)\ ,\ G^L_{\mu\alpha}(\tau) = \dot\Phi^L_{\mu\alpha}(\tau) Z^L_{\mu\alpha}(\tau)\ ,\ \dot\Phi^L_{\mu\alpha}(\tau) = \left\langle \dot\phi(h)\dot\phi(h')\right\rangle_{h,h'\sim\mathcal{N}(0,\boldsymbol{H}^L)} \\
&Z^\ell_{\mu\alpha}(\tau) = C_{\lambda,\beta}(\tau) G^{\ell+1}_{\mu\alpha}(\tau)\ ,\ G^\ell_{\mu\alpha}(\tau) = \dot\Phi^\ell_{\mu\alpha}(\tau) Z^\ell_{\mu\alpha}(\tau)\ ,\ \dot\Phi^\ell_{\mu\alpha}(\tau) = \left\langle \dot\phi(h)\dot\phi(h')\right\rangle_{h,h'\sim\mathcal{N}(0,\boldsymbol{H}^\ell)}
\end{aligned}
\tag{113}
$$

On the temporal diagonal $\tau = 0$, these equations give the usual recursions used to compute the NNGP kernels at initialization [5], though with initialization variance $C_{\lambda,\beta}(0) = \lambda^{-1}$, set by the weight decay term in the Langevin dynamics. This indicates that the long time Langevin dynamics at $\gamma_0 \to 0$ simply rescales the Gaussian weight variance based on λ. It would be interesting to explore fluctuation dissipation relationships at finite γ_0 within this framework which we leave to future work.

K.3 Equilibrium Analysis

The Langevin dynamics at finite N converges (possibly in a time extensive in N) to an equilibrium distribution with several interesting properties, as was recently studied by Aitchison et al [91] and implicitly by Seroussi et al [31] in a large sample size limit. This setting differs from the previous section where first $N \to \infty$ limit is taken, followed by a $t \to \infty$ limit in the DMFT. This, section, on the other hand studies for any N, the $t \to \infty$ limiting equilibrium distribution. This equilibrated distribution is then analyzed in the $N \to \infty$ limit. The relationship between these two orders of limits remains an open problem. The equilibrium distribution over parameters $p(\boldsymbol{\theta}|\mathcal{D}) \propto \exp\left(-\beta\gamma^2 L(\boldsymbol{\theta}) - \frac{\lambda}{2}|\boldsymbol{\theta}|^2\right)$ can be viewed as a Bayes posterior with log-likelihood $-\beta\gamma^2 L(\boldsymbol{\theta})$ and a Gaussian prior with scale $\lambda^{-1/2}$. In the mean field limit with $\gamma = \sqrt{N}\gamma_0$, we can express the density over pre-activations $\boldsymbol{h}^\ell$ and the output predictions f. This gives

$$
\begin{aligned}
p(\boldsymbol{f}|\mathcal{D}) \propto \exp&\left(-N\gamma_0^2\beta \sum_\mu \ell(f_\mu, y_\mu)\right) \\
&\times \int d\boldsymbol{h}^\ell_\mu \left\langle \prod_\mu \delta\left(f_\mu - \frac{1}{N\gamma_0}\boldsymbol{w}^L\cdot\phi(\boldsymbol{h}^L_\mu)\right)\prod_{\mu\ell}\delta\left(\boldsymbol{h}^{\ell+1}_\mu - \frac{1}{\sqrt{N}}\boldsymbol{W}^\ell\phi(\boldsymbol{h}^\ell_\mu)\right)\right\rangle_{\boldsymbol{\theta}\sim\mathcal{N}(0,\lambda^{-1}\mathbf{I})} \\
&\propto \int \prod_\mu d\hat f_\mu \prod_{\ell\mu\alpha} d\Phi^\ell_{\mu\alpha} d\hat\Phi^\ell_{\mu\alpha} \exp\left(-N\gamma_0^2\beta\sum_\mu \ell(f_\mu,y_\mu) - N\gamma_0\sum_\mu \hat f_\mu f_\mu + \frac{N}{2\lambda}\sum_{\mu\alpha}\hat f_\mu \Phi^L_{\mu\alpha}\hat f_\alpha\right) \\
&\exp\left(\frac{N}{2}\sum_{\ell\mu\alpha}\Phi^\ell_{\mu\alpha}\hat\Phi^\ell_{\mu\alpha} + N\sum_\ell \ln \mathcal{Z}[\Phi^{\ell-1},\hat\Phi^\ell]\right) \\
&\mathcal{Z}[\Phi^{\ell-1},\hat\Phi^\ell] = \int\prod_\mu \frac{dh_\mu d\hat h_\mu}{2\pi}\exp\left(-\frac{1}{2\lambda}\sum_{\mu\alpha}\hat h_\mu \Phi^{\ell-1}_{\mu\alpha}\hat h_\alpha - \frac{1}{2}\sum_{\mu\alpha}\phi(h_\mu)\hat\Phi^\ell_{\mu\alpha}\phi(h_\alpha) + i\sum_\mu \hat h_\mu h_\mu\right)
\end{aligned}
\tag{114}
$$

We see that $p(\boldsymbol{f}|\mathcal{D}) \propto \int d\Phi d\hat\Phi \exp\left(NS[\Phi,\hat\Phi]\right)$ where

$$
S = -\gamma_0^2\beta\sum_\mu \ell(f_\mu,y_\mu) - \gamma_0\sum_\mu \hat f_\mu f_\mu + \frac{1}{2\lambda}\sum_{\mu\alpha}\hat f_\mu\Phi^L_{\mu\alpha}\hat f_\alpha + \frac{1}{2}\sum_{\ell\mu\alpha}\Phi^\ell_{\mu\alpha}\hat\Phi^\ell_{\mu\alpha} + \sum_\ell \ln\mathcal{Z}[\Phi^{\ell-1},\hat\Phi^\ell]
\tag{115}
$$

Thus the predictions f_μ become non-random in this $N \to \infty$ limit and can be determined from the saddle point equations as in [91]. Again, letting $\Delta_\mu = -\frac{\partial}{\partial f_\mu}\ell(f_\mu, y_\mu)$, we find

$$\begin{aligned}
&\frac{\partial S}{\partial f_\mu} = \gamma_0 \hat{f}_\mu - \gamma_0^2 \beta \Delta_\mu = 0 \ , \ \frac{\partial S}{\partial \hat{f}_\mu} = -\gamma_0 f_\mu + \frac{1}{\lambda}\sum_\alpha \Phi^L_{\mu\alpha}\hat{f}_\alpha = 0 \\
&\frac{\partial S}{\partial \Phi^L_{\mu\alpha}} = \frac{1}{2\lambda}\hat{f}_\mu \hat{f}_\alpha + \frac{1}{2}\hat{\Phi}^L_{\mu\alpha} = 0 \ , \ \frac{\partial S}{\partial \hat{\Phi}^L_{\mu\alpha}} = \frac{1}{2}\Phi^\ell_{\mu\alpha} - \frac{1}{2}\left\langle \phi(h^L_\mu)\phi(h^L_\alpha)\right\rangle = 0 \\
&\frac{\partial S}{\partial \Phi^\ell_{\mu\alpha}} = -\frac{1}{2}\left\langle \hat{h}^{\ell+1}_\mu \hat{h}^{\ell+1}_\alpha \right\rangle + \frac{1}{2}\hat{\Phi}^\ell_{\mu\alpha} = 0 \ , \ \frac{\partial S}{\partial \hat{\Phi}^\ell_{\mu\alpha}} = \frac{1}{2}\Phi^\ell_{\mu\alpha} - \frac{1}{2}\left\langle \phi(h^\ell_\mu)\phi(h^\ell_\alpha)\right\rangle = 0
\end{aligned} \tag{116}$$

which implies that f_μ at the fixed point satisfies the following equations

$$f_\mu = \frac{\beta}{\lambda}\sum_\alpha \Phi^L_{\mu\alpha}\Delta_\alpha \ , \ \Delta_\alpha = -\frac{\partial \ell(f_\alpha, y_\alpha)}{\partial f_\alpha}. \tag{117}$$

The last layer's dual kernel has the form $\hat{\Phi}^L_{\mu\alpha} = -\frac{\gamma_0^2\beta^2}{2\lambda}\Delta_\mu\Delta_\alpha$, which we see vanishes as feature learning strength is taken to zero $\gamma_0 \to 0$, while for non-negligible γ_0, we see that the last layer features are non-Gaussian. We thus see that the moment generating function for the last layer field has the form

$$\mathcal{Z}[\Phi^{L-1}, \hat{\Phi}^L] = \int \prod_\mu \frac{dh_\mu d\hat{h}_\mu}{2\pi} \exp\left(-\frac{1}{2\lambda}\sum_{\mu\alpha}\hat{h}_\mu \hat{h}_\alpha \Phi^{L-1}_{\mu\alpha} + \frac{\gamma_0^2\beta^2}{2\lambda}\left[\sum_\mu \Delta_\mu \phi(h_\mu)\right]^2 + i\sum_\mu \hat{h}_\mu h_\mu \right) \tag{118}$$

In the $\gamma_0 \to 0$ limit, the non-Gaussian component of this density vanishes. Now that we have this form, we can compute Φ^L conditional on Φ^{L-1}. Next, we calculate $\hat{\Phi}^{L-1}_{\mu\alpha} = \left\langle \hat{h}^L_\mu \hat{h}^L_\alpha \right\rangle$, giving

$$\hat{\boldsymbol{\Phi}}^{L-1} = \lambda[\boldsymbol{\Phi}^{L-1}]^{-1} - \lambda^2[\boldsymbol{\Phi}^{L-1}]^{-1}\left\langle \boldsymbol{h}^L \boldsymbol{h}^{L\top}\right\rangle [\boldsymbol{\Phi}^{L-1}]^{-1} \tag{119}$$

Again, we note that in the $\gamma_0 \to 0$ limit, since $\left\langle \boldsymbol{h}^L \boldsymbol{h}^L \right\rangle \sim \lambda^{-1}\boldsymbol{\Phi}^{L-1}$, so that $\hat{\boldsymbol{\Phi}}^{L-1} = 0$, implying that the h^{L-1} fields are also Gaussian in this $\gamma_0 \to 0$ limit. For arbitrary γ_0, this recursive argument can be completed going backwards using

$$\boldsymbol{\Phi}^\ell = \left\langle \phi(\boldsymbol{h}^\ell)\phi(\boldsymbol{h}^\ell)^\top \right\rangle \ , \ \hat{\boldsymbol{\Phi}}^{\ell-1} = \lambda[\boldsymbol{\Phi}^{\ell-1}]^{-1} - \lambda^2[\boldsymbol{\Phi}^{\ell-1}]^{-1}\left\langle \boldsymbol{h}^\ell \boldsymbol{h}^{\ell\top}\right\rangle [\boldsymbol{\Phi}^{\ell-1}]^{-1} \tag{120}$$

For deep linear networks, the distributions are all Gaussian, allowing one to close algebraically, the saddle point equations for $\Phi, \hat{\Phi}$ [91].

L Momentum Dynamics

Standard gradient descent often converges slowly and requires careful tuning of learning rate. Momentum, in contrast can, be stable under a wider range of learning rates and can benefit from acceleration on certain problems [92–95]. In this section we show that our field theory is still valid when training with momentum; simply altering the field definitions appropriately gives the infinite-width feature learning behavior.

Momentum uses a low-pass filtered version of the gradients to update the weights. A continuous limt of momentum dynamics on the trainable parameters $\{\boldsymbol{W}^\ell\}$ would give the following differential equations.

$$\begin{aligned}
&\frac{\partial}{\partial t}\boldsymbol{W}^\ell(t) = \boldsymbol{Q}^\ell(t) \\
&\tau\frac{d}{dt}\boldsymbol{Q}^\ell(t) = -\boldsymbol{Q}^\ell + \frac{\gamma}{N}\sum_\mu \Delta_\mu(t)\boldsymbol{g}^{\ell+1}_\mu(t)\phi(\boldsymbol{h}^\ell_\mu(t))^\top
\end{aligned} \tag{121}$$

We write the expression this way so that the small time constant $\tau \to 0$ limit corresponds to classic gradient descent. Integrating out the $\boldsymbol{Q}^\ell(t)$ variable, this gives the following weight dynamics

$$\boldsymbol{W}^\ell(t) = \boldsymbol{W}^\ell(0) + \frac{\gamma}{N\tau}\int_0^t dt' \int_0^{t'} dt'' e^{-(t'-t'')/\tau}\sum_\mu \Delta_\mu(t'')\boldsymbol{g}^{\ell+1}_\mu(t'')\phi(\boldsymbol{h}^\ell_\mu(t''))^\top \tag{122}$$

which implies the following field evolution

$$h_\mu^{\ell+1}(t) = \chi_\mu^{\ell+1}(t) + \frac{\gamma_0}{\tau}\int_0^t dt' \int_0^{t'} dt'' e^{-(t'-t'')/\tau} \sum_\alpha \Delta_\alpha(t'') g_\alpha^{\ell+1}(t'') \Phi_{\mu\alpha}^\ell(t,t'')$$
$$z_\mu^\ell(t) = \xi_\mu^\ell(t) + \frac{\gamma_0}{\tau}\int_0^t dt' \int_0^{t'} dt'' e^{-(t'-t'')/\tau} \sum_\alpha dt'' \Delta_\alpha(t'') \phi(h_\alpha^\ell(t'')) G_{\mu\alpha}^{\ell+1}(t,t'') \tag{123}$$

We see that in the $\tau \to 0$ limit, the t'' integral is dominated by the contribution at $t'' \sim t'$ recovering usual gradient descent dynamics. For $\tau \gg 0$, we see that the integral accumulates additional contributions from the past values of fields and kernels.

