WEAK CORRELATIONS AS THE UNDERLYING PRINCIPLE FOR LINEARIZATION OF GRADIENT-BASED LEARNING SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning models, such as wide neural networks, can be viewed as nonlinear dynamical systems composed of numerous interacting degrees of freedom. When such systems approach the limit of infinite number of degrees of freedom, their dynamics tend to simplify. This paper investigates gradient descent-based learning algorithms that exhibit linearization in their parameters. We establish that this apparent linearity, arises from weak correlations between the first, and higher-order derivatives of the hypothesis function with respect to the parameters, at initialization. Our findings indicate that these weak correlations fundamentally underpin the observed linearization phenomenon of wide neural networks. Leveraging this connection, we derive bounds on the deviation from linearity during stochastic gradient descent training. To support our analysis, we introduce a novel technique for characterizing the asymptotic behavior of random tensors. We validate our theoretical insights through empirical studies, comparing the linearized dynamics to the observed correlations.

1 INTRODUCTION

Deep learning in general, and particularly over-parameterized neural networks, revolutionized various fields Graves et al. (2013); He et al. (2016); Krizhevsky et al. (2012); Silver et al. (2016), and they are likely to do much more. Yet, the underlying reason for their unprecedented success remains elusive. These systems can be interpreted as non-linear dynamical physical systems, characterized by a multitude of interacting degrees of freedom, which makes an exact description of their behavior exceedingly hard. However, it is well established that dynamical physical systems, when expanded to an infinite number of degrees of freedom tend to exhibit a simplified form of dynamics Anderson (1972), therefore, it seems plausible to consider such a limit in the context of deep learning systems.

A seminal study in 2018 Jacot et al. (2018), demonstrated that wide, fully connected neural networks, undergoing deterministic gradient descent, behave as though they were linear with respect to their parameters, (while maintaining a highly non-linear structure in their inputs). This structure has been denoted as the neural tangent kernel (NTK). The result sparked a plethora of subsequent research, generalizing it to other architectures, investigating the rate of convergence towards this linear limit, exploring the deviation of the parameters themselves from their initial configuration, decoding the structure of the kernels, and leveraging this knowledge to enhance our understanding of wide neural networks in general Lee et al. (2019); Li et al. (2019); Cao & Gu (2019); Karniadakis et al. (2021); Huang et al. (2021); Bartlett et al. (2021); Woodworth et al. (2020).

Subsequent discussions arose regarding the role of this limit in the exemplary performance of wide neural networks. Several studies have demonstrated that in certain contexts, infinitely wide neural networks converge to their global minimum at an exponential rate Jacot et al. (2018); Lee et al. (2019); Du et al. (2019); Allen-Zhu et al. (2019a;b); Daniely (2017); Li & Liang (2018); Du et al. (2018); Xu et al. (2020). Moreover, wide neural networks have been posited as effective tools for generalization, with connections drawn to the double descent phenomenon Belkin et al. (2019); Nakkiran et al. (2021); Mei & Montanari (2022). Although simplified, first-order approximations were shown to capture many of the critical properties of finite-width neural networks, making it a valuable framework for understanding neural networks behavior in general Li et al. (2019); Littwin et al. (2021); Yang & Hu (2020).

These conclusions however encounter some contention when juxtaposed with empirical evidence. Notably, several experiments indicate that for real-world data, NTK-based learning is less effective than its wide (albeit finite) neural network counterparts Lee et al. (2020); Fort et al. (2020). This apparent "*NTK inferiority paradox*" suggests that the relationship between the NTK limit and the success of finite neural networks may be more intricate than initially presumed.

A relatively understudied aspect within the framework of the neural tangent kernel pertains to the fundamental mechanisms underpinning the phenomenon of linearization. Previous research, such as Chizat et al. (2019), suggests that any gradient-based learning algorithm inherently possesses an intrinsic scale dictating its linearization behavior. Furthermore, incorporating an external parameter can modify this intrinsic scale, thereby directly influencing the extent to which linearization manifests.

In a related context, Liu et al. (2020) demonstrated that the ratio between the spectral norm of the Hessian and the Euclidean norm of the gradient governs the rate of linearization. Their analysis also established that, for wide neural networks, this ratio typically remains small, thus facilitating linearization.

