Structures of neural network effective theories

Ian Banta,[1] Tianji Cai,[1,2] Nathaniel Craig,[1,3] and Zhengkang Zhang[1,3,4]

[1]*Department of Physics, University of California, Santa Barbara, California 93106, USA*
[2]*SLAC National Accelerator Laboratory, Stanford University, Stanford, California 94309, USA*
[3]*Kavli Institute for Theoretical Physics, University of California, Santa Barbara, California 93106, USA*
[4]*Department of Physics and Astronomy, University of Utah, Salt Lake City, Utah 84112, USA*

(Received 24 August 2023; accepted 28 March 2024; published 7 May 2024)

We develop a diagrammatic approach to effective field theories (EFTs) corresponding to deep neural networks at initialization, which dramatically simplifies computations of finite-width corrections to neuron statistics. The structures of EFT calculations make it transparent that a single condition governs criticality of all connected correlators of neuron preactivations. Understanding of such EFTs may facilitate progress in both deep learning and field theory simulations.

DOI: 10.1103/PhysRevD.109.105007

I. INTRODUCTION

Machine learning (ML) has undergone a revolution in recent years, with applications ranging from image recognition and natural language processing, to self-driving cars and playing Go. Central to all these developments is the engineering of deep neural networks, a class of ML architectures consisting of multiple layers of artificial neurons. Such networks are apparently rather complex, with a deterring number of trainable parameters, which means practical applications have often been guided by expensive trial and error. Nevertheless, extensive research is underway toward opening the black box.

That a theoretical understanding of such complex systems is possible has to do with the observation that a wide range of neural network architectures actually admit a simple limit: they reduce to Gaussian processes when the network width (number of neurons per layer) goes to infinity [1–6], and evolve under gradient-based training as linear models governed by the neural tangent kernel [7–9]. However, an infinitely-wide network neither exists in practice, nor provides an accurate model for deep learning. It is therefore crucial to understand finite-width effects, which have recently been studied by a variety of methods [10–23].

This line of research in ML theory has an intriguing synergy with theoretical physics [24]. In particular, it has been realized that neural networks have a natural correspondence with (statistical or quantum) field theories [25–34]. Infinite-width networks—which are Gaussian processes—correspond to free theories, while finite-width corrections in wide networks can be calculated perturbatively as in weakly interacting theories. This allows for a systematically improvable characterization of neural networks beyond the (very few) exactly-solvable special cases [35–37]. Meanwhile, from an effective theory perspective [21], information propagation through a deep neural network can be understood as a renormalization group (RG) flow. Examining scaling behaviors near RG fixed points reveals strategies to tune the network to criticality [38–40], which is crucial for mitigating the notorious exploding and vanishing gradient problems in practical applications. In the reverse direction, this synergy also points to new opportunities to study field theories with neural networks [33].

Inspired by recent progress, in this paper we further explore the structures of effective field theories (EFTs) corresponding to archetypical deep neural networks. To this end, we develop a novel diagrammatic formalism.[1] Our approach largely builds on the frameworks of Refs. [21,22], which enable systematic calculations of finite-width corrections. The diagrammatic formalism dramatically simplifies these calculations, as we demonstrate by concisely reproducing known results in the main text and presenting further examples with new results in the Supplemental Material [42]. Interestingly, the structures of diagrams in the RG analysis suggest that neural network EFTs are of a quite special type, where a single condition governs the critical tuning of all neuron correlators. The study of these EFTs may lend new insights into both neural network properties and novel field-theoretic phenomena.

[1]See also Refs. [13,17,18,26,27,31,32,41] for Feynman diagram-inspired approaches to ML.

II. EFT OF DEEP NEURAL NETWORKS

The archetype of deep neural networks, the multilayer perceptron, can be defined by a collection of neurons whose values $\phi_i^{(\ell)}$ (called preactivations) are determined by the following operations given an input $\vec{x} \in \mathbb{R}^{n_0}$:

$$\phi_i^{(1)}(\vec{x}) = \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j + b_i^{(1)},$$
$$\phi_i^{(\ell)}(\vec{x}) = \sum_{j=1}^{n_{\ell-1}} W_{ij}^{(\ell)} \sigma(\phi_j^{(\ell-1)}(\vec{x})) + b_i^{(\ell)} \quad (\ell \geq 2). \tag{1}$$

Here superscripts in parentheses label layers, subscripts i, j label neurons within a layer (of which there are n_ℓ at the ℓth layer), and $\sigma(\phi)$ is the activation function [common choices include $\tanh(\phi)$ or $\mathrm{ReLU}(\phi) \equiv \max(0, \phi)$]. The weights $W_{ij}^{(\ell)}$ and biases $b_i^{(\ell)}$ ($\ell = 1, \ldots, L$) are the network parameters which are adjusted to minimize a loss function during training, such that the trained network can approximate the desired function.

The basic idea of an EFT of deep neural networks is to consider an ensemble of networks, where at initialization, each of the ℓ th-layer weights $W_{ij}^{(\ell)}$ (biases $b_i^{(\ell)}$) is drawn independently from a Gaussian distribution with mean zero and variance $C_W^{(\ell)}/n_{\ell-1}$ ($C_b^{(\ell)}$). The statistics of this ensemble encode both the typical behavior of neural networks initialized in this manner and how a particular network may fluctuate away from typicality. In the field theory language, these are captured by a Euclidean action, $\mathcal{S}[\phi] = -\log P(\phi)$, for all neuron preactivation fields $\phi_i^{(\ell)}(\vec{x})$, where $P(\phi)$ is the joint probability distribution. As we review in the Supplemental Material [42], at initialization the conditional probability distribution at each layer is Gaussian:

$$P(\phi^{(\ell)}|\phi^{(\ell-1)}) = [\det(2\pi\mathcal{G}^{(\ell)})]^{-\frac{n_\ell}{2}} e^{-\mathcal{S}_0^{(\ell)}}, \tag{2}$$

$$\mathcal{S}_0^{(\ell)} = \int d\vec{x}_1 d\vec{x}_2 \frac{1}{2} \sum_{i=1}^{n_\ell} \phi_i^{(\ell)}(\vec{x}_1)(\mathcal{G}^{(\ell)})^{-1}(\vec{x}_1, \vec{x}_2)\phi_i^{(\ell)}(\vec{x}_2), \tag{3}$$

where $\mathcal{G}^{(\ell)}(\vec{x}_1, \vec{x}_2) = \frac{1}{n_{\ell-1}} \sum_{j=1}^{n_{\ell-1}} \mathcal{G}_j^{(\ell)}(\vec{x}_1, \vec{x}_2)$, with

$$\mathcal{G}_j^{(\ell)}(\vec{x}_1, \vec{x}_2) = C_b^{(\ell)} + C_W^{(\ell)} \sigma(\phi_j^{(\ell-1)}(\vec{x}_1))\sigma(\phi_j^{(\ell-1)}(\vec{x}_2)) \equiv C_b^{(\ell)} + C_W^{(\ell)} \sigma_{j,\vec{x}_1}^{(\ell-1)} \sigma_{j,\vec{x}_2}^{(\ell-1)} \tag{4}$$