M Discrete Time

Our model can also be accommodated in discrete time, though we lose the NTK as a key player in the theory (note that $\frac{d}{dt} f_\mu = \frac{df_\mu}{d\theta} \cdot \frac{d\theta}{dt} = \sum_\alpha \Delta_\alpha K_{\mu\alpha}^{NTK}$ requires a continuous time limit of the gradient descent dynamics). For a discrete time analysis we let $t \in \mathbb{N}$ and define our network function as

$$f_\mu(t) = \frac{1}{N\gamma_0} \boldsymbol{w}^L(t) \cdot \phi(\boldsymbol{h}_\mu^L(t)) = \frac{1}{N\gamma_0}\left[\boldsymbol{w}^L(0) + \gamma_0 \sum_{s=0}^{t-1}\sum_\alpha \Delta_\alpha(s)\phi(\boldsymbol{h}_\alpha^L(s))\right] \cdot \phi(\boldsymbol{h}_\mu^L(t))$$
$$= \frac{1}{N\gamma_0} \boldsymbol{w}^L(0) \cdot \phi(\boldsymbol{h}_\mu^L(t)) + \sum_\alpha \sum_{s<t} \Delta_\alpha(s) \Phi_{\mu\alpha}^L(t,s) \tag{124}$$

We treat $f_\mu(t)$ as a potentially random variable and insert

$$1 = \int \frac{d\hat{f}_\mu(t) df_\mu(t)}{2\pi N^{-1}} \exp\left(i\hat{f}_\mu(t)\left[N f_\mu(t) - \frac{1}{\gamma_0}\boldsymbol{w}^L(0)\cdot\phi(\boldsymbol{h}^L(t)) - N\sum_\alpha\sum_{s<t}\Delta_\alpha(s)\Phi_{\mu\alpha}^L(t,s)\right]\right) \tag{125}$$

Noting that $\boldsymbol{w}^L(0)$ is involved in the definition of both $f_\mu(t)$ and $\boldsymbol{\xi}_\mu^L(t)$, we see that the average over $\boldsymbol{w}^L(0)$ now takes the form

$$\left\langle \exp\left(i\sum_{\mu t}[\hat{\boldsymbol{\xi}}_\mu^L(t) + \gamma_0^{-1}\hat{f}_\mu(t)\phi(\boldsymbol{h}_\mu^L(t))]\cdot \boldsymbol{w}^L(0)\right)\right\rangle_{\boldsymbol{w}^L(0)} = \exp\left(-\frac{1}{2}\sum_{\mu t\alpha s}\hat{\boldsymbol{\xi}}_\mu^L(t)\cdot\hat{\boldsymbol{\xi}}_\alpha^L(s)\right)$$
$$\exp\left(-\frac{N}{2\gamma_0^2}\sum_{\mu\alpha ts}\hat{f}_\mu(t)\hat{f}_\alpha(s)\Phi_{\mu\alpha}^L(t,s)\right)$$
$$\exp\left(-\frac{1}{\gamma_0}\sum_{\mu\alpha ts}\hat{f}_\mu(t)\phi(\boldsymbol{h}_\mu^L(t))\cdot\hat{\boldsymbol{\xi}}_\alpha^L(s)\right) \tag{126}$$

We extend our definition as before $iA_{\mu\alpha}^L(t,s) = \frac{1}{N\gamma_0}\phi(\boldsymbol{h}_\mu^L(t))\cdot\boldsymbol{\xi}_\alpha^L(s)$. Proceeding with the calculation as usual, we find that

$$Z \propto \int df_\mu(t) d\hat{f}_\mu(t) d\Phi^\ell ... dB^\ell \exp\left(NS[\{f,\hat{f},\Phi^\ell,\hat{\Phi}^\ell,...,A^\ell,B^\ell\}]\right)$$
$$S = i\sum_{\mu t}\hat{f}_\mu(t)f_\mu(t) - \frac{1}{2\gamma_0^2}\sum_{\mu\alpha ts}\hat{f}_\mu(t)\hat{f}_\alpha(s)\Phi_{\mu\alpha}^L(t,s) - i\sum_{\mu\alpha ts}\hat{f}_\mu(t)A_{\mu\alpha}^L(t,s) - i\sum_{\mu t\alpha s}\hat{f}_\mu(t)[\Theta(t-s)\Delta_\alpha(s)\Phi_{\mu\alpha}^L(t,s)]$$
$$+ \sum_{\ell\mu\alpha ts}[\Phi_{\mu\alpha}^\ell\hat{\Phi}_{\mu\alpha}^\ell(t,s) + G_{\mu\alpha}^\ell(t,s)\hat{G}_{\mu\alpha}(t,s) - A_{\mu\alpha}^\ell(t,s)B_{\mu\alpha}^\ell(t,s)]$$
$$+ \ln\mathcal{Z}[\{\Phi^\ell,\hat{\Phi}^\ell,...,A^\ell,B^\ell\}] \tag{127}$$

The saddle point equations can now be analyzed. In addition to the usual order parameters, we note that $f, \hat{f}$ also generate saddle point equations

$$\frac{\partial S}{\partial f_\mu(t)} = i\hat{f}_\mu(t) = 0$$
$$\frac{\partial S}{\partial i\hat{f}_\mu(t)} = f_\mu(t) + \frac{1}{\gamma_0^2}\sum_{\alpha s}\Phi^L_{\mu\alpha}(t,s)(i\hat{f}_\alpha(s)) - \sum_{\alpha s}A^L_{\mu\alpha}(t,s) - \sum_{\alpha s}\Theta(t-s)\Delta_\alpha(s)\Phi^L_{\mu\alpha}(t,s) \tag{128}$$

We also obtain saddle point equations for the new A^L, B^L order parameters.

$$\frac{\partial S}{\partial A^L_{\mu\alpha}(t,s)} = -B^L_{\mu\alpha}(t,s) - i\hat{f}_\mu(t) = 0 \tag{129}$$
$$\frac{\partial S}{\partial B^L_{\mu\alpha}(t,s)} = -A^L_{\mu\alpha}(t,s) + i\gamma_0^{-1}\left\langle \phi(h^L_\mu(t))\hat{\xi}^L_\alpha(s)\right\rangle = 0 \tag{130}$$

which implies $B^L_{\mu\alpha}(t,s) = 0$ and $A^L = \gamma_0^{-1}\left\langle \frac{\phi(h^L_\mu(t))}{\partial r^L_\alpha(s)}\right\rangle$. This gives the following DMFT

$$\begin{aligned}
f_\mu(t) &= \sum_{s<t}\sum_\alpha \Phi^L_{\mu\alpha}(t,s)\Delta_\alpha(s) + \sum_{\alpha s}A^L_{\mu\alpha}(t,s)\\
\boldsymbol{u}^\ell &\sim \mathcal{N}(0,\boldsymbol{\Phi}^{\ell-1})\ ,\ \boldsymbol{r}^\ell \sim \mathcal{N}(0,\boldsymbol{G}^{\ell+1})\\
h^\ell_\mu(t) &= u^\ell_\mu(t) + \gamma_0\sum_{\alpha s}[A^{\ell-1}_{\mu\alpha}(t,s) + \Theta(t-s)\Delta_\alpha(s)\Phi^{\ell-1}_{\mu\alpha}(t,s)]g^\ell_\alpha(s)\\
z^\ell_\mu(t) &= r^\ell_\mu(t) + \gamma_0\sum_{\alpha s}[B^\ell_{\mu\alpha}(t,s) + \Theta(t-s)\Delta_\alpha(s)G^{\ell+1}_{\mu\alpha}(t,s)]\phi(h^\ell_\alpha(s))\\
\Phi^\ell_{\mu\alpha}(t,s) &= \left\langle \phi(h^\ell_\mu(t))\phi(h^\ell_\alpha(s))\right\rangle\ ,\ G^\ell_{\mu\alpha}(t,s) = \left\langle g^\ell_\mu(t)g^\ell_\alpha(s)\right\rangle\\
A^\ell_{\mu\alpha}(t,s) &= \gamma_0^{-1}\left\langle \frac{\partial\phi(h^\ell_\mu(t))}{\partial u^\ell_\alpha(s)}\right\rangle\ ,\ B^\ell_{\mu\alpha}(t,s) = \gamma_0^{-1}\left\langle \frac{\partial g^{\ell+1}_\mu(t)}{\partial r^{\ell+1}_\alpha(s)}\right\rangle .
\end{aligned} \tag{131}$$

We leave it to future work to verify that a continuous time limit of the above DMFT recovers function evolution governed by the NTK.

N Equivalent Parameterizations

In this section, we show the equivalence of our parameterization scheme with many alternatives including the μP parameterization of Yang [1]. We also compare the derived stochastic processes obtained with DMFT and Tensor Programs in Appendix Section N.6. Following Yang we use a modified variant of abc parameterization (we assume one which defines the features $\boldsymbol{h}^{\ell+1} = N^{-a_\ell}\boldsymbol{W}^\ell\phi(\boldsymbol{h}^\ell)$ with $W^\ell_{ij} \sim \mathcal{N}(0, N^{-b_\ell})$ and $\eta = \gamma^2 N^{-c}$). Lastly, we will take $\gamma = \gamma_0 N^d$ and find that only $d = \frac{1}{2}$ will allow feature learning

$$\begin{aligned}
&\boldsymbol{h}^1 = D^{-a_0}\boldsymbol{W}^0\boldsymbol{x}_\mu\ ,\ W^0_{ij} \sim \mathcal{N}(0, D^{-b_0})\\
&\boldsymbol{h}^{\ell+1} = N^{-a_\ell}\boldsymbol{W}^\ell\phi(\boldsymbol{h}^\ell)\ ,\ W^\ell_{ij} \sim \mathcal{N}(0, N^{-b})\\
&f = \frac{1}{\gamma}h^{L+1}\ ,\ h^{L+1} = N^{-a_L}\boldsymbol{w}^L\cdot\phi(\boldsymbol{h}^L)\ , w^L_i \sim \mathcal{N}(0, N^{-b})\\
&\gamma = \gamma_0 N^d\ ,\ \gamma_0 = \mathcal{O}_N(1)
\end{aligned} \tag{132}$$

We will now derive constraints on (a, b, c, d) which give desired large width behavior. We will identify a one-dimensional family of parameterizations which satisfy three desiderata of network training 1. finite preactivations, 2. learning in finite time, 3. feature learning.