Another relevant result in this field presented by Liu et al. (2022), who proposed that the linear behavior observed in wide neural networks emerges fundamentally due to their structural composition as ensembles of numerous weak sub-models.

The closest study to this work is that of Dyer & Gur-Ari (2019), which introduced a methodology grounded in Feynman diagrams to systematically analyze wide neural networks. Their technique enables precise computation of the asymptotic behavior of correlation functions, notably the NTK, in the limit of infinite network width. By leveraging methods from theoretical physics, their work derives finite-width corrections to training dynamics, thus providing deeper insights into the evolutionary behavior of wide neural networks beyond the infinite-width approximation. However, their results are limited, as their setup is restricted and only considers the average values of these correlations.

1.1 Our Contributions

1. We establish that for gradient descent-based learning, linearity is equivalent to weak correlations between the first and subsequent derivatives of the hypothesis function, concerning its parameters at their initial values (3.3). This equivalence is suggested as the fundamental cause for the linearization observed in wide neural networks.

2. We prove directly that wide neural networks display this weak derivative correlations structure. By relying and extending the tensor programs formalism Yang & Littwin (2021), our approach uniformly addresses a broader spectrum of architectures at once than any other proof we are aware of (4.2).

3. Drawing from the same concepts, we demonstrate how modifications in the architecture of linearizing learning systems, and more specifically, wide neural networks, affect the rate of linearization. This finding is juxtaposed Chizat et al. (2019)'s result, regarding the implications of the introduction of an external scale (3.3.2,4.2).

4. Harnessing the formalism of weak derivatives correlations, we derive a bound on the deviation from linearization over time during learning, when utilizing stochastic gradient descent (4.1). This is a generalization of the traditional result for deterministic gradient descent Lee et al. (2019). This is crucial, as in most practical scenarios, stochastic gradient descent generalize better than deterministic gradient descent Lee et al. (2020); Fort et al. (2020).

5. We introduce the notion of *random tensor asymptotic behavior*, as an effective analytical tool to describe the asymptotic behavior of random tensors (2). Such tensors are not only integral to machine learning, but also serve a pivotal role in diverse mathematical and physical frameworks. Understanding the typical asymptotic behavior of these tensors is relevant for addressing many questions across these fields.

The overarching simplicity and broad applicability of our findings suggest that weak derivatives correlations could very well be the foundational cause for the prevalent linearization attributes observed in wide neural networks, and possibly for other linearizing systems.

2 Random Tensor Asymptotic Behavior

Random tensors play a fundamental role in machine learning in general, and in this work in particular. In this section, we demonstrate the effectiveness of employing the stochastic big O notation of the subordinate norm, to characterize the *asymptotic behavior* of a general random tensor sequence (hereinafter referred to as a random tensor). Addressing the asymptotic behavior of such tensors involves two inherent challenges: the complexity arising from their multitude of components, and the stochastic nature of these components. In this part, we will define an effective way to characterize their asymptotic behavior.

Our primary norm in this work will be the *Subordinate Tensor Norm*, defined as in Kreyszig (1991):

$$\|M\| = \sup\left\{ M \cdot \left(v^1 \times \ldots \times v^r\right) \,\middle|\, v^1, \ldots, v^r \in S_{N_1}, \ldots, S_{N_r} \right\} \, . \tag{1}$$

We provide a detailed explanation of this definition and discuss its advantages in Appendix B.1.

We combine this concept with the *Stochastic Big-O Notation*, introduced in Appendix B.2, which is defined for a sequence of random tensors, denoted by $M \equiv \{M_n\}_{n=1}^{\infty}$. Henceforth, we regard M as a random tensor depending on a limiting parameter $n \in \mathbb{N}$[1]. This leads us to the definition of a new asymptotic upper bound for random tensors.