for $\ell \geq 2$, and $\mathcal{G}_j^{(1)}(\vec{x}_1, \vec{x}_2) = C_b^{(1)} + C_W^{(1)} x_{1j} x_{2j}$. We have taken the continuum limit in input space to better parallel field theory analyses. $(\mathcal{G}^{(\ell)})^{-1}$ is understood as the pseudoinverse when $\mathcal{G}^{(\ell)}$ is not invertible. We see that for $\ell \geq 2$, $\mathcal{G}^{(\ell)}(\vec{x}_1, \vec{x}_2)$ is an operator of the $(\ell-1)$th-layer neurons, so Eq. (3) is actually an interacting theory with interlayer couplings. This also means the determinant in Eq. (2) is not a constant prefactor. To account for its effect, we introduce auxiliary anticommuting fields ψ, $\bar{\psi}$ which are analogs of ghosts and antighosts in the Faddeev-Popov procedure. Including all layers, we have

$$e^{-\mathcal{S}[\phi]} = \int \mathcal{D}\psi \mathcal{D}\bar{\psi} e^{-\sum_{\ell=1}^{L} (\mathcal{S}_0^{(\ell)}[\phi] + \mathcal{S}_\psi^{(\ell)}[\phi, \psi, \bar{\psi}])}, \tag{5}$$

where $\mathcal{S}_0^{(\ell)}$ is given by Eq. (3) above and

$$\mathcal{S}_\psi^{(\ell)} = -\int d\vec{x}_1 d\vec{x}_2 \sum_{i'=1}^{n_\ell/2} \bar{\psi}_{i'}^{(\ell)}(\vec{x}_1)(\mathcal{G}^{(\ell)})^{-1}(\vec{x}_1, \vec{x}_2)\psi_{i'}^{(\ell)}(\vec{x}_2). \tag{6}$$

The ℓth-layer neurons interact with the $(\ell-1)$th-layer and $(\ell+1)$th-layer neurons via $\mathcal{S}_0^{(\ell)}$ and $\mathcal{S}_0^{(\ell+1)}$, respectively, while their associated ghosts have opposite-sign couplings to the $(\ell-1)$th-layer neurons but do not couple to $(\ell+1)$ th-layer neurons. This means $\phi^{(\ell)}$ and $\psi^{(\ell)}$ loops cancel as far as their couplings to $\phi^{(\ell-1)}$ are concerned, which must be the case since the network has directionality—neurons at a given layer cannot be affected by what happens at deeper layers.

III. NEURON STATISTICS FROM FEYNMAN DIAGRAMS

We are interested in calculating neuron statistics, i.e., connected correlators of neuron preactivation fields $\phi_i^{(\ell)}(\vec{x})$ in the EFT above. More precisely, we would like to track the evolution of neuron correlators as a function of network layer ℓ, which encodes how information is processed through a deep neural network and has an analogous form to RG flows in field theory. To this end, we develop an efficient diagrammatic framework to recursively determine ℓ th-layer neuron correlators in terms of $(\ell-1)$th-layer neuron correlators.

To derive the Feynman rules, we first note that if $\phi_j^{(\ell-1)}(\vec{x})$ were classical background fields with no statistical fluctuations, we would simply have a free theory for $\phi_i^{(\ell)}(\vec{x})$ with propagator $\mathcal{G}^{(\ell)}(\vec{x}_1, \vec{x}_2)$. In this case, the two-point correlator is given by

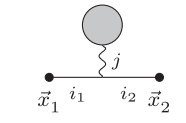
$$\langle \phi_{i_1}^{(\ell)}(\vec{x}_1)\phi_{i_2}^{(\ell)}(\vec{x}_2)\rangle \overset{?}{=} \delta_{i_1 i_2}\mathcal{G}^{(\ell)}(\vec{x}_1, \vec{x}_2) = \delta_{i_1 i_2} \frac{1}{n_{\ell-1}} \sum_j \mathcal{G}_j^{(\ell)}(\vec{x}_1, \vec{x}_2), \tag{7}$$

and, by simple Wick contraction, the four-point correlator is given by

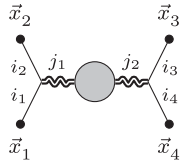
$$\begin{aligned}&\langle \phi_{i_1}^{(\ell)}(\vec{x}_1)\phi_{i_2}^{(\ell)}(\vec{x}_2)\phi_{i_3}^{(\ell)}(\vec{x}_3)\phi_{i_4}^{(\ell)}(\vec{x}_4)\rangle \\ &\overset{?}{=} \delta_{i_1 i_2}\delta_{i_3 i_4}\mathcal{G}^{(\ell)}(\vec{x}_1,\vec{x}_2)\mathcal{G}^{(\ell)}(\vec{x}_3,\vec{x}_4) + \text{perms} \\ &= \delta_{i_1 i_2}\delta_{i_3 i_4}\frac{1}{n_{\ell-1}^2}\sum_{j_1,j_2}\mathcal{G}_{j_1}^{(\ell)}(\vec{x}_1,\vec{x}_2)\mathcal{G}_{j_2}^{(\ell)}(\vec{x}_3,\vec{x}_4) + \text{perms}. \end{aligned} \tag{8}$$

We have used "$\overset{?}{=}$" to indicate that these equations may not hold when statistical fluctuations of $\phi_j^{(\ell-1)}(\vec{x})$ are taken into account. Indeed, since $\phi_j^{(\ell-1)}(\vec{x})$ are integrated over in the path integral, we should take the ensemble average of the expressions on the right-hand sides of Eqs. (7) and (8). This can be represented diagrammatically as:

$$\begin{aligned}\left\langle \phi_{i_1}^{(\ell)}(\vec{x}_1)\,\phi_{i_2}^{(\ell)}(\vec{x}_2)\right\rangle &= \delta_{i_1 i_2}\frac{1}{n_{\ell-1}}\sum_j \left\langle \mathcal{G}_j^{(\ell)}(\vec{x}_1,\vec{x}_2)\right\rangle \\ &= \sum_j \quad ,\end{aligned} \tag{9}$$



$$\begin{aligned}&\left\langle \phi_{i_1}^{(\ell)}(\vec{x}_1)\,\phi_{i_2}^{(\ell)}(\vec{x}_2)\,\phi_{i_3}^{(\ell)}(\vec{x}_3)\,\phi_{i_4}^{(\ell)}(\vec{x}_4)\right\rangle \\ &= \delta_{i_1 i_2}\delta_{i_3 i_4}\frac{1}{n_{\ell-1}^2}\left\langle \mathcal{G}_{j_1}^{(\ell)}(\vec{x}_1,\vec{x}_2)\,\mathcal{G}_{j_2}^{(\ell)}(\vec{x}_3,\vec{x}_4)\right\rangle + \text{perms}. \\ &= \sum_{j_1,j_2} \quad + \text{perms}.\end{aligned} \tag{10}$$



In our diagrammatic notation, all external fields are ℓ th-layer preactivations $\phi_i^{(\ell)}(\vec{x})$ since we are interested in calculating their correlators. A double wavy internal line labeled by j and connected to external points $\vec{x}_1, \vec{x}_2$ represents $\mathcal{G}_j^{(\ell)}(\vec{x}_1,\vec{x}_2)$ [an operator made of the $(\ell-1)$ th-layer fields $\phi_j^{(\ell-1)}(\vec{x}_1), \phi_j^{(\ell-1)}(\vec{x}_2)$], and a blob means taking the expectation value of the product of all operators attached to it. Each $\phi^2\mathcal{G}$ vertex comes with a factor of $\frac{1}{n_{\ell-1}}$, and external $\phi_i^{(\ell)}$ fields meeting at the same vertex share the same neuron index. We can similarly write down higher-point correlators in this diagrammatic notation.