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

As before we let the NTK be the matrix which defines network prediction dynamics $\partial_t f_\mu = \sum_\alpha K^{NTK}_{\mu\alpha}\Delta_\alpha$. We demand that this matrix be $\mathcal{O}_N(1)$ so that the network predictions have $\mathcal{O}_N(1)$ evolution

$$\begin{aligned}
K^{NTK}_{\mu\alpha} &= \gamma^2 N^{-c}\sum_\ell \frac{\partial f_\mu}{\partial \boldsymbol{W}^\ell}\cdot\frac{\partial f_\alpha}{\partial \boldsymbol{W}^\ell} \\
&= N^{-c}\left[\frac{\phi(\boldsymbol{h}^L_\mu)\cdot\phi(\boldsymbol{h}^L_\alpha)}{N^{2a_L}} + \sum_\ell \frac{\partial h^{L+1}_\mu}{\partial \boldsymbol{h}^{\ell+1}_\mu}\cdot\frac{\partial h^{L+1}_\alpha}{\partial \boldsymbol{h}^{\ell+1}_\alpha}\frac{\phi(\boldsymbol{h}^\ell_\mu)\cdot\phi(\boldsymbol{h}^\ell_\alpha)}{N^{2a_\ell}} + \frac{\partial h^{L+1}_\mu}{\partial \boldsymbol{h}^1_\mu}\cdot\frac{\partial h^{L+1}_\alpha}{\partial \boldsymbol{h}^1_\alpha}\frac{\boldsymbol{x}_\mu\cdot\boldsymbol{x}_\alpha}{D^{2a_0}}\right] \\
&= N^{-c}\left[N^{1-2a_L}\Phi^L_{\mu\alpha} + \sum_\ell N^{1-2a_\ell}\frac{\partial h^{L+1}_\mu}{\partial \boldsymbol{h}^{\ell+1}_\mu}\cdot\frac{\partial h^{L+1}_\alpha}{\partial \boldsymbol{h}^{\ell+1}_\alpha}\Phi^\ell_{\mu\alpha} + D^{1-2a_0}\frac{\partial h^{L+1}_\mu}{\partial \boldsymbol{h}^1_\mu}\cdot\frac{\partial h^{L+1}_\alpha}{\partial \boldsymbol{h}^1_\alpha}\boldsymbol{K}^x\right]
\end{aligned} \tag{133}$$

where we used the usual definition of the kernels $\Phi^\ell = \frac{1}{N}\phi(\boldsymbol{h}^\ell)\cdot\phi(\boldsymbol{h}^\ell)$ which are $\mathcal{O}_N(1)$ provided each neuron's preactivation $h^\ell_i = \mathcal{O}_N(1)$. We see that the choice $a_\ell = \frac{1}{2}$ recovers the parameterization discussed in Appendix D. Further to have $\mathcal{O}_N(1)$ evolution of the output predictions f_μ we need $K^{NTK} = \mathcal{O}_N(1)$. Now, to enforce the $\mathcal{O}_N(1)$ evolution of predictions we demand

$$\begin{aligned}
N^{1-c-2a_\ell} &= \mathcal{O}_N(1)\ ,\ \ell\in\{1,...,L\} \\
N^{-c}D^{1-2a_0} &= \mathcal{O}_N(1)
\end{aligned} \tag{134}$$

If, on the other hand, we take $D\sim\mathcal{O}_N(N)$, then this simply demands the constraint that $c = 2a_\ell - 1$ for all $\ell\in\{0,...,L\}$.

N.2 Fields Are $\mathcal{O}_N(1)$

Having fields which are $\mathcal{O}_N(1)$ can be ensured at initialization provided that

$$\begin{aligned}
\left\langle h^{\ell+1}_i h^{\ell+1}_j\right\rangle &= N^{-2a_\ell}\sum_{k,k'}\left\langle W^\ell_{ik}(0)W^\ell_{jk'}(0)\right\rangle\phi(h^\ell_k)\phi(h^\ell_{k'}) \\
&= \delta_{ij}N^{1-2a_\ell-b_\ell}\Phi^\ell = \mathcal{O}_N(1)
\end{aligned} \tag{135}$$

which implies that $2a_\ell + b_\ell = 1$. Again we see that $a_\ell = \frac{1}{2}, b_\ell = 0$ works, but this is not the only possible scaling. Alternatively standard parameterization $a_\ell = 0, b_\ell = 1$ will also preserve the $\mathcal{O}_N(1)$ scale of the features. We next need to analyze the scale of the feature gradients $\frac{\partial h^{L+1}}{\partial \boldsymbol{h}^\ell}$. We start with the last layer

$$\begin{aligned}
\frac{\partial h^{\ell+1}}{\partial \boldsymbol{h}^L} &= N^{-a_L}\boldsymbol{w}^L\odot\dot\phi(\boldsymbol{h}^L) \\
\implies \frac{\partial h^{\ell+1}}{\partial \boldsymbol{h}^L}\cdot\frac{\partial h^{\ell+1}}{\partial \boldsymbol{h}^L} &= \mathcal{O}_N(N^{1-2a_\ell-b_\ell})
\end{aligned} \tag{136}$$

Since we already demanded that $2a_L + b_L = 1$, this inner product will be $\mathcal{O}_N(1)$. Now we will see whether it remains $\mathcal{O}_N(1)$ under its recursion

$$\frac{\partial h^{L+1}}{\partial \boldsymbol{h}^\ell} = \left(\frac{\partial \boldsymbol{h}^{\ell+1}}{\partial \boldsymbol{h}^\ell}\right)^\top\frac{\partial h^{L+1}}{\partial \boldsymbol{h}^{\ell+1}} = \dot\phi(\boldsymbol{h}^\ell)\odot\left[N^{-a_\ell}\boldsymbol{W}^\ell(0)^\top\frac{\partial h^{L+1}}{\partial \boldsymbol{h}^\ell}\right] \tag{137}$$

Now, letting $\boldsymbol{g}^\ell = \sqrt{N}\frac{\partial h^{L+1}}{\partial \boldsymbol{h}^\ell}$ and $\boldsymbol{z}^\ell = N^{-a_\ell}\boldsymbol{W}^\ell(0)^\top\boldsymbol{g}^{\ell+1}$ we have

$$\langle z_i z_j\rangle = \delta_{ij}N^{1-2a_\ell-b_\ell}G^{\ell+1} \tag{138}$$

which is indeed $\mathcal{O}_N(1)$ as desired provided that $2a + b = 1$.

N.3 $\mathcal{O}_N(1)$ Feature Evolution

Now, we desire that the fields h_i, z_i all evolve by an $\mathcal{O}_N(1)$ amount during network training, which is equivalent to stable feature learning. The update equation for $\boldsymbol{W}^\ell$ and $\boldsymbol{h}^\ell$ give

$$\frac{d}{dt}\boldsymbol{W}^\ell = \gamma N^{-c-a_\ell} \sum_\mu \Delta_\mu \frac{\partial h^{L+1}}{\partial \boldsymbol{h}_\mu^{\ell+1}} \boldsymbol{\phi}_\mu^{\ell\top} = \gamma_0 N^{d-c-a_\ell-\frac{1}{2}} \sum_\mu \Delta_\mu \boldsymbol{g}_\mu^{\ell+1} \boldsymbol{\phi}_\mu^{\ell\top}$$

$$\boldsymbol{h}_\mu^{\ell+1}(t) = \boldsymbol{\chi}_\mu^{\ell+1}(t) + \gamma_0 N^{d-c-2a_\ell+\frac{1}{2}} \sum_\alpha \int_0^t ds \Delta_\alpha(s) \boldsymbol{g}_\alpha^{\ell+1}(s) \Phi_{\mu\alpha}^\ell(t,s) \tag{139}$$

where we used $\gamma = \gamma_0 N^d$. This equation implies that $d - c - 2a_\ell + \frac{1}{2} = 0$ is necessary and sufficient for $\mathcal{O}_N(1)$ feature evolution.

N.4 Putting Constraints Together

We now let $\gamma = \gamma_0 N^d$. We see that the set of parameterizations which yield $\mathcal{O}(1)$ feature evolution are those for which

1. Features h, z are $\mathcal{O}_N(1) \implies 2a_\ell + b_\ell = 1$
2. Outputs predictions evolve in $\mathcal{O}_N(1)$ time $\implies c + 2a_\ell = 1$
3. Features h, z have $\mathcal{O}_N(1)$ evolution $\implies d = c + 2a_\ell - \frac{1}{2} = \frac{1}{2}$.

We see that the parameterization discussed in Appendix D 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. The set of parameterizations which meet these three requirements is one dimensional with $d = \frac{1}{2}$, and $(a_\ell, b_\ell, c_\ell) \in \{(a_\ell, 1 - 2a_\ell, 1 - 2a_\ell) : a_\ell \in \mathbb{R}\}$. However, in the next section, we show that if one demands $\mathcal{O}_N(1)$ learning rate, then the parameterization is unique and is the μP parameterization of Yang and Hu [1].

N.5 $\mathcal{O}_N(1)$ Learning Rate

We are also interested in a parameterization for which we can have $\mathcal{O}(1)$ learning rate which are those for which $\gamma^2 N^{-c} = \mathcal{O}_N(N^{2d-c}) = \mathcal{O}_N(1) \implies c = 2d = 1$. Under this constraint, $a_\ell = 0$ and $b_\ell = 1$, which corresponds to standard parameterization, modified by $\gamma = \gamma_0\sqrt{N}$ in the last layer. In a computational algorithm, the learning rate would be $\eta = \gamma^2 N^{-c} = \gamma_0^2$. This is equivalent to the μP parameterization of Yang and Hu [1].

N.6 Equivalence of DMFT at $\gamma_0 = 1$ and Tensor Programs derived Stochastic Process

Now that we have established that the parameterization we consider here (modified NTK parameterization) is equivalent to μP, (modified standard parameterization), we will now demonstrate that the stochastic process which we obtained through a stationary action principle applied to our DMFT action S is equivalent to the stochastic process derived from the Tensor Programs framework of Yang [90, 1]. Using the notation from Appendix H of Yang and Hu [1], they give the following evolution equations for the preactivations in a hidden layer in one pass SGD

$$Z^{h_t} = \hat{Z}^{Wx_t} + \dot{Z}^{Wx_t} - \sum_{s=0}^{t-1} \chi_s Z^{dh_s} \mathbb{E}[Z^{x_s} Z^{x_t}]$$

$$Z^{dx_t} = \hat{Z}^{W^\top dh_t} + \dot{Z}^{W^\top dh_t} - \sum_{s=0}^{t-1} \chi_s Z^{x_s} \mathbb{E}[Z^{dh_t} Z^{dh_s}] \tag{140}$$

where $\hat{Z}^{Wx_t}$ is mean zero Gaussian variable with covariance $\mathbb{E}[Z^{x_t} Z^{x_s}]$ and $\hat{Z}^{W^\top dh_t}$ is mean zero Gaussian with covariance $\mathbb{E}[Z^{dh_t} Z^{dh_s}]$. We can switch to the notation of this work by making the substitutions $Z^{h_t} \to h(t)$, $\hat{Z}^{Wx_t} \to u(t)$, $\chi_s \to -\Delta(s)$, $\dot{Z}^{Wx} \to \sum_s \Delta(s) A(t,s)$ and $\mathbb{E}[Z^{x_s} Z^{x_t}] \to \Phi(t,s)$, and so on. A summary of the full set of notational substitutions between this work and TP are summarized in Table 2.
After these substitutions are made, we see that the equations above match the one-pass SGD version

DMFT	$h(t)$	$\chi(t)$	$g(t)$	$\xi(t)$	$\Phi^\ell(t,s)$	$G^\ell(t,s)$	$A^\ell(t,s), B^\ell(t,s)$	$\Delta(t)$
TP	Z^{h_t}	Z^{Wx_t}	Z^{dx_t}	$Z^{W^\top dh_t}$	$\mathbb{E}[Z^{x_t}Z^{x_s}]$	$\mathbb{E}[Z^{dh_t}Z^{dh_s}]$	θ_{ts}	$-\chi_t$

Table 2: A dictionary relating the notation of the Tensor Programs (TP) framework [1] and this work.

of the DMFT Equations in Appendix M. A similar identification can be made for the backward pass. This shows that both Tensor Programs and DMFT, though alternative derivations, give identical descriptions of the stochastic processes induced by random initializations + GD in infinite neural networks.