Denoting $\mathcal{N} = \{f : \mathbb{N} \to \mathbb{R}^{0+}\}$ as the set of all functions from $\mathbb{N}$ to the non-negative real numbers, we introduce the following definition:

Definition 2.1 (Asymptotic Upper Bound of Random Tensors)**.** A random tensor M, as defined above, is said to be asymptotically upper bounded by $f \in \mathcal{N}$ as follows:

$$M = O\left(f\right) \, , \tag{2}$$

if and only if:

$$\forall g \in \mathcal{N} \; s.t \; f = o\left(g\right) : \lim_{n \to \infty} P\left(\|M_n\| \leq g\left(n\right)\right) = 1 \, . \tag{3}$$

The lower asymptotic bound, $f = \Omega(M)$, is defined analogously but with the inequality reversed and $g = o(f)$.

Like with an infinite number of deterministic sequences, where pointwise convergence often falls short and uniform convergence is required, we demand a definition of a uniform asymptotic bound for discussing an infinite number of random tensors. This concept is rigorously defined in appendix C.1.

Remark 2.1. For a finite number of tensors, it can simply be demonstrated that the uniform bound aligns with the pointwise asymptotic bound, analogous to sequences convergence.

We demonstrate in lemma C.6 that this notation inherits many of the norm's properties it is defined above, including all of the properties of the subordinate norm, delineated in lemma C.1. Furthermore, it satisfies several other useful properties, outlined in appendix C.3.

2.1 Proprieties

Remark 2.2. We denote $f \leq g$ or $O(f) \leq O(g)$ iff $f = O(g)$. We also denote $f < g$ or $O(f) < O(g)$ iff $f = O(g)$ and $f \nsim g$, where $f \sim g \Leftrightarrow O\left(f\right) = O\left(g\right) \Leftrightarrow f = O\left(g\right) \wedge g = O\left(f\right)$. It is important to note that $f < g$ can hold without necessitating $f = o(g)$.

It can be readily shown that for any random tensor M, there exist upper and lower bounds such that $O\left(h_-\right) \leq O\left(M\right) \leq O\left(h_+\right)$, and that they satisfy $h_- \leq h_+$. Furthermore, if h_+ and h_- satisfy $h_+ \sim h_-$, their asymptotic behavior is unique. Meaning that for any other pair h'_+, h'_-, the relationship $h_+ \sim h'_+ \sim h'_- \sim h_-$ still holds (C.5). In such scenarios, we assert that M possesses an exact asymptotic behavior, denoted as $O\left(h_+\right) = O\left(h_-\right)$.

[1]The results are applicable not only to $\mathbb{N}$, but also to any other set endowed with a total order.

The existence of such a pair however is not guaranteed, as illustrated by a random variable that for every $n \in \mathbb{N}$, has equal probability of one-half to yield either 1 or n. For this variable, the optimal upper bound is $O(n)$, and the optimal lower bound is $O(1)$, but these do not exhibit the same limiting behavior. Analogously, deterministic sequence may exhibit similar behavior, featuring multiple distinct partial limits. However, in the deterministic case, the *limsup* and *liminf* serve as the appropriate upper and lower limits respectively. This observation leads to the question of whether an appropriate asymptotic bound exists for the random case. It turns out, it does.

Theorem 2.1 (Definite Asymptotic Bounds for Tensors)**.** Consider a random tensor M with a limiting parameter n as described earlier. There exists $f \in \mathcal{N}$ serving as a tight/definite upper bound for M, satisfying:

$$M = O\left(f\right) \wedge \forall f \nless g : M \neq O\left(g\right) \ . \tag{4}$$

Furthermore, the asymptotic behavior of f is unique.

Explanation. Although the theorem's result is intuitive, the challenge arises from the fact that our order above $\mathcal{N}$ is not a total one, even when considering only the asymptotic behavior of the functions. For example, none of the following equations hold true:

$$\sin\left(\pi n\right) < \cos\left(\pi n\right), \quad \cos\left(\pi n\right) < \sin\left(\pi n\right), \quad \sin\left(\pi n\right) \sim \cos\left(\pi n\right) \ . \tag{5}$$

We address this issue by employing Zorn's lemma, as demonstrated in appendix C.2. □