Since we are mostly interested in connected correlators, it is convenient to decompose $\mathcal{G}_j^{(\ell)}(\vec{x}_1,\vec{x}_2)$ into an expectation value piece and a fluctuating piece:

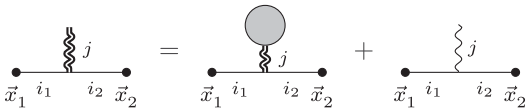
$$\mathcal{G}_j^{(\ell)}(\vec{x}_1,\vec{x}_2) = \langle \mathcal{G}_j^{(\ell)}(\vec{x}_1,\vec{x}_2)\rangle + C_W^{(\ell)}\Delta_j^{(\ell-1)}(\vec{x}_1,\vec{x}_2), \tag{11}$$

where

$$\Delta_j^{(\ell-1)}(\vec{x}_1,\vec{x}_2) \equiv \sigma_{j,\vec{x}_1}^{(\ell-1)}\sigma_{j,\vec{x}_2}^{(\ell-1)} - \langle \sigma_{j,\vec{x}_1}^{(\ell-1)}\sigma_{j,\vec{x}_2}^{(\ell-1)}\rangle \tag{12}$$

for $\ell \geq 2$, and $\Delta_j^{(0)}(\vec{x}_1,\vec{x}_2) = 0$. Diagrammatically, we can represent this as:

$$= \quad + \quad . \tag{13}$$

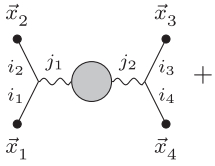


A single wavy internal line labeled by j and connected to external points $\vec{x}_1, \vec{x}_2$ represents $\Delta_j^{(\ell-1)}(\vec{x}_1,\vec{x}_2)$ [an operator made of $\phi_j^{(\ell-1)}(\vec{x}_1), \phi_j^{(\ell-1)}(\vec{x}_2)$], and the $\phi^2\Delta$ vertex represents a factor of $\frac{C_W^{(\ell)}}{n_{\ell-1}}$. By definition, the fluctuation operators are tadpole-free:

$$= \delta_{i_1 i_2}\frac{C_W^{(\ell)}}{n_{\ell-1}}\left\langle \Delta_j^{(\ell-1)}(\vec{x}_1,\vec{x}_2)\right\rangle = 0\,. \tag{14}$$

The diagrammatic representation Eq. (13) of the decomposition Eq. (11) makes it straightforward to extract the connected part of neuron correlators. Starting from a full correlator like Eq. (10), we separate each double wavy line according to Eq. (13). Taking the first term on the right-hand side of Eq. (13) for any double wavy line would disconnect the diagram. Therefore, we can replace each double wavy line by a single wavy line (representing $\Delta_j^{(\ell-1)}$) when calculating connected correlators. For example, the connected four-point correlator is given by:

$$\begin{aligned}&\left\langle \phi_{i_1}^{(\ell)}(\vec{x}_1)\,\phi_{i_2}^{(\ell)}(\vec{x}_2)\,\phi_{i_3}^{(\ell)}(\vec{x}_3)\,\phi_{i_4}^{(\ell)}(\vec{x}_4)\right\rangle_{\mathrm{C}} \\ &= \sum_{j_1,j_2} \quad + \text{perms}. \\ &= \delta_{i_1 i_2}\delta_{i_3 i_4}\frac{\left(C_W^{(\ell)}\right)^2}{n_{\ell-1}^2}\left\langle \Delta_{j_1}^{(\ell-1)}(\vec{x}_1,\vec{x}_2)\,\Delta_{j_2}^{(\ell-1)}(\vec{x}_3,\vec{x}_4)\right\rangle \\ &\quad + \text{perms}.\end{aligned} \tag{15}$$



Note that the blob representing $\langle \Delta_{j_1}^{(\ell-1)}(\vec{x}_1,\vec{x}_2) \times \Delta_{j_2}^{(\ell-1)}(\vec{x}_3,\vec{x}_4)\rangle$ is automatically connected due to Eq. (14). The only exception to this double → single wavy line replacement rule is when calculating the two-point correlator in Eq. (9), which by definition involves the expectation value piece [first term on the right-hand side of Eq. (13)] but is trivially connected (note that since we have normalized $\int \mathcal{D}\phi e^{-\mathcal{S}} = 1$, disconnected diagrams involving vacuum bubbles sum to zero).

From the discussion above it is clear that each $\phi^2\Delta$ vertex comes with a factor of $\frac{1}{n}$ (where n collectively denotes $n_1, \ldots, n_{L-1}$). In the infinite-width limit, $n \to \infty$, the EFT is a free theory, whereas for large but finite n, we have a weakly-interacting theory where higher-point connected correlators can be perturbatively calculated as a $\frac{1}{n}$ expansion. Meanwhile, each sum over internal neuron indices (like $\sum_j$ in Eq. (9) above) gives rise to a factor of n. We will see below that no $\frac{n}{n}$ terms arise, so we have a valid perturbative expansion.

To see how this $\frac{1}{n}$ expansion works, let us first take a closer look at the two-point correlator Eq. (9). We can expand it as:

$$\langle \mathcal{G}^{(\ell)}(\vec{x}_1, \vec{x}_2)\rangle = \sum_{p=0}^{\infty} \frac{1}{n_{\ell-1}^{p}} \mathcal{K}_p^{(\ell)}(\vec{x}_1, \vec{x}_2). \tag{16}$$

The leading-order (LO) term $\mathcal{K}_0^{(\ell)}$ is known as the kernel; it is the propagator for $\phi_i^{(\ell)}$ in the free-theory limit $n \to \infty$. Evaluating $\langle \mathcal{G}^{(\ell)}(\vec{x}_1, \vec{x}_2)\rangle$ in this limit amounts to using free-theory propagators $\mathcal{K}_0^{(\ell-1)}$ for the previous-layer neurons $\phi_j^{(\ell-1)}$ in the blob in Eq. (13):

$$\begin{aligned}\mathcal{K}_0^{(\ell)}(\vec{x}_1, \vec{x}_2) &= \sum_j \frac{1}{n_{\ell-1}} \langle \mathcal{G}_j^{(\ell)}(\vec{x}_1, \vec{x}_2)\rangle_{\mathcal{K}_0^{(\ell-1)}} \\ &= C_b^{(\ell)} + C_W^{(\ell)} \langle \sigma_{\vec{x}_1}\sigma_{\vec{x}_2}\rangle_{\mathcal{K}_0^{(\ell-1)}}.\end{aligned} \tag{17}$$

Here subscript $\mathcal{K}_0^{(\ell-1)}$ means the expectation value is computed with the free-theory propagator $\mathcal{K}_0^{(\ell-1)}$, namely:

$$\langle \mathcal{O}(\phi_i^{(\ell-1)}(\vec{x}_1), \phi_i^{(\ell-1)}(\vec{x}_2), \ldots)\rangle_{\mathcal{K}_0^{(\ell-1)}} \equiv \frac{\int \mathcal{D}\phi \mathcal{O}(\phi(\vec{x}_1), \phi(\vec{x}_2), \ldots) e^{-\int d\vec{y}_1 d\vec{y}_2 \frac{1}{2}\phi(\vec{y}_1)(\mathcal{K}_0^{(\ell-1)})^{-1}(\vec{y}_1,\vec{y}_2)\phi(\vec{y}_2)}}{\int \mathcal{D}\phi e^{-\int d\vec{y}_1 d\vec{y}_2 \frac{1}{2}\phi(\vec{y}_1)(\mathcal{K}_0^{(\ell-1)})^{-1}(\vec{y}_1,\vec{y}_2)\phi(\vec{y}_2)}}. \tag{18}$$

On the right-hand side of Eq. (17), we have dropped both neuron and layer indices on σ because the $\mathcal{K}_0^{(\ell-1)}$ subscript already indicates the layer, and the expectation value is identical for all neurons in that layer. One can further evaluate $\langle \sigma_{\vec{x}_1}\sigma_{\vec{x}_2}\rangle_{\mathcal{K}_0^{(\ell-1)}}$ for specific choices of activation functions σ, but we stay activation-agnostic for the present analysis.