O Gradient Independence

The gradient independence approximation treats the random initial weight matrix $\boldsymbol{W}^\ell(0)$ as a *independently sampled Gaussian matrix* when used in the backward pass. We let this second matrix be $\tilde{\boldsymbol{W}}^\ell(0)$. As before, we have $\boldsymbol{\chi}^{\ell+1} = \frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(0)\boldsymbol{\phi}(\boldsymbol{h}^\ell)$, however we now define $\boldsymbol{\xi}^\ell = \frac{1}{\sqrt{N}}\tilde{\boldsymbol{W}}^\ell(0)^\top \boldsymbol{g}^{\ell+1}$. Now, when computing the moment generating function Z, the integrals over $\boldsymbol{W}^\ell(0)$ and $\tilde{\boldsymbol{W}}^\ell(0)$ factorize

$$\begin{aligned}
&\left\langle \exp\left(\frac{i}{\sqrt{N}}\int_0^\infty dt\left[\sum_\mu \hat{\boldsymbol{\chi}}^{\ell+1}(t)\boldsymbol{W}^\ell(0)\phi(\boldsymbol{h}^\ell_\mu(t)) + \boldsymbol{g}^{\ell+1}_\mu(t)^\top \tilde{\boldsymbol{W}}^\ell(0)\boldsymbol{\xi}^\ell_\mu(t)\right]\right)\right\rangle \\
&= \exp\left(-\frac{1}{2}\sum_{\mu\alpha}\int_0^\infty dt'\int_0^\infty ds'\left[\hat{\boldsymbol{\chi}}^{\ell+1}_\mu(t)\cdot\hat{\boldsymbol{\chi}}^{\ell+1}_\alpha(s)\boldsymbol{\Phi}^\ell_{\mu\alpha}(t,s) + \hat{\boldsymbol{\xi}}^\ell_\mu(t)\cdot\hat{\boldsymbol{\xi}}^\ell_\alpha(s)G^{\ell+1}_{\mu\alpha}(t,s)\right]\right).
\end{aligned} \tag{141}$$

We see that in this field theory, the fields χ, ξ are all independent Gaussian processes $\{\chi^{\ell+1}_\mu(t)\} \sim \mathcal{GP}(0, \boldsymbol{\Phi}^\ell)$ and $\{\xi^\ell_\mu(t)\} \sim \mathcal{GP}(0, \boldsymbol{G}^{\ell+1})$. This corresponds to making the assumption that $\boldsymbol{A}^\ell = \boldsymbol{B}^\ell = 0$ so that $\chi = u$ and $\xi = r$ within the full DMFT.

P Perturbation Theory

P.1 Small γ_0 Expansion

In this section we analyze the leading corrections in a small γ_0 expansion of our DMFT theory. All fields are expanded in power series in γ_0.

$$\begin{aligned}
h^\ell_\mu(t) - u^\ell_\mu(t) &= \sum_{n=1}^\infty \gamma_0^n h^{\ell,(n)}_\mu(t) \\
z^\ell_\mu(t) - r^\ell_\mu(t) &= \sum_{n=1}^\infty \gamma_0^n z^{\ell,(n)}_\mu(t)
\end{aligned} \tag{142}$$

Our goal is to calculate all corrections to the kernels up to $\mathcal{O}(\gamma_0^3)$ to show that the leading correction is $\mathcal{O}(\gamma_0^2)$ and the subleading correction is $\mathcal{O}(\gamma_0^4)$. It will again be convenient to utilize the vector notation defined in D.

We note that unlike other works on perturbation theory in wide networks, we do not attempt to characterize fluctuation effects in the kernels due to finite width, but rather operate in a regime where the kernels are concentrating and their variance is negligible. For a more thorough discussion of perturbative field theory in finite width networks, see [27, 28, 35].

P.1.1 Linear Network

The kernels in deep linear networks can be expanded in powers of γ_0^2 giving a leading order correction of size $\mathcal{O}(\gamma_0^2)$ and can be computed explicitly from the closed saddle point equations. We use the symmetrizer $\{\boldsymbol{X}, \boldsymbol{Y}\}_{sym} = \boldsymbol{X}\boldsymbol{Y} + \boldsymbol{Y}^\top\boldsymbol{X}^\top$ as shorthand. The leading order behavior of

$\boldsymbol{C}^\ell \sim \boldsymbol{C}^{(0)} + \mathcal{O}(\gamma_0^2)$, $\boldsymbol{D}^\ell \sim \boldsymbol{D}^{(0)} + \mathcal{O}(\gamma_0^2), \boldsymbol{H}^{\ell,0} = \boldsymbol{H}^{(0)} = \boldsymbol{K}^x \otimes \boldsymbol{1}\boldsymbol{1}^\top, \boldsymbol{G}^{\ell,(0)} = \boldsymbol{G}^{(0)} = \boldsymbol{1}\boldsymbol{1}^\top$ is independent of layer index so we find the following leading order corrections

$$\begin{aligned}
\boldsymbol{H}^\ell &\sim \boldsymbol{H}^{(0)} + \ell\gamma_0^2 \left(\{\boldsymbol{C}^{(0)}\boldsymbol{D}^{(0)}, \boldsymbol{H}^{(0)}\}_{sym} + \boldsymbol{C}^{(0)}\boldsymbol{1}\boldsymbol{1}^\top \boldsymbol{C}^{(0)\top} \right) + \mathcal{O}(\gamma_0^4) \\
\boldsymbol{G}^\ell &\sim \boldsymbol{1}\boldsymbol{1}^\top + (L+1-\ell)\gamma_0^2 (\{\boldsymbol{D}^{(0)}\boldsymbol{C}^{(0)}, \boldsymbol{1}\boldsymbol{1}^\top\}_{sym} + \boldsymbol{D}^{(0)}\boldsymbol{H}^{(0)}[\boldsymbol{D}^{(0)}]^\top) + \mathcal{O}(\gamma_0^4) \\
\boldsymbol{K}^{NTK} &\sim L\boldsymbol{H}^0 + \gamma_0^2 \frac{L(L+1)}{2} \left(\{\boldsymbol{C}^{(0)}\boldsymbol{D}^{(0)}, \boldsymbol{K}^x\}_{sym} + \boldsymbol{C}^{(0)}\boldsymbol{1}\boldsymbol{1}^\top \boldsymbol{C}^{(0)\top} \right) \\
&\quad + \gamma_0^2 \frac{L(L+1)}{2} \boldsymbol{K}^x \otimes (\{\boldsymbol{D}^{(0)}\boldsymbol{C}^{(0)}, \boldsymbol{1}\boldsymbol{1}^\top\}_{sym} + \boldsymbol{D}^{(0)}\boldsymbol{H}^{(0)}[\boldsymbol{D}^{(0)}]^\top) + O(\gamma_0^4)
\end{aligned} \tag{143}$$

Note that $[\boldsymbol{C}^0\boldsymbol{g}]_{\mu t} = \int_0^t dt' \sum_\beta H^0_{\mu\beta}(t,t')\Delta_\beta(t')g(t') = \sum_\beta K^x_{\mu\beta} \int_0^t dt' \Delta_\beta(t')g(t')$ and note that $[\boldsymbol{D}\boldsymbol{h}]_t = \int_0^t dt' G^0(t,t') \sum_\alpha \Delta_\alpha(t')h_\alpha(t') = \sum_\alpha \int_0^t dt' \Delta_\alpha(t')h_\alpha(t')$.

$$\begin{aligned}
H^\ell_{\mu\nu}(t,s) &= K^x_{\mu\nu} \\
&\quad + \ell\gamma_0^2 \sum_{\alpha\beta} K^x_{\mu\alpha}K^x_{\nu\beta} \int_0^t dt' \Delta_\alpha(t') \int_0^{t'} dt'' \Delta_\beta(t'') + ((\mu,t) \leftrightarrow (\nu,s)) \\
&\quad + \ell\gamma_0^2 \sum_{\alpha\beta} K^x_{\mu\alpha}K^x_{\nu\beta} \left[\int_0^t dt' \Delta_\alpha(t') \right] \left[\int_0^s ds' \Delta_\beta(s') \right] \\
G^\ell(t,s) &= 1 + \gamma_0^2 (L+1-\ell) \sum_{\alpha\beta} K^x_{\alpha\beta} \int_0^t dt' \Delta_\alpha(t') \int_0^{t'} dt'' \Delta_\beta(t'') + (t \leftrightarrow s) \\
&\quad + \gamma_0^2 (L+1-\ell) \sum_{\alpha\beta} K^x_{\mu\alpha} \left[\int_0^t dt' \Delta_\alpha(t') \right] \left[\int_0^s ds' \Delta_\alpha(s') \right]
\end{aligned} \tag{144}$$

We can simplify the notation by introducing functions $v_\alpha(t) = \int_0^t \Delta_\alpha(t')$ and $v_{\alpha\beta}(t) = \int_0^t dt' \Delta_\alpha(t') \int_0^{t'} dt'' \Delta_\beta(t'')$.

$$\begin{aligned}
H^\ell_{\mu\nu}(t,s) &= K^x_{\mu\nu} + \ell\gamma_0^2 \sum_{\alpha\beta} K^x_{\mu\alpha}K^x_{\nu\beta}[v_{\alpha\beta}(t) + v_{\beta\alpha}(s)] + \ell\gamma_0^2 \sum_{\alpha\beta} K^x_{\mu\alpha}K^x_{\nu\beta} v_\alpha(t)v_\beta(s) \\
G^\ell(t,s) &= 1 + \gamma_0^2 (L+1-\ell) \sum_{\alpha\beta} K^x_{\alpha\beta}[v_{\alpha\beta}(t) + v_{\beta\alpha}(s) + v_\alpha(t)v_\beta(s)]
\end{aligned} \tag{145}$$

Using the fact that

$$\begin{aligned}
K^{NTK}_{\mu\alpha}(t,s) &= \sum_{\ell=0}^{L} G^{\ell+1}(t,s) H^\ell_{\mu\alpha}(t,s) \\
&\sim (L+1)K^x_{\mu\alpha} + \gamma_0^2 \sum_{\ell=1}^{L} H^{\ell,2}_{\mu\alpha}(t,s) + \gamma_0^2 \sum_{\ell=1}^{L} G^{\ell,2}(t,s) K^x_{\mu\alpha} + \mathcal{O}(\gamma_0^4)
\end{aligned} \tag{146}$$

and utilizing the identity $\sum_{\ell=1}^L \ell = \frac{1}{2}L(L+1)$, we recover the result provided in the main text.

P.2 Nonlinear Perturbation Theory

We start with the formula which implicitly defines $\boldsymbol{h}, \boldsymbol{z}$

$$\boldsymbol{h}^\ell = \boldsymbol{u}^\ell + \gamma_0 \boldsymbol{C}^\ell [\dot{\phi}(\boldsymbol{h}^\ell) \odot \boldsymbol{z}^\ell] \, , \; \boldsymbol{z}^\ell = \boldsymbol{r}^\ell + \gamma_0 \boldsymbol{D}^\ell \phi(\boldsymbol{h}^\ell) \tag{147}$$

We proceed under the assumption of a power series in γ_0

$$
\begin{aligned}
\boldsymbol{h}^\ell - \boldsymbol{u}^\ell &= \gamma_0 \boldsymbol{h}^{\ell,1} + \gamma_0^2 \boldsymbol{h}^{\ell,2} + ... \\
\boldsymbol{z}^\ell - \boldsymbol{r}^\ell &= \gamma_0 \boldsymbol{z}^{\ell,1} + \gamma_0^2 \boldsymbol{z}^{\ell,2} + ... \\
\boldsymbol{\Phi}^\ell - \boldsymbol{\Phi}^{\ell,0} &= \gamma_0 \boldsymbol{\Phi}^{\ell,1} + \gamma_0^2 \boldsymbol{\Phi}^{\ell,2} + ... \\
\boldsymbol{G}^\ell - \boldsymbol{G}^{\ell,0} &= \gamma_0 \boldsymbol{G}^{\ell,1} + \gamma_0^2 \boldsymbol{G}^{\ell,2} + ... \\
\boldsymbol{C}^\ell - \boldsymbol{C}^{\ell,0} &= \gamma_0 \boldsymbol{C}^{\ell,1} + \gamma_0^2 \boldsymbol{C}^{\ell,2} + ... \\
\boldsymbol{D}^\ell - \boldsymbol{D}^{\ell,0} &= \gamma_0 \boldsymbol{D}^{\ell,1} + \gamma_0^2 \boldsymbol{D}^{\ell,2} + ...
\end{aligned} \tag{148}
$$

Expanding both sides of the implicit equation for $\boldsymbol{z}^\ell$ we have

$$
\begin{aligned}
\gamma_0 \boldsymbol{z}^{\ell,1} + \gamma_0^2 \boldsymbol{z}^{\ell,2} + ... =& \gamma_0 \boldsymbol{D}^{\ell,0} \phi(\boldsymbol{u}^\ell) \\
&+ \gamma_0^2 \left[\boldsymbol{D}^{\ell,0} \dot{\phi}(\boldsymbol{u}) \odot \boldsymbol{h}^{\ell,1} + \boldsymbol{D}^{\ell,1} \phi(\boldsymbol{u}) \right] \\
&+ \gamma_0^3 \left[\boldsymbol{D}^{\ell,0} \dot{\phi}(\boldsymbol{u}) \odot \boldsymbol{h}^{\ell,2} + \boldsymbol{D}^{\ell,0} \ddot{\phi}(\boldsymbol{u}) \odot [\boldsymbol{h}^{\ell,1}]^2 + \boldsymbol{D}^{\ell,1} \dot{\phi}(\boldsymbol{u}) \odot \boldsymbol{h}^{\ell,1} + \boldsymbol{D}^{\ell,2} \phi(\boldsymbol{u}) \right] \\
&+ \mathcal{O}(\gamma_0^4)
\end{aligned} \tag{149}
$$