Since every such random tensor M has precisely one definite asymptotic bound f, we can consider this bound as the *random tensor's asymptotic behavior*, represented as:

$$O(M) = O(f) \ . \tag{6}$$

3 Weak Correlations and Linearization

3.1 Notations for Supervised Learning

3.1.1 General notations

Supervised learning involves learning a *classifier*: a function $\hat{y} : X \to Y$ that maps an input set (here $X \subseteq \mathbb{R}^{d_X}$), to an output set (here $Y \subseteq \mathbb{R}^{d_Y}$), given a dataset of its values $X' \subseteq X$, denoted as the "*target function*". This is achieved by using an *hypothesis function*, in our case of the form $F : \mathbb{R}^N \to \{f : X \to Y\}$ which depends on certain parameters $\theta \in \mathbb{R}^N$ (in the case of fully connected neural networks for example, the weights and biases). The objective of supervised learning is to find the optimal values for these parameters such that F captures $\hat{y}$ best, with respect to a cost function $\mathcal{C}$ assumed here convex. We use $x \in X$ to denote elements in the input set, and $i, j = 1...d_Y$ to denote the output vector indices. The parameters θ are enumerated as $\theta_\alpha, \alpha = 1, ..., |\theta| = N$, and their initial values are denoted by $\theta_0 = \theta(0)$.

We work within the optimization framework of single input batches gradient descent-based training, which is defined such that for every learning step $s \in \mathbb{N}$:

$$\begin{aligned} \Delta^{x_s}\theta\left(s\right) = \theta\left(s+1\right) - \theta\left(s\right) = -\eta\nabla\mathcal{C}\left(F\left(\theta\right)\left(x_s\right), \hat{y}\left(x_s\right)\right)|_{\theta=\theta(s)} = \\ -\eta\nabla F\left(\theta\left(s\right)\right)\left(x_s\right)\mathcal{C}'\left(F\left(\theta\left(s\right)\right)\left(x_s\right), \hat{y}\left(x_s\right)\right) \ . \end{aligned} \tag{7}$$

Here, $\nabla_\alpha = \frac{\partial}{\partial\theta_\alpha}$ represents the gradient operator, x_s denotes the $s \in \mathbb{N}$th input data, and $\mathcal{C}'\left(y\right) = \nabla_y C\left(y\right)$ refers to the derivative of the cost function. The derivative matrix/the Jacobian ∇F is defined such that for every indices i, α, $\left(\nabla F\right)_{\alpha i} = \nabla_\alpha F_i$. We denote η as the learning rate and $\left(x_s, \hat{y}\left(x_s\right)\right)$ as the inputs and labels, respectively. The training path is defined as the sequence of inputs upon which we trained our system, represented by $\{x_s \in X'\}_{s=0}^{\infty}$. We assume that each input along this path is drawn from the same random distribution $\mathcal{P}$, neglecting the possibility of drawing the same input multiple times. The same distribution will be used for both training and testing. Moreover, we assume that the hypothesis function and the cost function $F, \mathcal{C}$ are analytical in their parameters. We study learning in the limit where the number of parameters $N \equiv |\theta| \to \infty$, with $N \equiv N(n)$ being a function of some other parameter $n \in \mathbb{N}$, denoted as the "limiting parameter". For neural networks, n is typically chosen as the width of the smallest layer, but we can choose any parameter governs the system's linearization.

Remark 3.1. This framework can be greatly generalised, as we discussed in appendix I.

3.1.2 NEURAL TANGENT KERNEL NOTATIONS

Numerous gradient descent learning systems (GDML) with different neural network architectures, display a linear-like structure in their parameters in the large width limit. In this linear limit, the hypothesis function takes the following form:

$$\begin{gathered} F_{lin}\left(0\right)=F\left(\theta_0\right),\forall s\in\mathbb{N}_0:F_{lin}\left(s+1\right)= \\ F_{lin}\left(s\right)-\Theta_0\left(\cdot,x_s\right)\mathcal{C}'\left(F_{lin}\left(s\right)\left(x_s\right),\hat{y}\left(x_s\right)\right)\ , \end{gathered} \tag{8}$$

with the kernel Θ defined such as:

$$\forall x,x'\in X:\Theta\left(\theta\right)\left(x,x'\right)=\eta\nabla F\left(\theta\right)\left(x\right)^T\nabla F\left(\theta\right)\left(x'\right),\Theta_0\equiv\Theta\left(\theta_0\right)\ , \tag{9}$$

where ∇F^T is the transpose of ∇F, the Jacobian.