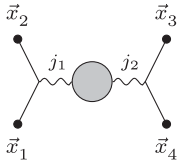
Equation (17) allows us to recursively determine $\mathcal{K}_0^{(\ell)}$ from $\mathcal{K}_0^{(\ell-1)}$, and has been well known from studies of infinite-width networks. It may also be viewed as the RG flow of $\mathcal{K}_0$, with ultraviolet boundary condition $\mathcal{K}_0^{(1)}(\vec{x}_1, \vec{x}_2) = C_b^{(1)} + \frac{C_W^{(1)}}{n_0}\vec{x}_1 \cdot \vec{x}_2$. It is straightforward to extend the diagrammatic calculation to $\mathcal{K}_{p\geq 1}$. We present a simple derivation of the RG flow of $\mathcal{K}_1$ in the Supplemental Material [42].

Next, consider the connected four-point correlator in Eq. (15). Following Ref. [21], we separate the neuron index structures and define the vertex function $V_4^{(\ell)}$ as follows:

$$\begin{aligned}&\langle \phi_{i_1}^{(\ell)}(\vec{x}_1)\phi_{i_2}^{(\ell)}(\vec{x}_2)\phi_{i_3}^{(\ell)}(\vec{x}_3)\phi_{i_4}^{(\ell)}(\vec{x}_4)\rangle_{\mathrm{C}} \\ &= \delta_{i_1 i_2}\delta_{i_3 i_4} \frac{1}{n_{\ell-1}} V_4^{(\ell)}(\vec{x}_1, \vec{x}_2; \vec{x}_3, \vec{x}_4) + \text{perms},\end{aligned} \tag{19}$$

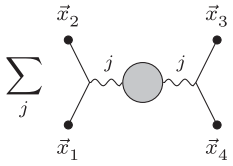
We can obtain $\frac{1}{n_{\ell-1}} V_4^{(\ell)}(\vec{x}_1, \vec{x}_2; \vec{x}_3, \vec{x}_4)$ by setting $i_1 = i_2$, $i_3 = i_4$ in the diagram in Eq. (15). For simplicity we will leave implicit the pairwise-equal neuron index labels for the external fields, writing:

$$\frac{1}{n_{\ell-1}} V_4^{(\ell)}(\vec{x}_1, \vec{x}_2; \vec{x}_3, \vec{x}_4) = \sum_{j_1, j_2} \cdot \tag{20}$$



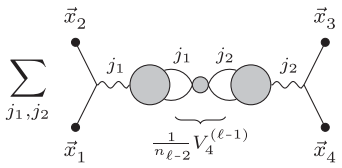
To evaluate the diagram, we need to consider two cases, $j_1 = j_2$ and $j_1 \neq j_2$. For $j_1 = j_2 \equiv j$, we can use the free theory to evaluate the blob at LO:

$$\sum_j \cdot \tag{21}$$



$$= \frac{(C_W^{(\ell)})^2}{n_{\ell-1}} \langle \Delta(\vec{x}_1, \vec{x}_2)\Delta(\vec{x}_3, \vec{x}_4)\rangle_{\mathcal{K}_0^{(\ell-1)}} + \mathcal{O}\left(\frac{1}{n^2}\right).$$

As in Eq. (17), we have dropped the layer and neuron indices on Δ. Note that each $\phi^2\Delta$ vertex comes with a factor of $\frac{C_W^{(\ell)}}{n_{\ell-1}}$, while the neuron index sum yields a factor of $n_{\ell-1}$, so the final result is $\mathcal{O}(\frac{1}{n})$. For $j_1 \neq j_2$, free-theory propagators cannot connect Δ_{j_1} and Δ_{j_2}, and the leading contribution is from inserting a connected four-point correlator of the $(\ell-1)$th layer:



$$\sum_{j_1,j_2} = \frac{\left(C_W^{(\ell)}\right)^2}{4\, n_{\ell\text{-}2}} \left[\prod_{\alpha=1}^{4} \int d\vec{y}_\alpha d\vec{z}_\alpha \left(\mathcal{K}_0^{(\ell-1)}\right)^{-1}(\vec{y}_\alpha, \vec{z}_\alpha)\right] V_4^{(\ell-1)}(\vec{y}_1, \vec{y}_2; \vec{y}_3, \vec{y}_4)$$
$$\left\langle \Delta(\vec{x}_1, \vec{x}_2)\, \phi(\vec{z}_1)\, \phi(\vec{z}_2) \right\rangle_{\mathcal{K}_0^{(\ell-1)}} \left\langle \Delta(\vec{x}_3, \vec{x}_4)\, \phi(\vec{z}_3)\, \phi(\vec{z}_4) \right\rangle_{\mathcal{K}_0^{(\ell-1)}} + \mathcal{O}\left(\frac{1}{n^2}\right) \tag{22}$$

$$= \frac{\left(C_W^{(\ell)}\right)^2}{4\, n_{\ell\text{-}2}} \left(\prod_{\alpha=1}^{4} \int d\vec{y}_\alpha\right) V_4^{(\ell-1)}(\vec{y}_1, \vec{y}_2; \vec{y}_3, \vec{y}_4) \left\langle \frac{\delta^2 \Delta(\vec{x}_1, \vec{x}_2)}{\delta\phi(\vec{y}_1)\delta\phi(\vec{y}_2)} \right\rangle_{\mathcal{K}_0^{(\ell-1)}} \left\langle \frac{\delta^2 \Delta(\vec{x}_3, \vec{x}_4)}{\delta\phi(\vec{y}_3)\delta\phi(\vec{y}_4)} \right\rangle_{\mathcal{K}_0^{(\ell-1)}} + \mathcal{O}\left(\frac{1}{n^2}\right).$$

In this diagram, internal solid lines (labeled with neuron indices j_1, j_2) denote $\phi_{j_1}^{(\ell-1)}$, $\phi_{j_2}^{(\ell-1)}$ propagators. Exchanging the two $\phi_{j_1}^{(\ell-1)}$ lines or the two $\phi_{j_2}^{(\ell-1)}$ lines results in the same diagram, hence a symmetry factor $\frac{1}{2^2} = \frac{1}{4}$. The smaller blob at the center (together with the attached propagators) represents a connected four-point correlator of the $(\ell-1)$th layer, $\frac{1}{n_{\ell-2}} V_4^{(\ell-1)}$. The larger blobs give rise to the correlators in the second line of Eq. (22); they are automatically connected since $\langle \Delta_j^{(\ell-1)}(\vec{x}_1, \vec{x}_2)\rangle = \langle \phi_j^{(\ell-1)}(\vec{x})\rangle = 0$. A correlator $\langle \phi(\vec{x})\ldots\rangle$ by its standard definition includes the propagators $\mathcal{K}_0^{(\ell-1)}(\vec{x}, \vec{y})\ldots$ (with $\vec{y}$ to be integrated over), so when we use correlators to build up diagrams, each internal propagator connecting two correlators (blobs) is counted twice. To avoid double-counting we thus insert an *inverse* propagator for each internal line in the diagram. This explains the factors of $(\mathcal{K}_0^{(\ell-1)})^{-1}$ in the first line of Eq. (22), which effectively amputate the connected four-point correlator (or equivalently the larger blobs in the diagram). The final expression in Eq. (22) is obtained by Wick contraction, which yields factors of $\mathcal{K}_0^{(\ell-1)}$ that cancel $(\mathcal{K}_0^{(\ell-1)})^{-1}$. Compared to the $j_1 = j_2$ case in Eq. (21), we now have an extra factor of $\frac{1}{n_{\ell-2}}$ from the insertion of the $(\ell-1)$ th-layer connected four-point correlator and an extra factor of $n_{\ell-1}$ from neuron index summation, so the result is again $\mathcal{O}(\frac{1}{n})$.