Performing a similar exercise for $\boldsymbol{h}^\ell$, we get the following first three leading terms for $\boldsymbol{z}^\ell, \boldsymbol{h}^\ell$, we find

$$
\begin{aligned}
&\boldsymbol{z}^{\ell,1} = \boldsymbol{D}^{\ell,0} \phi(\boldsymbol{u}) \\
&\boldsymbol{z}^{\ell,2} = \boldsymbol{D}^{\ell,0} \dot{\phi}(\boldsymbol{u}) \odot \boldsymbol{h}^{\ell,1} + \boldsymbol{D}^{\ell,1} \phi(\boldsymbol{u}) \\
&\boldsymbol{z}^{\ell,3} = \boldsymbol{D}^{\ell,0} \left[\frac{1}{2} \ddot{\phi}(\boldsymbol{u}) \odot [\boldsymbol{h}^{\ell,1}]^2 + \dot{\phi}(\boldsymbol{u}) \odot \boldsymbol{h}^{\ell,2} \right] + \boldsymbol{D}^{\ell,1} [\dot{\phi}(\boldsymbol{u}) \odot \boldsymbol{h}^{\ell,1}] + \boldsymbol{D}^{\ell,2} \phi(\boldsymbol{u}) \\
&\boldsymbol{h}^{\ell,1} = \boldsymbol{C}^{\ell,0} \boldsymbol{g}^{\ell,0} = \boldsymbol{C}^{\ell,0} [\dot{\phi}(\boldsymbol{u}) \odot \boldsymbol{r}] \\
&\boldsymbol{h}^{\ell,2} = \boldsymbol{C}^{\ell,1} \boldsymbol{g}^{\ell,1} + \boldsymbol{C}^{\ell,0} \boldsymbol{g}^{\ell,2} \\
&= \boldsymbol{C}^{\ell,0} \left[\dot{\phi}(\boldsymbol{u}) \boldsymbol{z}^{\ell,1} + \ddot{\phi}(\boldsymbol{u}) \boldsymbol{h}^{\ell,1} \boldsymbol{r} \right] + \boldsymbol{C}^{\ell,1} \left[\dot{\phi}(\boldsymbol{u}) \boldsymbol{z}^{\ell,2} + \ddot{\phi}(\boldsymbol{u}) \boldsymbol{h}^{\ell,1} \boldsymbol{z}^{\ell,1} + \frac{1}{2} \dddot{\phi}(\boldsymbol{u}) [\boldsymbol{h}^{\ell,1}]^2 \boldsymbol{r} + \ddot{\phi}(\boldsymbol{u}) \boldsymbol{h}^{\ell,2} \boldsymbol{r} \right] \\
&\boldsymbol{h}^{\ell,3} = \boldsymbol{C}^{\ell,0} \boldsymbol{g}^{\ell,2} + \boldsymbol{C}^{\ell,1} \boldsymbol{g}^{\ell,1} + \boldsymbol{C}^{\ell,2} \boldsymbol{g}^{\ell,0} \\
&= \boldsymbol{C}^{\ell,0} \left[\dot{\phi}(\boldsymbol{u}) \boldsymbol{z}^{\ell,2} + \ddot{\phi}(\boldsymbol{u}) \boldsymbol{h}^{\ell,1} \boldsymbol{z}^{\ell,1} + \ddot{\phi}(\boldsymbol{u}) \boldsymbol{h}^{\ell,2} \boldsymbol{r} + \frac{1}{2} \dddot{\phi}(\boldsymbol{u}) [\boldsymbol{h}^{\ell,1}]^2 \boldsymbol{r} \right] \\
&+ \boldsymbol{C}^{\ell,1} \left[\dot{\phi}(\boldsymbol{u}) \boldsymbol{z}^{\ell,1} + \ddot{\phi}(\boldsymbol{u}) \boldsymbol{h}^{\ell,1} \boldsymbol{r} \right] + \boldsymbol{C}^{\ell,2} \left[\dot{\phi}(\boldsymbol{u}) \boldsymbol{r} \right]
\end{aligned} \tag{150}
$$

As will become apparent soon, it is crucially important to identify the dependence of each of these terms on $\boldsymbol{r}$. We note that $z^{\ell,1}$ does not depend on r and $h^{\ell,1}$ is linear in r. In the next section, we use this fact to show that $\Phi^{\ell,1} = 0$ and $G^{\ell,1} = 0$. These conditions imply that $C^{\ell,0}$ and $D^{\ell,1} = 0$. As a consequence, $z^{\ell,2}$ is linear in r and $h^{\ell,2}$ only contains even powers of r. Lastly, this implies that $z^{\ell,3}$ only contains even powers of r and $h^{\ell,3}$ contains only odd powers of r.

P.2.1 Leading Corrections to Φ^1 Kernel is $\mathcal{O}(\gamma_0^2)$

We start in the first layer where $\boldsymbol{u}^1 \sim \mathcal{GP}(0, \boldsymbol{K}^x \otimes \mathbf{1}\mathbf{1}^\top)$ (note that this is $\mathcal{O}_{\gamma_0}(1)$) and compute the expansion of Φ^1 in γ_0

$$
\begin{aligned}
\boldsymbol{\Phi}^1 =& \left\langle \phi(\boldsymbol{h}^1)\phi(\boldsymbol{h}^1)^\top \right\rangle = \left\langle \phi(\boldsymbol{u}^1)\phi(\boldsymbol{u}^1)^\top \right\rangle \\
&+ \gamma_0 \left\langle \left[\dot\phi(\boldsymbol{u}^1)\boldsymbol{h}^{1,1}\right] \phi(\boldsymbol{u}^1)^\top \right\rangle + \gamma_0 \left\langle \phi(\boldsymbol{u}^1) \left[\dot\phi(\boldsymbol{u}^1)\boldsymbol{h}^{1,1}\right]^\top \right\rangle \\
&+ \gamma_0^2 \left\langle \left[\dot\phi(\boldsymbol{u}^1)\boldsymbol{h}^{1,1}\right] \left[\dot\phi(\boldsymbol{u}^1)\boldsymbol{h}^{1,1}\right]^\top \right\rangle \\
&+ \frac{\gamma_0^2}{2} \left\langle \left[\ddot\phi(\boldsymbol{u}^1)\boldsymbol{h}^{1,2}\right] \phi(\boldsymbol{u}^1) \right\rangle + \frac{\gamma_0^2}{2} \left\langle \left[\ddot\phi(\boldsymbol{u}^1)\boldsymbol{h}^{1,2}\right] \phi(\boldsymbol{u}^1) \right\rangle \\
&+ \gamma_0^3 \left\langle \left[\dot\phi(\boldsymbol{u}^1)\boldsymbol{h}^{1,3} + \ddot\phi(\boldsymbol{u})\boldsymbol{h}^{1,1}\boldsymbol{h}^{1,2} + \frac{1}{6}\dddot\phi(\boldsymbol{u})(\boldsymbol{h}^{1,1})^3\right] \phi(\boldsymbol{u})^\top \right\rangle \\
&+ \gamma_0^3 \left\langle \phi(\boldsymbol{u}) \left[\dot\phi(\boldsymbol{u}^1)\boldsymbol{h}^{1,3} + \ddot\phi(\boldsymbol{u})\boldsymbol{h}^{1,1}\boldsymbol{h}^{1,2} + \frac{1}{6}\dddot\phi(\boldsymbol{u})(\boldsymbol{h}^{1,1})^3\right]^\top \right\rangle \\
&+ \gamma_0^3 \left\langle \left[\dot\phi(\boldsymbol{u})\boldsymbol{h}^{1,2} + \frac{1}{2}\ddot\phi(\boldsymbol{u})(\boldsymbol{h}^{1,1})^2\right] \left[\dot\phi(\boldsymbol{u})\boldsymbol{h}^{1,1}\right]^\top \right\rangle \\
&+ \gamma_0^3 \left\langle \left[\dot\phi(\boldsymbol{u})\boldsymbol{h}^{1,1}\right] \left[\dot\phi(\boldsymbol{u})\boldsymbol{h}^{1,2} + \frac{1}{2}\ddot\phi(\boldsymbol{u})(\boldsymbol{h}^{1,1})^2\right]^\top \right\rangle \\
&+ \mathcal{O}(\gamma_0^4)
\end{aligned} \tag{151}
$$

where powers and multiplications of vectors are taken elementwise. Now, note that, as promised, the terms linear in γ_0 vanish since $\boldsymbol{h}^{1,1}$ is linear the Gaussian random variable $\boldsymbol{r}^1$, which is a mean zero and independent of $\boldsymbol{u}^1$ so an average like $\left\langle \boldsymbol{r}^1 F(\boldsymbol{u}^1) \right\rangle = \left\langle \boldsymbol{r}^{1,0} \right\rangle \left\langle F(\boldsymbol{u}^1) \right\rangle = 0$ must vanish for any function F. Thus we see that $\boldsymbol{\Phi}^\ell$'s leading correction is $\mathcal{O}(\gamma_0^2)$.

We also obtain, by a similar argument, that the cubic $\mathcal{O}(\gamma_0^3)$ term vanishes. To see this, note that $\boldsymbol{h}^{1,3}$ only contains odd powers of $\boldsymbol{r}^1$. Next, $\boldsymbol{h}^{1,1}\boldsymbol{h}^{1,2}$ contains only odd powers of $\boldsymbol{r}$, and $(\boldsymbol{h}^{1,1})^3$ is cubic in $\boldsymbol{r}$. Since all odd moments of a mean-zero Gaussian vanish, all averages of these terms over $\boldsymbol{r}$ annihilate, causing the γ_0^3 terms to vanish. Thus $\boldsymbol{\Phi}^1 = \boldsymbol{\Phi}^{1,0} + \gamma_0^2 \boldsymbol{\Phi}^{1,2} + \mathcal{O}(\gamma_0^4)$.

P.3 Forward Pass Induction for Φ^ℓ

We now assume the inductive hypothesis that for some $\ell \in \{1, ..., L-1\}$ that

$$
\boldsymbol{\Phi}^\ell = \boldsymbol{\Phi}^{\ell,0} + \gamma_0^2 \boldsymbol{\Phi}^{\ell,2} + \mathcal{O}(\gamma_0^4) \tag{152}
$$

and we will show that this will imply that the next layer must have a similar expansion $\boldsymbol{\Phi}^{\ell+1} = \boldsymbol{\Phi}^{\ell+1,0} + \gamma_0^2 \boldsymbol{\Phi}^{\ell+1,2} + \mathcal{O}(\gamma_0^4)$. First, we note that $\boldsymbol{u}^{\ell+1} \sim \mathcal{GP}(0, \boldsymbol{\Phi}^{\ell,0} + \gamma_0^2 \boldsymbol{\Phi}^{\ell,2} + ...)$. As before, we compute the leading terms in the expansion of $\boldsymbol{\Phi}^{\ell+1}$

$$
\begin{aligned}
\boldsymbol{\Phi}^{\ell+1} =& \left\langle \phi(\boldsymbol{h}^{\ell+1})\phi(\boldsymbol{h}^{\ell+1})^\top \right\rangle \\
=& \left\langle \phi(\boldsymbol{u}^{\ell+1})\phi(\boldsymbol{u}^{\ell+1}) \right\rangle + \gamma_0^2 \left\langle \left[\dot\phi(\boldsymbol{u}^{\ell+1})\boldsymbol{h}^{\ell+1,1}\right] \left[\dot\phi(\boldsymbol{u}^{\ell+1})\boldsymbol{h}^{\ell+1,1}\right]^\top \right\rangle \\
&+ \frac{\gamma_0^2}{2} \left\langle \left[\ddot\phi(\boldsymbol{u}^{\ell+1})\boldsymbol{h}^{\ell+1,2}\right] \phi(\boldsymbol{u}^{\ell+1})^\top \right\rangle + \frac{\gamma_0^2}{2} \left\langle \phi(\boldsymbol{u}^{\ell+1}) \left[\ddot\phi(\boldsymbol{u}^{\ell+1})\boldsymbol{h}^{\ell+1,2}\right]^\top \right\rangle + \mathcal{O}(\gamma_0^4)
\end{aligned} \tag{153}
$$