3.2 THE DERIVATIVES CORRELATIONS

3.2.1 THE DERIVATIVES CORRELATIONS DEFINITION

In the following, we prove that linearization is equivalent to having weak correlations between the first, and higher derivatives of the hypothesis function, with respect to the initial parameters. We define the *derivative correlations* as follows:

Definition 3.1 (Derivatives Correlations)**.** We define the derivatives correlations of the hypothesis function for any positive integer $d\in\mathbb{N}$ and non-negative integer $D\in\mathbb{N}^0$ as:

$$\mathfrak{C}^{D,d}\left(\theta\right)=\frac{\eta^{\frac{D}{2}+d}}{D!d!}\nabla^{\times D+d}F\left(\theta\right)^T\left(\nabla F\left(\theta\right)\right)^{\times d}\ , \tag{10}$$

where the higher order derivatives defined such that for every $d\in\mathbb{N}$ and indices $i,\alpha_1...\alpha_d$, $\left(\nabla^{\times d}F\right)_{\alpha_1...\alpha_d,i}=\nabla_{\alpha_1}\cdots\nabla_{\alpha_d}F_i$.

More explicitly, presenting the inputs and indices of these tensors:

$$\begin{gathered} \mathfrak{C}^{D,d}\left(\theta\right)^{\alpha_{1+d}\cdots\alpha_{D+d}}_{i_0,i_1...i_d}\left(x_0,x_1\ldots x_d\right)= \\ \frac{\eta^{\frac{D}{2}+d}}{D!d!}\sum_{\alpha_1...\alpha_d=1}^{N}\nabla^{\times D+d}_{\alpha_1...\alpha_{D+d}}F_{i_0}\left(\theta\right)\left(x_0\right)\cdot\left(\nabla_{\alpha_1}F_{i_1}\left(\theta\right)\left(x_1\right)\cdots\nabla_{\alpha_d}F_{i_d}\left(\theta\right)\left(x_d\right)\right)\ . \end{gathered} \tag{11}$$

The objects in (10) are the correlation of the derivatives in the sense that $\alpha_1\ldots\alpha_d$ can be viewed as random variables, drawn from a uniform distribution of $\{1...N\}$, while θ and all other indices are fixed instances and hence deterministic. In this context, $\nabla^{\times D+d}F$ and $\nabla F\times\ldots\times\nabla F$ in (10) can be viewed as random vectors of the variables $\alpha_1\ldots\alpha_d$, and the summation in (10) represents the (unnormalized) form of the "Pearson correlation" between the two random vectors. The overall coefficient of the learning rate $\eta^{\frac{D}{2}+d}$ serves as the appropriate normalization, as we will demonstrate in appendix E and F. We will also denote: $\mathfrak{C}^d\left(\theta\right)\equiv\mathfrak{C}^{0,d}\left(\theta\right),\mathfrak{C}^{D,d}\equiv\mathfrak{C}^{D,d}\left(\theta_0\right),\mathfrak{C}^d\equiv\mathfrak{C}^d\left(\theta_0\right)$.

Essentially, $D+d$ represents the degree of the derivative under consideration when interacting with the first derivative, whereas d specifies the number of copies of the first derivative involved in the interaction.

An example for these correlations is the $D=0,d=1$ correlation, the correlation of the first derivative with itself, the kernel (9):

$$\mathfrak{C}^1\left(\theta\right)=\eta\nabla F\left(\theta\right)^T\nabla F\left(\theta\right)=\Theta\left(\theta\right)\ . \tag{12}$$

The definition for the asymptotic behavior for these derivative correlations is slightly nuanced due to the many different potential combinations of distinct inputs. We rigorously define it in appendix D.1.

In the remainder of the paper we will show how these correlations serve as an effective tool for the theoretical analysis of the linearization of wide neural networks. While this is their main purpose, they can also be used the evaluate linearization rate numerically.