Adding up Eqs. (21) and (22) gives the final result for $V_4^{(\ell)}$ in terms $V_4^{(\ell-1)}$ and $\mathcal{K}_0^{(\ell-1)}$, i.e. the RG flow of V_4, which agrees with Refs. [14,21]. Both equations are $\mathcal{O}(\frac{1}{n})$, so $V_4^{(\ell)}$ defined by Eq. (19) is $\mathcal{O}(1)$.

We would like to note that the way we use Feynman diagrams in neural network EFT calculations is perhaps slightly different from what one is used to in other contexts. Usually one would derive Feynman rules for propagators and interaction vertices, and use them to build diagrams from which one can calculate correlators in terms of parameters of the theory. In the present case, however, our goal is to derive RG flows, which are *relations* between correlators. As we have seen above, the general strategy is to first write ℓ th-layer ϕ correlators in terms of $(\ell-1)$ th-layer Δ correlators, i.e. expectation values of (products of) $\Delta_j^{(\ell-1)}$'s, using the $\phi^2\Delta$ vertex [last diagram in Eq. (13)], and then calculate these $(\ell-1)$ th-layer Δ correlators in terms of $(\ell-1)$ th-layer ϕ correlators. In the second step, if a Δ correlator involves identical neuron indices [e.g. in Eq. (21)], it simply takes its free-theory value expressed in terms of free propagators (i.e. two-point ϕ correlators) at LO; if distinct neuron indices are involved, we need to insert mixed Δ-ϕ correlators [e.g. the larger blobs in Eq. (22)] to bridge the Δ's and four- or higher-point ϕ correlators. In either case, we can express the result in terms of free-theory expectation values of $(\ell-1)$th-layer single neuron operators, Eq. (18), with $\mathcal{O}$ a product of Δ's and ϕ's [with the exception of the LO two-point correlator $\mathcal{K}_0$ where $\mathcal{O} = \sigma_{\vec{x}_1}\sigma_{\vec{x}_2}$; see Eq. (17)]. By Wick contractions we can then rewrite these expectation values in terms of those of functional derivatives of Δ's [as in e.g. Eq. (22)].

The diagrammatic calculation extends straightforwardly to higher-point connected correlators, and provides a concise framework to systematically analyze finite-width effects in deep neural networks. In the Supplemental Material [42] we present new results for the connected six-point and eight-point correlators as further examples.

The RG flow can also be formulated at the level of the EFT action. The idea is to consider a tower of EFTs, $\mathcal{S}_{\text{eff}}^{(\ell)}$ $(\ell = 1, \ldots, L)$, obtained by integrating out the neurons and ghosts in all but the ℓth layer. They take the form:

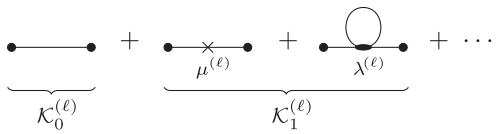
$$\mathcal{S}^{(\ell)}_{\text{eff}} = \int d\vec{x}_1 d\vec{x}_2 (\mathcal{K}_0^{(\ell)} + \mu^{(\ell)})^{-1}(\vec{x}_1, \vec{x}_2)\left[\frac{1}{2}\phi_i^{(\ell)}(\vec{x}_1)\phi_i^{(\ell)}(\vec{x}_2) - \bar{\psi}_{i'}^{(\ell)}(\vec{x}_1)\psi_{i'}^{(\ell)}(\vec{x}_2)\right]$$
$$- \int d\vec{x}_1 d\vec{x}_2 d\vec{x}_3 d\vec{x}_4 \lambda^{(\ell)}(\vec{x}_1, \vec{x}_2; \vec{x}_3, \vec{x}_4)\left[\frac{1}{8}\phi_i^{(\ell)}(\vec{x}_1)\phi_i^{(\ell)}(\vec{x}_2)\phi_j^{(\ell)}(\vec{x}_3)\phi_j^{(\ell)}(\vec{x}_4)\right.$$
$$\left. - \frac{1}{2}\phi_i^{(\ell)}(\vec{x}_1)\phi_i^{(\ell)}(\vec{x}_2)\bar{\psi}_{j'}^{(\ell)}(\vec{x}_3)\psi_{j'}^{(\ell)}(\vec{x}_4) + \frac{1}{2}\bar{\psi}_{i'}^{(\ell)}(\vec{x}_1)\psi_{i'}^{(\ell)}(\vec{x}_2)\bar{\psi}_{j'}^{(\ell)}(\vec{x}_3)\psi_{j'}^{(\ell)}(\vec{x}_4)\right] + \cdots \tag{23}$$

where summation over repeated indices is assumed. To determine the EFT couplings $\mu^{(\ell)}, \lambda^{(\ell)} \sim \mathcal{O}(\frac{1}{n})$, we can calculate the connected correlators $\langle \phi_{i_1}^{(\ell)}(\vec{x}_1)\phi_{i_2}^{(\ell)}(\vec{x}_2)\rangle$, $\langle \phi_{i_1}^{(\ell)}(\vec{x}_1)\phi_{i_2}^{(\ell)}(\vec{x}_2)\phi_{i_3}^{(\ell)}(\vec{x}_3)\phi_{i_4}^{(\ell)}(\vec{x}_4)\rangle_{\text{C}}$ using Eq. (23), and require the results match those derived above from Eq. (5). Then their RG flows directly follow from those of $\langle \mathcal{G}^{(\ell)}(\vec{x}_1, \vec{x}_2)\rangle$, $V_4^{(\ell)}(\vec{x}_1, \vec{x}_2; \vec{x}_3, \vec{x}_4)$ discussed above. The calculation of neuron correlators from Eq. (23) follows a separate set of Feynman rules which in fact closely resemble the familiar ones from standard field theory calculations and should be evident from the equations below. For example, matching the connected four-point correlator relates $\lambda^{(\ell)}$ to $V_4^{(\ell)}$ and $\mathcal{K}_0^{(\ell)}$:

$$\frac{1}{n_{\ell-1}} V_4^{(\ell)}(\vec{x}_1, \vec{x}_2; \vec{x}_3, \vec{x}_4) =$$

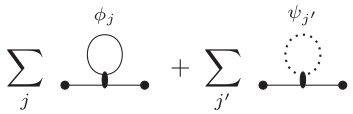
$$+ \mathcal{O}\left(\frac{1}{n^2}\right)$$
$$= \prod_{\alpha=1}^{4} \int d\vec{y}_\alpha \mathcal{K}_0^{(\ell)}(\vec{x}_\alpha, \vec{y}_\alpha)\lambda^{(\ell)}(\vec{y}_1, \vec{y}_2; \vec{y}_3, \vec{y}_4) + \mathcal{O}\left(\frac{1}{n^2}\right), \tag{24}$$

where we use an elongated vertex to indicate pairing of the four arguments of $\lambda^{(\ell)}$. For the two-point correlator, the calculation involves the following diagrams:



(25)

The alternative pairing of legs at the quartic vertex in the last diagram results in an $\mathcal{O}(\frac{n}{n})$ contribution, which, however, is canceled by diagrams with ghost loops due to the opposite-sign coupling:



(26)

Similar cancellations also explain the exclusion of $\mathcal{O}(\frac{n}{n^2})$ loop diagrams in the calculation of the connected four-point correlator in Eq. (24).[2]

IV. STRUCTURES OF RG FLOW AND CRITICALITY

The RG analysis of neuron statistics is highly relevant for the critical tuning of deep neural networks. The necessity of tuning has long been appreciated in practical applications of deep learning, especially in the context of mitigating the infamous exploding and vanishing gradient problems which make it difficult to train deep networks given finite machine precision. In the EFT framework, this is related to the fact that generic choices of hyperparameters $C_b^{(\ell)}, C_W^{(\ell)}$ lead to exponential scaling of neuron correlators under RG. Taming the exponential behaviors requires tuning the network to criticality by judiciously setting these hyperparameters [38–40]. At the kernel level, the criticality analysis of Ref. [21] reveals two prominent universality classes which networks with a variety of activation functions fall into: scale-invariant (including e.g. ReLU) and $\mathcal{K}^\star = 0$ (including e.g. tanh). In each case, $\mathcal{K}_0^{(\ell)}$ flows toward a nontrivial fixed point as ℓ increases; crucially, the scaling near the fixed point is power-law rather than exponential, which allows information to propagate through the layers so the network can learn nontrivial features from data.

While previous criticality analyses have mostly focused on the two-point correlator, it is important to also consider higher-point correlators because they encode fluctuations across the ensemble. In other words, it is not sufficient to require the networks are well-behaved on average, but the scaling behavior of each network must be close to the average. At first sight, criticality seems to

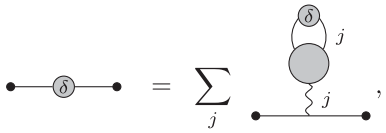
[2]We can reproduce the effective theory of Ref. [21] by integrating out the ghosts from Eq. (23). Calculations within this ghostless effective theory give rise to $\mathcal{O}(\frac{n_{\ell-1}}{n_\ell})$ terms, necessitating either working in the regime $n_\ell \gg n_{\ell-1} \gg 1$ or marginalizing the action over all but an $\mathcal{O}(1)$ number of neurons in the $(\ell-1)$th layer to have a perturbative $\frac{1}{n}$ expansion. Retaining the ghosts avoids such subtleties, rendering $\mu^{(\ell)}$ genuinely $\mathcal{O}(\frac{1}{n})$ and the difference between $\lambda^{(\ell)}$ and the (amputated) connected four-point correlator genuinely $\mathcal{O}(\frac{1}{n^2})$.

impose more constraints than the number of tunable hyperparameters, if we require power-law scaling of all higher-point correlators at arbitrary input points. However, as we will show, the structures of RG flow, manifest in the diagrammatic formulation, are such that tuning $\mathcal{K}_0$ to criticality actually ensures power-law scaling of all higher-point connected correlators near the fixed point.

Let us start with the two-point correlator. The asymptotic scaling behavior (exponential vs power-law) can be inferred from the following question: upon an infinitesimal variation at the $(\ell-1)$th layer,

$$\langle \mathcal{G}^{(\ell-1)}(\vec{x}_1,\vec{x}_2)\rangle \to \langle \mathcal{G}^{(\ell-1)}(\vec{x}_1,\vec{x}_2)\rangle + \delta\langle \mathcal{G}^{(\ell-1)}(\vec{x}_1,\vec{x}_2)\rangle, \tag{27}$$

how does $\langle \mathcal{G}^{(\ell)}(\vec{x}_1,\vec{x}_2)\rangle$ change? Diagrammatically, this can be calculated as follows:



(28)

where a blob labeled "δ" denotes the variation of the (two-point) correlator. The result is

$$\delta\langle \mathcal{G}^{(\ell)}(\vec{x}_1,\vec{x}_2)\rangle = \int d\vec{y}_1 d\vec{y}_2 \chi^{(\ell)}(\vec{x}_1,\vec{x}_2;\vec{y}_1,\vec{y}_2) \times \delta\langle \mathcal{G}^{(\ell-1)}(\vec{y}_1,\vec{y}_2)\rangle, \tag{29}$$

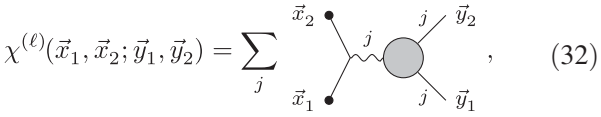
or, equivalently:

$$\frac{\delta\langle \mathcal{G}^{(\ell)}(\vec{x}_1,\vec{x}_2)\rangle}{\delta\langle \mathcal{G}^{(\ell-1)}(\vec{y}_1,\vec{y}_2)\rangle} = \chi^{(\ell)}(\vec{x}_1,\vec{x}_2;\vec{y}_1,\vec{y}_2), \tag{30}$$

where

$$\begin{aligned}
&\chi^{(\ell)}(\vec{x}_1,\vec{x}_2;\vec{y}_1,\vec{y}_2)\\
&= \frac{C_W^{(\ell)}}{2}\int d\vec{z}_1 d\vec{z}_2 (\mathcal{K}_0^{(\ell-1)})^{-1}(\vec{y}_1,\vec{z}_1)(\mathcal{K}_0^{(\ell-1)})^{-1}(\vec{y}_2,\vec{z}_2)\\
&\quad\times \langle \Delta(\vec{x}_1,\vec{x}_2)\phi(\vec{z}_1)\phi(\vec{z}_2)\rangle_{\mathcal{K}_0^{(\ell-1)}} + \mathcal{O}\left(\frac{1}{n}\right)\\
&= \frac{C_W^{(\ell)}}{2}\left\langle \frac{\delta^2\Delta(\vec{x}_1,\vec{x}_2)}{\delta\phi(\vec{y}_1)\delta\phi(\vec{y}_2)}\right\rangle_{\mathcal{K}_0^{(\ell-1)}} + \mathcal{O}\left(\frac{1}{n}\right).
\end{aligned} \tag{31}$$

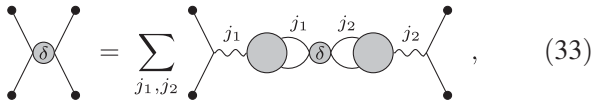
We can clearly see the structural similarity to Eq. (22) above. Ultimately, the same subdiagram enters both equations, and we can write:



(32)

where the $\phi_j^{(\ell-1)}$ legs on the right side are amputated, and an exchange symmetry between them in the full diagram is assumed [hence a symmetry factor $\frac{1}{2}$ in Eq. (31)].