where, as before the γ_0 and γ_0^3 terms vanish by the fact that odd moments of $\boldsymbol{r}^{\ell+1}$ vanish. Now, note that all averages are performed over $\boldsymbol{u}^{\ell+1} \sim \mathcal{GP}(0, \boldsymbol{\Phi}^{\ell,0} + \gamma_0^2 \boldsymbol{\Phi}^{\ell,2} + ...)$, which depends on the perturbed kernel of the previous layer. How can we calculate the contribution of the correction which is due to the previous layer's kernel movement? This can be obtained easily from the following identity. Let $F(\boldsymbol{u}, \boldsymbol{r})$ be an arbitrary observable which depends on Gaussian fields $\boldsymbol{u}$ and $\boldsymbol{r}$ which

have covariances $\boldsymbol{\Phi}^{\ell,0} + \gamma_0^2 \boldsymbol{\Phi}^{\ell,2} + \mathcal{O}(\gamma_0^4)$ and $\boldsymbol{G}^{\ell+2,0} + \gamma_0^2 \boldsymbol{G}^{\ell+2,2} + \mathcal{O}(\gamma_0^3)$ (note this only requires that the linear in γ_0 terms of G vanish which is easy to verify). Then

$$\begin{aligned}
\langle F(\boldsymbol{u}, \boldsymbol{r}) \rangle_{\boldsymbol{u},\boldsymbol{r}} = \int d\boldsymbol{k} d\boldsymbol{u} d\boldsymbol{v} d\boldsymbol{r} F(\boldsymbol{u}, \boldsymbol{r}) \exp\left(-\frac{1}{2}\boldsymbol{k}^\top [\boldsymbol{\Phi}^{\ell,0} + \gamma_0^2 \boldsymbol{\Phi}^{\ell,2} + ...]\boldsymbol{k} + i\boldsymbol{k}\cdot\boldsymbol{u}\right) \\
\exp\left(-\frac{1}{2}\boldsymbol{v}^\top [\boldsymbol{G}^{\ell+2,0} + \gamma_0^2 \boldsymbol{G}^{\ell+2,2} + ...]\boldsymbol{v} + i\boldsymbol{v}\cdot\boldsymbol{r}\right) \qquad (154) \\
\sim \langle F(\boldsymbol{u}, \boldsymbol{r}) \rangle_{\boldsymbol{u}_0 \boldsymbol{r}_0} \\
+ \frac{\gamma_0^2}{2} \mathrm{Tr}\left[\boldsymbol{\Phi}^{\ell-1,2} \left\langle \frac{\partial^2}{\partial \boldsymbol{u} \partial \boldsymbol{u}^\top} f(\boldsymbol{u}, \boldsymbol{r}) \right\rangle_{\boldsymbol{u}_0 \boldsymbol{r}_0}\right] \\
+ \frac{\gamma_0^2}{2} \mathrm{Tr}\left[\boldsymbol{G}^{\ell+1,2} \left\langle \frac{\partial^2}{\partial \boldsymbol{r} \partial \boldsymbol{r}^\top} f(\boldsymbol{u}, \boldsymbol{r}) \right\rangle_{\boldsymbol{u}_0 \boldsymbol{r}_0}\right] + \mathcal{O}(\gamma_0^3) \qquad (155)
\end{aligned}$$

where $\boldsymbol{u}_0 \sim \mathcal{GP}(0, \boldsymbol{\Phi}^{\ell,0}), \boldsymbol{r}_0 \sim \mathcal{N}(0, \boldsymbol{G}^{\ell+2,0})$. Thus, the leading order behavior of $\boldsymbol{\Phi}^{\ell+1}$ can easily be obtained in terms of averages over the original unperturbed covariances

$$\begin{aligned}
\boldsymbol{\Phi}^{\ell+1} = \left\langle \phi(\boldsymbol{u}_0)\phi(\boldsymbol{u}_0)^\top \right\rangle_{\boldsymbol{u}_0} + \frac{\gamma_0^2}{2} \mathrm{Tr}\left[\boldsymbol{\Phi}^{\ell,2} \left\langle \frac{\partial^2}{\partial \boldsymbol{u}_0 \partial \boldsymbol{u}_0^\top} \phi(\boldsymbol{u}_0)\phi(\boldsymbol{u}_0)^\top \right\rangle_{\boldsymbol{u}_0}\right] \\
+ \frac{\gamma_0^2}{2} \frac{\partial^2}{\partial \gamma_0^2}|_{\gamma_0=0} \langle \phi(\boldsymbol{h}(\boldsymbol{u}_0, \boldsymbol{r}_0, \gamma_0))\phi(\boldsymbol{h}(\boldsymbol{u}_0, \boldsymbol{r}_0, \gamma_0)) \rangle_{\boldsymbol{u}_0, \boldsymbol{r}_0} + \mathcal{O}(\gamma_0^4), \qquad (156)
\end{aligned}$$

where the trace is taken against the Hessian indices and the indices on $\boldsymbol{\Phi}^{\ell,2}$. This gives us the desired result by induction that for all $\ell \in \{1, ..., L\}$, we have $\boldsymbol{\Phi}^\ell = \boldsymbol{\Phi}^{\ell,0} + \gamma_0^2 \boldsymbol{\Phi}^{\ell,2} + \mathcal{O}(\gamma_0^4)$. We see that Φ^ℓ accumulates corrections from the previous layers' corrections through the forward pass recursion.

P.4 Leading Corrections to G^L Kernel is $\mathcal{O}(\gamma_0^2)$

The analogous argument for $\boldsymbol{G}^L$ now can be provided. First note that $\boldsymbol{r}^L$ is independent of $\boldsymbol{u}^L$ and of γ_0. Thus we can find that $\boldsymbol{G}^L$ has no linear-in-γ_0 term in its expansion since

$$\boldsymbol{G}^{L,1} = \left\langle [\dot{\phi}(\boldsymbol{u}^L)\boldsymbol{r}^L] \left[\dot{\phi}(\boldsymbol{u}^L)\boldsymbol{z}^{L,1} + \ddot{\phi}(\boldsymbol{u}^L)\boldsymbol{h}^{L,1}\boldsymbol{r}^L\right] \right\rangle + \left\langle [\dot{\phi}(\boldsymbol{u}^L)\boldsymbol{r}^L] \left[\dot{\phi}(\boldsymbol{u}^L)\boldsymbol{z}^{L,1} + \ddot{\phi}(\boldsymbol{u}^L)\boldsymbol{h}^{L,1}\boldsymbol{r}^L\right] \right\rangle = 0$$

each term contains only odd powers of $\boldsymbol{r}^L$ and odd moments of Gaussian variables vanish. After much more work, one can verify that $\boldsymbol{G}^{L,3}$ also must vanish since all terms contain odd powers of $\boldsymbol{r}$.

$$\boldsymbol{G}^{L,3} = \left\langle \boldsymbol{g}^{L,3}\boldsymbol{g}^{L,0\top} \right\rangle + \left\langle \boldsymbol{g}^{L,0}\boldsymbol{g}^{L,3\top} \right\rangle + \left\langle \boldsymbol{g}^{L,2}\boldsymbol{g}^{L,1\top} \right\rangle + \left\langle \boldsymbol{g}^{L,1}\boldsymbol{g}^{L,2\top} \right\rangle \qquad (157)$$

First, note that $\boldsymbol{g}^{L,0}$ is linear in $\boldsymbol{r}$. Next, note that $\boldsymbol{g}^{L,1}$ only depends on even powers of $\boldsymbol{r}$ since $\boldsymbol{g}^{L,1} = \dot{\phi}(\boldsymbol{u})\boldsymbol{z}^{L,1} + \ddot{\phi}(\boldsymbol{u})\boldsymbol{h}^{L,1}\boldsymbol{r}$. Next, we have

$$\boldsymbol{g}^{L,2} = \dot{\phi}(\boldsymbol{u})\boldsymbol{z}^{L,2} + \ddot{\phi}(\boldsymbol{u})[\boldsymbol{h}^{L,2}\boldsymbol{r} + \boldsymbol{h}^{L,1}\boldsymbol{z}^{L,1}] + \frac{1}{2}\dddot{\phi}(\boldsymbol{u})[\boldsymbol{h}^{L,1}]^2 \qquad (158)$$

which only depends on odd powers of $\boldsymbol{r}$. Lastly, we have $\boldsymbol{g}^{L,3}$

$$\begin{aligned}
\boldsymbol{g}^{L,3} = \dot{\phi}(\boldsymbol{u})\boldsymbol{z}^{L,3} + \ddot{\phi}(\boldsymbol{u})[\boldsymbol{h}^{L,3}\boldsymbol{r} + \boldsymbol{h}^{L,2}\boldsymbol{z}^{L,1} + \boldsymbol{h}^{L,1}\boldsymbol{z}^{L,2}] \\
+ \frac{1}{2}\dddot{\phi}(\boldsymbol{u})[2\boldsymbol{h}^{L,1}\boldsymbol{h}^{L,2}\boldsymbol{r} + [\boldsymbol{h}^{L,1}]^2\boldsymbol{z}^{L,1}] + \frac{1}{6}\phi^{(4)}(\boldsymbol{u})[\boldsymbol{h}^{L,1}]^3\boldsymbol{r} \qquad (159)
\end{aligned}$$

which we see only contains even powers of $\boldsymbol{r}$. Thus $\boldsymbol{g}^{L,3}\boldsymbol{g}^{L,0}$ will be odd in $\boldsymbol{r}$. Looking at the expansion for $\boldsymbol{G}^{L,3}$, we see that all terms are odd in $\boldsymbol{r}$ and so the averages vanish under the Gaussian integrals.

P.5 Backward Pass Recursion for G^ℓ

We can derive a similar recursion on the backward pass for $\boldsymbol{G}^\ell$'s leading order corrections. Using the same idea from the previous section, we find the following expressions

$$\boldsymbol{G}^\ell = \left\langle \left[\dot\phi(\boldsymbol{u}_0)\boldsymbol{r}_0\right]\left[\dot\phi(\boldsymbol{u}_0)\boldsymbol{r}_0\right]^\top \right\rangle_{\boldsymbol{u}_0,\boldsymbol{r}_0} + \frac{\gamma_0^2}{2}\left\langle \dot\phi(\boldsymbol{u}_0)\dot\phi(\boldsymbol{u}_0)\right\rangle_{\boldsymbol{u}_0} \odot \boldsymbol{G}^{\ell+1,2} + \frac{\gamma_0^2}{2}\frac{\partial^2}{\partial\gamma_0^2}|_{\gamma_0=0}\left\langle \left[\dot\phi(\boldsymbol{h}(\boldsymbol{u}_0,\boldsymbol{r}_0,\gamma_0))\boldsymbol{r}_0\right]\left[\dot\phi(\boldsymbol{h}(\boldsymbol{u}_0,\boldsymbol{r}_0,\gamma_0))\boldsymbol{r}_0\right]^\top \right\rangle_{\boldsymbol{u}_0,\boldsymbol{r}_0} + \mathcal{O}(\gamma_0^4)$$

This time, we see that $\boldsymbol{G}^\ell$ accumulates corrections from succeeding layers through the backward pass recursion.