The same pattern persists for higher-point connected correlators. For the connected four-point correlator, an infinitesimal variation of $V_4^{(\ell-1)}(\vec{x}_1,\vec{x}_2;\vec{x}_3,\vec{x}_4)$ results in a change in $V_4^{(\ell)}(\vec{x}_1,\vec{x}_2;\vec{x}_3,\vec{x}_4)$:



(33)

and we find:

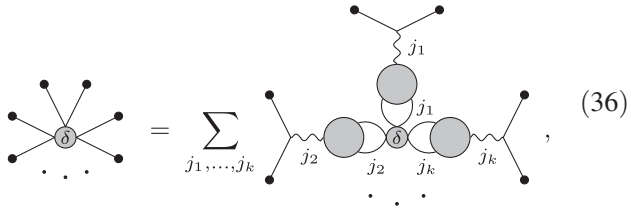
$$\begin{aligned}
&\frac{n_{\ell-2}}{n_{\ell-1}}\frac{\delta V_4^{(\ell)}(\vec{x}_1,\vec{x}_2;\vec{x}_3,\vec{x}_4)}{\delta V_4^{(\ell-1)}(\vec{y}_1,\vec{y}_2;\vec{y}_3,\vec{y}_4)}\\
&= \frac{1}{2}[\chi^{(\ell)}(\vec{x}_1,\vec{x}_2;\vec{y}_1,\vec{y}_2)\chi^{(\ell)}(\vec{x}_3,\vec{x}_4;\vec{y}_3,\vec{y}_4)\\
&\quad + \chi^{(\ell)}(\vec{x}_1,\vec{x}_2;\vec{y}_3,\vec{y}_4)\chi^{(\ell)}(\vec{x}_3,\vec{x}_4;\vec{y}_1,\vec{y}_2)],
\end{aligned} \tag{34}$$

where the symmetrized form arises because $V_4^{(\ell)}(\vec{x}_1,\vec{x}_2;\vec{x}_3,\vec{x}_4) = V_4^{(\ell)}(\vec{x}_3,\vec{x}_4;\vec{x}_1,\vec{x}_2)$.

Generally, the connected $2k$-point correlator is defined by

$$\begin{aligned}
&\langle \phi_{i_1}^{(\ell)}(\vec{x}_1)\ldots\phi_{i_{2k}}^{(\ell)}(\vec{x}_{2k})\rangle_{\mathrm{C}}\\
&= \delta_{i_1 i_2}\cdots\delta_{i_{2k-1}i_{2k}}\frac{1}{n_{\ell-1}^{k-1}}V_{2k}^{(\ell)}(\vec{x}_1,\vec{x}_2;\ldots;\vec{x}_{2k-1},\vec{x}_{2k}) + \text{perms},
\end{aligned} \tag{35}$$

where $V_{2k}^{(\ell)}$ can be shown to be $\mathcal{O}(1)$ [22]. Its variation follows from:



(36)

and we have

$$\begin{aligned}
&\left(\frac{n_{\ell-2}}{n_{\ell-1}}\right)^{k-1}\frac{\delta V_{2k}^{(\ell)}(\vec{x}_1,\vec{x}_2;\ldots;\vec{x}_{2k-1},\vec{x}_{2k})}{\delta V_{2k}^{(\ell-1)}(\vec{y}_1,\vec{y}_2;\ldots;\vec{y}_{2k-1},\vec{y}_{2k})}\\
&= \mathrm{sym}\left[\prod_{k'=1}^{k}\chi^{(\ell)}(\vec{x}_{2k'-1},\vec{x}_{2k'};\vec{y}_{2k'-1},\vec{y}_{2k'})\right],
\end{aligned} \tag{37}$$

where "sym." means symmetrizing the expression in the same way as in Eq. (34).

The quantity $\chi^{(\ell)}(\vec{x}_1,\vec{x}_2;\vec{y}_1,\vec{y}_2)$ in the equations above is a generalization of the parallel and perpendicular susceptibilities, $\chi_\parallel$ and $\chi_\perp$, introduced in Ref. [21] when analyzing the special case of two nearby inputs. In the nearby-inputs limit, tuning the network to criticality means adjusting the hyperparameters $C_W^{(\ell)}$, $C_b^{(\ell)}$ such that the kernel recursion Eq. (17) has a fixed point $\mathcal{K}^\star$ where $\chi_\parallel=\chi_\perp=1$. In the Supplemental Material [42], we show that, at least for the scale-invariant and $\mathcal{K}^\star=0$ universality classes, this tuning actually implies a stronger condition is satisfied (at LO in $\frac{1}{n}$):

$$\chi^{(\ell)}(\vec{x}_1,\vec{x}_2;\vec{y}_1,\vec{y}_2)|_{\mathcal{K}_0^{(\ell-1)}=\mathcal{K}^\star} = \frac{1}{2}[\delta(\vec{x}_1-\vec{y}_1)\delta(\vec{x}_2-\vec{y}_2)+\delta(\vec{x}_1-\vec{y}_2)\delta(\vec{x}_2-\vec{y}_1)]. \quad (38)$$

Equation (38) ensures perturbations around the fixed point stay constant through the layers, not just for the two-point correlator, $\delta\langle\mathcal{G}^{(\ell)}(\vec{x}_1,\vec{x}_2)\rangle=\delta\langle\mathcal{G}^{(\ell-1)}(\vec{x}_1,\vec{x}_2)\rangle$, but for the entire tower of higher-point connected correlators, $\frac{1}{n_{\ell-1}^{k-1}}\delta V_{2k}^{(\ell)}(\vec{x}_1,\ldots,\vec{x}_{2k})=\frac{1}{n_{\ell-2}^{k-1}}\delta V_{2k}^{(\ell-1)}(\vec{x}_1,\ldots,\vec{x}_{2k})$. This in turn implies that for all of them, RG flow toward the fixed point is power-law instead of exponential, once the single condition Eq. (38) is satisfied. The discussion above makes it transparent that power-law scaling of higher-point connected correlators at criticality (previously observed in Refs. [21,22] up to eight-point level in the degenerate-input limit) has its roots in the structures of EFT interactions, as manifested by the common structure shared by the diagrams in Eqs. (28), (33) and (36).

V. SUMMARY AND OUTLOOK

In this paper, we introduced a diagrammatic formalism that significantly simplifies perturbative calculations of finite-width effects in EFTs corresponding to archetypical deep neural networks. The concise reproduction of known results and derivation of new results highlights the efficiency of the diagrammatic approach, while the incorporation of ghosts vastly simplifies $\frac{1}{n}$ counting in the EFT action. Our analysis also made transparent the structures of such EFTs which underlie the success of critical tuning in deep neural networks. In fact, a universal diagrammatic structure emerges in the RG analysis of all higher-point connected correlators of neuron preactivations, which means criticality (i.e. power-law as opposed to exponential scaling) of all the neuron statistics at initialization is governed by a single condition, Eq. (38).

From the deep learning point of view, the neuron correlators at initialization that we calculated provide the initial conditions for understanding training of neural networks. It is worth noting that the statistics at initialization already contain useful information as criticality is known to ensure trainability of neural networks [21,40,43]. An obvious next step is to extend the diagrammatic formalism to incorporate gradient-based training and simplify perturbative calculations involving the neural tangent kernel [7,8] and its differentials [11–13,21].

From the fundamental physics point of view, we are hopeful that much more can be learned from the intimate connection between neural networks and field theories. Understanding the structures of EFTs corresponding to other neural network architectures (e.g. recurrent neural networks [32,44] and transformers [45]) will allow us to gain further insights into this connection and potentially point to novel ML architecture designs for simulating field theories (see Refs. [27,33,46] for recent progress in this direction).

ACKNOWLEDGMENTS

We are particularly grateful to Sho Yaida for helpful conversations throughout the course of this work. We thank Hannah Day, Marat Freytsis, Boris Hanin, Yonatan Kahn, and Anindita Maiti for useful discussions and comments on a preliminary draft, and Guy Gur-Ari for related discussions. We thank Johannes Mosig and Pengpeng Xiao for pointing out typos in the preprint version. Feynman diagrams in this work were drawn using `tikz-feynman` [47]. This work was supported in part by the U.S. Department of Energy under the Grant No. DE-SC0011702. This work was performed in part at the Aspen Center for Physics, supported by the National Science Foundation under Grant No. NSF PHY-2210452, and the Kavli Institute for Theoretical Physics, supported by the National Science Foundation under Grant No. NSF PHY-1748958.