P.6 Form of the Leading Corrections

We can expand the $\boldsymbol{h}^\ell$ and $\boldsymbol{z}^\ell$ fields around $\boldsymbol{u}^{\ell,0}, \boldsymbol{r}^{\ell,0}$ to find the leading order corrections to each feature kernel

$$\boldsymbol{\Phi}^{\ell,2} = \frac{1}{2}\frac{\partial^2}{\partial\gamma_0^2}|_{\gamma_0=0}\left\langle \phi(\boldsymbol{h}^\ell(\boldsymbol{u}_0,\boldsymbol{r}_0,\gamma_0))\phi(\boldsymbol{h}^\ell(\boldsymbol{u}_0,\boldsymbol{r}_0,\gamma_0))^\top\right\rangle_{\boldsymbol{u}_0,\boldsymbol{r}_0} + \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Phi}^{\ell-1,2}\left\langle \frac{\partial^2}{\partial\boldsymbol{u}_0\partial\boldsymbol{u}_0^\top}\left[\phi(\boldsymbol{u}_0)\phi(\boldsymbol{u}_0)^\top\right]\right\rangle_{\boldsymbol{u}_0}\right] \tag{160}$$

The first term requires additional expansion to extract the corrections in γ_0^2

$$\begin{aligned}
\phi(\boldsymbol{u}+\gamma_0\boldsymbol{C}^\ell\boldsymbol{g}^\ell) &\sim \phi(\boldsymbol{u}) + \gamma_0\dot\phi(\boldsymbol{u})\odot[\boldsymbol{C}^\ell\boldsymbol{g}^\ell] + \frac{\gamma_0^2}{2}\ddot\phi(\boldsymbol{u})\odot[\boldsymbol{C}^\ell\boldsymbol{g}^\ell]^2 \\
&\sim \phi(\boldsymbol{u}) + \gamma_0\dot\phi(\boldsymbol{u})\odot[\boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,0}] + \gamma_0^2\dot\phi(\boldsymbol{u})\odot[\boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,1}] + \frac{\gamma_0^2}{2}\ddot\phi(\boldsymbol{u})\odot[\boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,0}]^2 \\
\dot\phi(\boldsymbol{h}^\ell)\odot\boldsymbol{z}^\ell &\sim \dot\phi(\boldsymbol{u})\odot\boldsymbol{r} + \gamma_0\ddot\phi(\boldsymbol{u})\odot[\boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,0}]\odot\boldsymbol{r} + \gamma_0\dot\phi(\boldsymbol{u})\odot[\boldsymbol{D}^{\ell,0}\phi(\boldsymbol{u})] + \mathcal{O}(\gamma_0^2) \\
C^{\ell,0}_{\mu\alpha}(t,s) &= A^{\ell-1,1}_{\mu\alpha}(t,s) + \Theta(t-s)\Delta^0_\alpha(s)\Phi^{\ell-1,0}_{\mu\alpha}(t,s) \\
D^{\ell,0}_{\mu\alpha}(t,s) &= B^{\ell,1}_{\mu\alpha}(t,s) + \Theta(t-s)\Delta^0_\alpha(s)\Phi^{\ell-1,0}_{\mu\alpha}(t,s)
\end{aligned} \tag{161}$$

where we used the fact that $\boldsymbol{C}^{\ell,1}=0$ which follows from the fact that $\Phi^{\ell-1,1}=0$, and $\Delta^{\ell,1}=0$. Now, expanding out term by term

$$\begin{aligned}
\boldsymbol{\Phi}^\ell =& \boldsymbol{\Phi}^{\ell,0} + \gamma_0^2\left\langle[\dot\phi(\boldsymbol{u})\odot(\boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,0})][\dot\phi(\boldsymbol{u})\odot(\boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,0})]^\top\right\rangle \\
&+ \gamma_0^2\left\langle\left[\dot\phi(\boldsymbol{u})\odot(\boldsymbol{C}^{\ell,0}[\ddot\phi(\boldsymbol{u})\odot[\boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,0}]\odot\boldsymbol{r}])\right]\phi(\boldsymbol{u})^\top\right\rangle + \text{transpose} \\
&+ \gamma_0^2\left\langle\left[\dot\phi(\boldsymbol{u})\odot(\boldsymbol{C}^{\ell,0}[\dot\phi(\boldsymbol{u})\odot[\boldsymbol{D}^{\ell,0}\phi(\boldsymbol{u})]])\right]\phi(\boldsymbol{u})^\top\right\rangle + \text{transpose} \\
&+ \frac{\gamma_0^2}{2}\left\langle\left[\ddot\phi(\boldsymbol{u})\odot[\boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,0}]^2\right]\phi(\boldsymbol{u})^\top\right\rangle + \text{transpose} \\
&+ \frac{\gamma_0^2}{2}\mathrm{Tr}\left[\boldsymbol{\Phi}^{\ell-1,2}\left\langle\frac{\partial^2}{\partial\boldsymbol{u}\partial\boldsymbol{u}^\top}\left[\phi(\boldsymbol{u})\phi(\boldsymbol{u})^\top\right]\right\rangle_{\boldsymbol{u}\sim\mathcal{GP}(0,\boldsymbol{\Phi}^{\ell-1,0})}\right] + \mathcal{O}(\gamma_0^4)
\end{aligned} \tag{162}$$

We see that the corrections for the Φ^ℓ kernels accumulate on the forward pass through the final term so $\Phi^{\ell,2}\sim\mathcal{O}(\ell)$. Now we will perform the same analysis for $\boldsymbol{G}^\ell$.

$$\begin{aligned}
\boldsymbol{G}^\ell =& \left\langle\boldsymbol{g}^\ell(\boldsymbol{u},\boldsymbol{r})\boldsymbol{g}^\ell(\boldsymbol{u},\boldsymbol{r})^\top\right\rangle_{\boldsymbol{u}\sim\mathcal{GP}(0,\boldsymbol{\Phi}^{\ell-1,0})\boldsymbol{r}\sim\mathcal{GP}(0,\boldsymbol{G}^{\ell+1,0})} \\
&+ \frac{\gamma_0^2}{2}\mathrm{Tr}\left[\boldsymbol{G}^{\ell+1,2}\left\langle\frac{\partial^2}{\partial\boldsymbol{r}\partial\boldsymbol{r}^\top}\left[(\dot\phi(\boldsymbol{u})\odot\boldsymbol{r})(\dot\phi(\boldsymbol{u})\odot\boldsymbol{r})^\top\right]\right\rangle_{\boldsymbol{u}\sim\mathcal{GP}(0,\boldsymbol{\Phi}^{\ell-1,0})\boldsymbol{r}\sim\mathcal{GP}(0,\boldsymbol{G}^{\ell+1,0})}\right] + \mathcal{O}(\gamma_0^4) \\
=& \left\langle\boldsymbol{g}^\ell(\boldsymbol{u},\boldsymbol{r})\boldsymbol{g}^\ell(\boldsymbol{u},\boldsymbol{r})^\top\right\rangle_{\boldsymbol{u}\sim\mathcal{GP}(0,\boldsymbol{\Phi}^{\ell-1,0})\boldsymbol{r}\sim\mathcal{GP}(0,\boldsymbol{G}^{\ell+1,0})} \\
&+ \frac{\gamma_0^2}{2}\boldsymbol{G}^{\ell+1,2}\odot\left\langle\dot\phi(\boldsymbol{u})\dot\phi(\boldsymbol{u})\right\rangle_{\boldsymbol{u}\sim\mathcal{GP}(0,\boldsymbol{\Phi}^{\ell-1,0})} + \mathcal{O}(\gamma_0^4)
\end{aligned} \tag{163}$$

We see that, through the second term, the $\boldsymbol{G}^\ell$ kernels accumulate on the backward pass so that $\boldsymbol{G}^{\ell,2} \sim \mathcal{O}(L+1-\ell)$. As before the difficult term is the first expression which requires a full expansion of $\boldsymbol{g}^\ell$ to second order

$$\begin{aligned}\boldsymbol{g}^\ell \sim &\dot{\phi}(\boldsymbol{u}) \odot \boldsymbol{r} + \gamma_0 \dot{\phi}(\boldsymbol{u}) \odot [\boldsymbol{D}^{\ell,0}\phi(\boldsymbol{u}) + \gamma_0 \boldsymbol{D}^{\ell,0}\dot{\phi}(\boldsymbol{u})\boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,0}] \\ &+ \gamma_0 \ddot{\phi}(\boldsymbol{u})[\boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,0} + \gamma_0 \boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,1}] \odot \boldsymbol{r}\end{aligned} \tag{164}$$

From these terms we find

$$\begin{aligned}\boldsymbol{G}^\ell =& \boldsymbol{G}^{\ell,0} + \gamma_0^2 \left\langle [\dot{\phi}(\boldsymbol{u}) \odot (\boldsymbol{D}^{\ell,0}\phi(\boldsymbol{u}))][\dot{\phi}(\boldsymbol{u}) \odot (\boldsymbol{D}^{\ell,0}\phi(\boldsymbol{u}))]^\top \right\rangle \\ &+ \gamma_0^2 \left\langle [\ddot{\phi}(\boldsymbol{u})(\boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,0})][\ddot{\phi}(\boldsymbol{u})(\boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,0})]^\top \right\rangle \\ &+ \gamma_0^2 \left\langle \left[\dot{\phi}(\boldsymbol{u}) \odot \left(\boldsymbol{D}^{\ell,0}\dot{\phi}(\boldsymbol{u})\boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,0}\right)\right] \boldsymbol{g}^{\ell,0} \right\rangle + \text{transpose} \\ &+ \gamma_0^2 \left\langle [\ddot{\phi}(\boldsymbol{u}) \odot \boldsymbol{C}^{\ell,0}(\ddot{\phi}(\boldsymbol{u}) \odot \boldsymbol{C}^{\ell,0}\boldsymbol{g}^{\ell,0})]\boldsymbol{g}^{\ell,0} \right\rangle + \text{transpose} \\ &+ \frac{\gamma_0^2}{2}\boldsymbol{G}^{\ell+1,2} \odot \left\langle \dot{\phi}(\boldsymbol{u})\dot{\phi}(\boldsymbol{u}) \right\rangle_{\boldsymbol{u}\sim\mathcal{GP}(0,\boldsymbol{\Phi}^{\ell-1,0})} + \mathcal{O}(\gamma_0^4)\end{aligned} \tag{165}$$

Now the correction to the NTK has the form

$$\boldsymbol{K}^{NTK,2} = \boldsymbol{\Phi}^{L,2} + \sum_{\ell=1}^{L-1} \boldsymbol{G}^{\ell,0}\boldsymbol{\Phi}^{\ell,2} + \sum_{\ell=1}^{L-1} \boldsymbol{G}^{\ell,2}\boldsymbol{\Phi}^{\ell,0} + \boldsymbol{G}^{1,2} \odot (\boldsymbol{K}^x \otimes \boldsymbol{1}\boldsymbol{1}^\top) \tag{166}$$

Since each $\Phi^{\ell,2}, G^{L+1-\ell,2} \sim \mathcal{O}(\ell)$, each of the two sums from $\ell \in \{1, ..., L-1\}$ gives a depth scaling of the form $\sim \sum_{\ell=1}^{L-1} \ell = \frac{L(L-1)}{2}$. Since the original NTK has scale $\boldsymbol{K}^{NTK,0} \sim \mathcal{O}(L)$, the relative change in the kernel is $\frac{|\boldsymbol{K}^2|}{|\boldsymbol{K}^0|} = \mathcal{O}(\gamma_0^2 L)$. In a finite width N, network, our definition $\gamma = \gamma_0\sqrt{N}$ would indicate that a width N network would have corrections of scale $\gamma_0^2 L = \frac{\gamma^2 L}{N}$ in the NTK regime where $\gamma = \mathcal{O}_N(1)$ provided the network is sufficiently wide to disregard initialization dependent fluctuations in the kernels.

P.7 Perturbation Theory in N

Finite size corrections to the DMFT can also be obtained within our field theoretic framework. Let $\boldsymbol{k} = \text{Vec}\{\boldsymbol{\Phi}^\ell, \hat{\boldsymbol{\Phi}}^\ell, \boldsymbol{G}^\ell, \hat{\boldsymbol{G}}^\ell, \boldsymbol{A}^\ell, \boldsymbol{B}^\ell\}$ denote the collection of kernel order parameters of the DMFT. The partition function with a kernel source term can be written as

$$Z[\boldsymbol{j}_k] = \int d\boldsymbol{k} \exp\left(NS[\boldsymbol{k}] + \boldsymbol{j}_k \cdot \boldsymbol{k}\right) \tag{167}$$

The infinite-width DMFT is characterized by the set of saddle point equations which are $\nabla_{\boldsymbol{k}} S[\boldsymbol{k}]|_{\boldsymbol{k}=\boldsymbol{k}^*} = 0$. Let the saddle point be $\boldsymbol{k}^*$. To identify corrections to the kernels due to finite size, we now expand S to second order

$$S[\boldsymbol{k}] \approx S[\boldsymbol{k}^*] + \frac{1}{2}(\boldsymbol{k}-\boldsymbol{k}^*)\nabla_{\boldsymbol{k}}^2 S[\boldsymbol{k}]|_{\boldsymbol{k}=\boldsymbol{k}^*}(\boldsymbol{k}-\boldsymbol{k}^*) + ... \tag{168}$$

The linear component vanishes at the saddle point since $\nabla_{\boldsymbol{k}} S[\boldsymbol{k}]|_{\boldsymbol{k}=\boldsymbol{k}^*} = 0$. It is also easy to verify that $S[\boldsymbol{k}^*] = 0$, which gives appropriate normalization in the $N \to \infty$ limit. Thus the moment generating function for the kernels has the form

$$Z[\boldsymbol{j}] \sim \int d\boldsymbol{k} \exp\left(\frac{N}{2}(\boldsymbol{k}-\boldsymbol{k}^*)\nabla_{\boldsymbol{k}}^2 S[\boldsymbol{k}]|_{\boldsymbol{k}=\boldsymbol{k}^*}(\boldsymbol{k}-\boldsymbol{k}^*) + \boldsymbol{j}_k \cdot \boldsymbol{k}\right) \tag{169}$$

This suggests that the leading order correction of the kernel distributions can be approximated as mean zero Gaussian distribution determined by the saddle point solution $\boldsymbol{k}^*$ and covariance given by $\frac{1}{N}\left[-\nabla_{\boldsymbol{k}}^2 S[\boldsymbol{k}]|_{\boldsymbol{k}=\boldsymbol{k}^*}\right]^{-1}$. We derive expressions for the components of this Hessian in P.7.1. These fluctuations have standard deviation $\mathcal{O}(N^{-1/2})$. This technique is a common approach to identifying finite size effects [56] and was recently employed in Bayesian inference setting [29].