[1] R. M. Neal, Priors for infinite networks, in *Bayesian Learning for Neural Networks* (Springer, New York, NY, 1996), pp. 29–53.

[2] C. Williams, Computing with infinite networks, in *Advances in Neural Information Processing Systems*, edited by M. Mozer, M. Jordan, and T. Petsche (MIT Press, Cambridge, MA, 1996), Vol. 9.

[3] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, Deep neural networks as Gaussian processes, arXiv:1711.00165.

[4] A. G. d. G. Matthews, M. Rowland, J. Hron, R. E. Turner, and Z. Ghahramani, Gaussian process behaviour in wide deep neural networks, arXiv:1804.11271.
[5] G. Yang, Tensor programs I: Wide feedforward or recurrent neural networks of any architecture are Gaussian processes, arXiv:1910.12478.
[6] B. Hanin, Random neural networks in the infinite width limit as Gaussian processes, arXiv:2107.01562.
[7] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, Adv. Neural Inf. Process. Syst. **31**, 8571 (2018), arXiv:1806.07572.
[8] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, Wide neural networks of any depth evolve as linear models under gradient descent, J. Stat. Mech. (2020) 124002.
[9] G. Yang, Tensor programs II: Neural tangent kernel for any architecture, arXiv:2006.14548.
[10] J. M. Antognini, Finite size corrections for neural network Gaussian processes, arXiv:1908.10030.
[11] B. Hanin and M. Nica, Finite depth and width corrections to the neural tangent kernel, arXiv:1909.05989.
[12] J. Huang and H.-T. Yau, Dynamics of deep neural networks and neural tangent hierarchy, in *International Conference on Machine Learning* (PMLR, 2020), pp. 4542–4551, arXiv:1909.08156.
[13] E. Dyer and G. Gur-Ari, Asymptotics of wide networks from Feynman diagrams, arXiv:1909.11304.
[14] S. Yaida, Non-Gaussian processes and neural networks at finite widths, arXiv:1910.00019.
[15] G. Naveh, O. B. David, H. Sompolinsky, and Z. Ringel, Predicting the outputs of finite deep neural networks trained with noisy gradients, Phys. Rev. E **104**, 064301 (2021).
[16] I. Seroussi, G. Naveh, and Z. Ringel, Separation of scales and a thermodynamic description of feature learning in some CNNs, arXiv:2112.15383.
[17] K. Aitken and G. Gur-Ari, On the asymptotics of wide networks with polynomial activations, arXiv:2006.06687.
[18] A. Andreassen and E. Dyer, Asymptotics of wide convolutional neural networks, arXiv:2008.08675.
[19] J. Zavatone-Veth, A. Canatar, B. Ruben, and C. Pehlevan, Asymptotics of representation learning in finite Bayesian neural networks, Adv. Neural Inf. Process. Syst. **34**, 24765 (2021), arXiv:2106.00651.
[20] G. Naveh and Z. Ringel, A self consistent theory of Gaussian processes captures feature learning effects in finite CNNs, Adv. Neural Inf. Process. Syst. **34**, 21352 (2021), arXiv:2106.04110.
[21] D. A. Roberts, S. Yaida, and B. Hanin, *The Principles of Deep Learning Theory* (Cambridge University Press, Cambridge, England, 2022).
[22] B. Hanin, Correlation functions in random fully connected neural networks at finite width, arXiv:2204.01058.
[23] S. Yaida, Meta-principled family of hyperparameter scaling strategies, arXiv:2210.04909.
[24] D. A. Roberts, Why is AI hard and physics simple?, arXiv:2104.00008.
[25] S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, A correspondence between random neural networks and statistical field theory, arXiv:1710.06570.
[26] O. Cohen, O. Malka, and Z. Ringel, Learning curves for overparametrized deep neural networks: A field theory perspective, Phys. Rev. Res. **3**, 023034 (2021).
[27] J. Halverson, A. Maiti, and K. Stoner, Neural networks and quantum field theory, Mach. Learn. Sci. Tech. **2**, 035002 (2021).
[28] D. Bachtis, G. Aarts, and B. Lucini, Quantum field-theoretic machine learning, Phys. Rev. D **103**, 074510 (2021).
[29] A. Maiti, K. Stoner, and J. Halverson, Symmetry-via-duality: Invariant neural network densities from parameter-space correlators, arXiv:2106.00694.
[30] J. Erdmenger, K. T. Grosvenor, and R. Jefferson, Towards quantifying information flows: Relative entropy in deep neural networks and the renormalization group, SciPost Phys. **12**, 041 (2022).
[31] H. Erbin, V. Lahoche, and D. O. Samary, Non-perturbative renormalization for the neural network-QFT correspondence, Mach. Learn. Sci. Tech. **3**, 015027 (2022).
[32] K. T. Grosvenor and R. Jefferson, The edge of chaos: Quantum field theory and deep neural networks, SciPost Phys. **12**, 081 (2022).
[33] J. Halverson, Building quantum field theories out of neurons, arXiv:2112.04527.
[34] H. Erbin, V. Lahoche, and D. O. Samary, Renormalization in the neural network-quantum field theory correspondence.
[35] J. Zavatone-Veth and C. Pehlevan, Exact marginal prior distributions of finite Bayesian neural networks, Adv. Neural Inf. Process. Syst. **34**, 3364 (2021), arXiv:2104.11734.
[36] L. Noci, G. Bachmann, K. Roth, S. Nowozin, and T. Hofmann, Precise characterization of the prior predictive distribution of deep ReLU networks, Adv. Neural Inf. Process. Syst. **34**, 20851 (2021), arXiv:2106.06615.
[37] B. Hanin and A. Zlokapa, Bayesian interpolation with deep linear networks, arXiv:2212.14457.
[38] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, On the expressive power of deep neural networks, in *International Conference on Machine Learning* (PMLR, 2017), pp. 2847–2854, arXiv:1606.05336.
[39] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, Adv. Neural Inf. Process. Syst. **29** (2016), arXiv:1606.05340.
[40] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, Deep information propagation, arXiv:1611.01232.
[41] A. Maloney, D. A. Roberts, and J. Sully, A solvable model of neural scaling laws, arXiv:2210.16859.
[42] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevD.109.105007 for additional technical details, further demonstration of diagrammatic calculations and numerical results.
[43] H. Day, Y. Kahn, and D. A. Roberts, Feature learning and generalization in deep networks with orthogonal weights, arXiv:2310.07765.
[44] K. Segadlo, B. Epping, A. van Meegen, D. Dahmen, M. Krämer, and M. Helias, Unified field theory for deep and recurrent neural networks, J. Stat. Mech. (2022) 103401.

[45] E. Dinan, S. Yaida, and S. Zhang, Effective theory of transformers at initialization, arXiv:2304.02034.
[46] M. Demirtas, J. Halverson, A. Maiti, M. D. Schwartz, and K. Stoner, Neural network field theories: Non-Gaussianity, actions, and locality, Mach. Learn. Sci. Tech. **5**, 015002 (2024).
[47] J. Ellis, TikZ-Feynman: Feynman diagrams with TikZ, Comput. Phys. Commun. **210**, 103 (2017).