Before computing the Hessian terms, we can compare finite size effects under NTK scaling $\gamma = \mathcal{O}_N(1)$, and the mean field scaling $\gamma = \mathcal{O}(\sqrt{N})$. Concretely, we are interested in the feature learning component of the kernel change which is $\mathcal{O}(\frac{\gamma^2}{N})$. Let $\langle \Delta \boldsymbol{k} \rangle$ represent the change in the kernel through training, which we showed in P.2 is of size $\mathcal{O}(\gamma_0^2) = \mathcal{O}\left(\frac{\gamma^2}{N}\right)$. We will now define the signal to noise ratio of feature learning as

$$\text{SNR} = \frac{\langle \Delta \boldsymbol{k} \rangle}{\sqrt{\text{Var}(\boldsymbol{k})}} = \mathcal{O}\left(\frac{\gamma^2}{\sqrt{N}}\right) \tag{170}$$

For NTK regime, this is vanishing as $N \to \infty$, while for the DMFT regime, this is goes as $\mathcal{O}(\sqrt{N})$ since kernel evolution is always $\mathcal{O}(1)$ but variance is $\mathcal{O}(N^{-1})$.

P.7.1 DMFT Action Hessian

We now compute the various blocks of the Hessian of the DMFT action $-\nabla^2_{\boldsymbol{k}} S[\boldsymbol{k}^*]$. The various necessary derivatives are given below. We will utilize the the factorization of the single site MGF $\sum_\ell \ln \mathcal{Z}_\ell$ to simplify many of the expressions. Below we provide exhaustive expressions for each of the relevant terms. These expressions are included for completeness but we did not yet attempt computing all of them as we did the saddle point equations which define $\boldsymbol{k}^*$. First, we list the collection of Hessian terms which do not involve A, B below.

$$\begin{aligned}
-\frac{\partial^2 S}{\partial \hat{\Phi}^\ell_{\mu\nu}(t,s) \partial \hat{\Phi}^{\ell'}_{\alpha\beta}(t',s')} &= \delta_{\ell\ell'} \left[\left\langle \phi(h^\ell_\mu(t)) \phi(h^\ell_\nu(s)) \phi(h^\ell_\alpha(t')) \phi(h^\ell_\beta(s')) \right\rangle - \Phi^\ell_{\mu\nu}(t,s) \Phi^\ell_{\alpha\beta}(t',s') \right] \\
-\frac{\partial^2 S}{\partial \Phi^\ell_{\mu\nu}(t,s) \partial \Phi^{\ell'}_{\alpha\beta}(t',s')} &= 0 \\
-\frac{\partial^2 S}{\partial \hat{\Phi}^\ell_{\mu\nu}(t,s) \partial \Phi^{\ell'}_{\alpha\beta}(t',s')} &= -\delta_{\ell\ell'} \delta_{\mu\alpha} \delta_{\nu\beta} \delta(t-t') \delta(s-s') + \frac{\partial}{\partial \Phi^{\ell'}_{\alpha\beta}(t',s')} \Phi^\ell_{\mu\nu}(t,s) \\
-\frac{\partial^2 S}{\partial \hat{G}^\ell_{\mu\nu}(t,s) \partial \hat{G}^{\ell'}_{\alpha\beta}(t',s')} &= \delta_{\ell,\ell'} \left[\left\langle g^\ell_\mu(t) g^\ell_\nu(s) g^\ell_\alpha(t') g^\ell_\beta(s') \right\rangle - G^\ell_{\mu\nu}(t,s) G^\ell_{\alpha\beta}(t',s') \right] \\
-\frac{\partial^2 S}{\partial G^\ell_{\mu\nu}(t,s) \partial G^{\ell'}_{\alpha\beta}(t',s')} &= 0 \\
-\frac{\partial^2 S}{\partial \hat{G}^\ell_{\mu\nu}(t,s) \partial G^{\ell'}_{\alpha\beta}(t',s')} &= -\delta_{\ell\ell'} \delta_{\mu\alpha} \delta_{\nu\beta} \delta(t-t') \delta(s-s') + \frac{\partial}{\partial G^{\ell'}_{\alpha\beta}(t',s')} G^\ell_{\mu\nu}(t,s) \\
-\frac{\partial^2 S}{\partial \hat{\Phi}^\ell_{\mu\nu}(t,s) \partial \hat{G}^{\ell'}_{\alpha\beta}(t',s')} &= \delta_{\ell,\ell'} \left[\left\langle \phi(h^\ell_\mu(t)) \phi(h^\ell_\nu(s)) g^\ell_\alpha(t') g^\ell_\beta(s') \right\rangle - \Phi^\ell_{\mu\nu}(t,s) G^\ell_{\alpha\beta}(t',s') \right] \\
-\frac{\partial^2 S}{\partial \hat{\Phi}^\ell_{\mu\nu}(t,s) \partial G^\ell_{\alpha\beta}(t',s')} &= \frac{\partial}{\partial G^\ell_{\alpha\beta}(t',s')} \Phi^\ell_{\mu\nu}(t,s) \\
-\frac{\partial^2 S}{\partial \hat{G}^\ell_{\mu\nu}(t,s) \partial \Phi^\ell_{\alpha\beta}(t',s')} &= \frac{\partial}{\partial \Phi^\ell_{\alpha\beta}(t',s')} G^\ell_{\mu\nu}(t,s) \\
-\frac{\partial^2 S}{\partial G^\ell_{\mu\nu}(t,s) \partial \Phi^\ell_{\alpha\beta}(t',s')} &= 0
\end{aligned}$$

Now for the terms involving A^ℓ, B^ℓ, we find the following expressions.

$$
\begin{aligned}
-\frac{\partial^2 S}{\partial A^\ell_{\mu\nu}(t,s)\partial A^{\ell'}_{\alpha\beta}(t',s')} &= \left\langle \frac{\partial^2}{\partial u^{\ell+1}_\nu(s)\partial u^{\ell'+1}_\beta(s')}[g^{\ell+1}_\mu(t)g^{\ell'+1}_\alpha(t')] \right\rangle - B^\ell_{\mu\nu}(t,s)B^\ell_{\alpha\beta}(t',s') \\
-\frac{\partial^2 S}{\partial B^\ell_{\mu\nu}(t,s)\partial B^{\ell'}_{\alpha\beta}(t',s')} &= \left\langle \frac{\partial^2}{\partial r^\ell_\nu(s)\partial r^{\ell'}_\beta(s')}[\phi(h^\ell_\mu(t))\phi(h^{\ell'}_\alpha(t'))] \right\rangle \\
-\frac{\partial^2 S}{\partial A^\ell_{\mu\nu}(t,s)\partial B^{\ell'}_{\alpha\beta}(t',s')} &= \delta_{\ell\ell'}\delta_{\mu\alpha}\delta_{\nu\beta}\delta(t-t')\delta(s-s') \\
&\quad + \left\langle \frac{\partial^2}{\partial u^{\ell+1}_\nu(s)\partial r^{\ell'}_\beta(s')}[g^{\ell+1}_\mu(t)\phi(h^{\ell'}_\alpha(s))] \right\rangle - B^\ell_{\mu\nu}(t,s)A^{\ell'}_{\alpha\beta}(t',s') \\
-\frac{\partial^2 S}{\partial A^\ell_{\mu\nu}(t,s)\partial \Phi^{\ell'}_{\alpha\beta}(t',s')} &= \delta_{\ell,\ell'}\left\langle \frac{\partial^3}{\partial u^{\ell+1}_\nu(t)\partial u^{\ell+1}_\alpha(t')\partial u^{\ell+1}_\beta(s')} g^{\ell+1}_\mu(t) \right\rangle \\
-\frac{\partial^2 S}{\partial A^\ell_{\mu\nu}(t,s)\partial \hat{\Phi}^{\ell'}_{\alpha\beta}(t',s')} &= \delta_{\ell+1,\ell'}\left[\left\langle \frac{\partial}{\partial u^{\ell+1}_\nu(t)}[g^{\ell+1}_\mu(t)\phi(h^{\ell+1}_\alpha(t'))\phi(h^{\ell+1}_\beta(s'))] \right\rangle - B^\ell_{\mu\nu}(t,s)\Phi^{\ell+1}_{\alpha\beta}(t',s')\right] \\
-\frac{\partial^2 S}{\partial A^\ell_{\mu\nu}(t,s)\partial G^{\ell'}_{\alpha\beta}(t',s')} &= \delta_{\ell+2,\ell'}\left\langle \frac{\partial^3}{\partial u^{\ell+1}_\nu(t)\partial r^{\ell+1}_\alpha(t')\partial r^{\ell+1}_\beta(s')} g^{\ell+1}_\mu(t) \right\rangle \\
-\frac{\partial^2 S}{\partial A^\ell_{\mu\nu}(t,s)\partial \hat{G}^{\ell'}_{\alpha\beta}(t',s')} &= \delta_{\ell+1,\ell'}\left[\left\langle \frac{\partial}{\partial u^{\ell+1}_\nu(t)}[g^{\ell+1}_\mu(t)g^{\ell+1}_\alpha(t')g^{\ell+1}_\beta(s')] \right\rangle - B^\ell_{\mu\nu}(t,s)G^{\ell+1}_{\alpha\beta}(t',s')\right] \\
-\frac{\partial^2 S}{\partial B^\ell_{\mu\nu}(t,s)\partial \Phi^{\ell'}_{\alpha\beta}(t',s')} &= \delta_{\ell-1,\ell'}\left\langle \frac{\partial^3}{\partial r^\ell_\nu(t)\partial u^\ell_\alpha(t')\partial u^\ell_\beta(s')} \phi(h^\ell_\mu(t)) \right\rangle \\
-\frac{\partial^2 S}{\partial B^\ell_{\mu\nu}(t,s)\partial \hat{\Phi}^{\ell'}_{\alpha\beta}(t',s')} &= \delta_{\ell,\ell'}\left[\left\langle \frac{\partial}{\partial r^\ell_\nu(t)}[\phi(h^\ell_\mu(t))\phi(h^\ell_\alpha(t'))\phi(h^\ell_\beta(s'))] \right\rangle - A^\ell_{\mu\nu}(t,s)\Phi^\ell(t',s')\right] \\
-\frac{\partial^2 S}{\partial B^\ell_{\mu\nu}(t,s)\partial G^{\ell'}_{\alpha\beta}(t',s')} &= \delta_{\ell+1,\ell'}\left\langle \frac{\partial^3}{\partial r^\ell_\nu(t)\partial r^\ell_\alpha(t')\partial r^\ell_\beta(s')} \phi(h^\ell_\mu(t)) \right\rangle \\
-\frac{\partial^2 S}{\partial B^\ell_{\mu\nu}(t,s)\partial \hat{G}^{\ell'}_{\alpha\beta}(t',s')} &= \delta_{\ell,\ell'}\left[\left\langle \frac{\partial}{\partial r^\ell_\nu(t)}[\phi(h^\ell_\mu(t))g^\ell_\alpha(t')g^\ell_\beta(s')] \right\rangle - A^\ell_{\mu\nu}(t,s)G^\ell_{\alpha\beta}(t',s')\right]
\end{aligned}
$$

From these block matrices which comprise the Hessian, we can obtain the finite width covariance structure in our order parameters $\boldsymbol{k} = \text{Vec}\{\Phi^\ell, G^\ell, A^\ell, B^\ell\}$ by computing the inverse $\boldsymbol{C} = \left(-\nabla^2_{\boldsymbol{k}} S[\boldsymbol{k}^*]\right)^{-1}$. In this approximation scheme, we have $\boldsymbol{k} \sim \mathcal{N}(\boldsymbol{k}^*, \frac{1}{N}\left(-\nabla^2_{\boldsymbol{k}} S[\boldsymbol{k}^*]\right)^{-1})$. Many questions about this expansion remain including: can it be proven that these $\mathcal{O}(N^{-1/2})$ fluctuations always lead to higher expected test loss?