At first glance, this may seem computationally impractical, as computing the D, d-th derivative requires summing $O\left(N^d\right)$ elements. However, we do not actually need to compute the full d-th gradient of F to obtain its correlation. Instead, we can use the chain rule:

$$\mathfrak{C}^{D,d}\left(\theta\right)^{\alpha_{1+d}\ldots\alpha_{D+d}}_{i_0,i_1\ldots i_d}\left(x_0, x_1\ldots x_d\right) = \frac{\eta^{\frac{D}{2}+d}}{D!d!}\partial_{a_1}\cdots\partial_{a_{D+d}}F_{i_0}\left(\begin{array}{c}\theta_0 + a_1\nabla F_{i_1}\left(\theta_0, x_1\right) + \ldots + a_d\nabla F_{i_d}\left(\theta_0, x_D\right) + \\ a_{1+d}e_{\alpha_{1+d}} + \ldots + a_{D+d}e_{\alpha_{D+d}}, x_0\end{array}\right). \tag{13}$$

computed for $a_1 = ... = a_D = 0$, where e_α is the the α-th standard basis vector. This approach reduces the computation to summing only $O\left(N\right)$ elements.

3.3 Equivalence of Linearity and Weak Derivatives Correlations

Our main theorems concern the equivalence of linearity and weak derivative correlations. In other words, weak correlations can be regarded as the fundamental reason for the linear structure of wide neural networks. These theorems are applicable for systems that are properly scaled in the initial condition, meaning that when taking $n \to \infty$ the different components of the system remain finite. We define in rigour exactly what it means in appendix D.2. We denote such systems as properly normalised GDMLs or *PGDML*s.

3.3.1 Our Main Theorems

The relationship between linearization and weak derivative correlations, is formalized through the equivalence theorems, which characterized by a monotonically increasing sequence $m(n)$, where $\lim_{n\to\infty} m(n) = \infty$. This sequence captures the rate of linearization or correlation decay, and constitutes an intrinsic parameter of the system. For instance, in the case of wide neural networks, one typically has $m(n) = \sqrt{n}$. Nevertheless, $m(n)$ may take any form that satisfies the stated conditions, with its mathematical role lying in defining the equivalence relation.

Theorem 3.1 (Fixed Weak Correlations and Linearization Equivalence)**.** Given the setup described in this section, for a sufficiently small learning rate $\eta < \eta_{the}$, the two properties are equivalent, where the asymptotic bounds are uniform for every $d, D \in \mathbb{N}$:

1. $m(n)$ - fixed weak derivatives correlation:

$$\mathfrak{C}^d = O\left(\frac{1}{m\left(n\right)}\right), \mathfrak{C}^{D,d} = O\left(\frac{1}{\sqrt{m\left(n\right)}}\right) \tag{14}$$

2. Simple linearity: for every fixed training step $s \in \mathbb{N}$:

$$F\left(\theta\left(s\right)\right) - F_{lin}\left(s\right) = O\left(\frac{1}{m\left(n\right)}\right), \tag{15}$$

$$\eta^{\frac{D}{2}}\left(\nabla^{\times D}F\left(\theta\left(s\right)\right) - \nabla^{\times D}F\left(\theta_0\right)\right) = O\left(\frac{1}{\sqrt{m\left(n\right)}}\right). \tag{16}$$

η_{the} is defined such as all the correlations are uniformly bounded by $O(1)$, to ensure the sum converges, as shown in Appendix E.2. Any system that does not satisfy this condition will diverge within only a few training steps, as we show in Appendix E.2. For fully connected networks for example, $\eta_{the} \sim \frac{1}{n}$.

The next theorem delineates an even stronger equivalence, which is also relevant for wide neural networks. It also encompasses the scaling of the learning rate.

Theorem 3.2 (Exponential Weak Correlations and Linearization Equivalence)**.** Given the setup described in this section, the two properties are equivalent, where the asymptotic bounds are uniform for every $D \in \mathbb{N}_0, d \in \mathbb{N}$:

1. $m(n)$ - power weak derivatives correlation: for $(D, d) \neq (0, 1)$:

$$\mathfrak{C}^{D,d} = O\left(\frac{1}{\sqrt{m\left(n\right)}}\right)^d. \tag{17}$$

2. Strong linearity: for every reparametrisation of the learning rate $\eta \to r(n)\eta$, $r(n) > 0$ and for every fixed training step $s \in \mathbb{N}$:

$$F\left(\theta\left(s\right)\right) - F_{lin}\left(s\right) = O\left(\frac{r\left(n\right)}{m\left(n\right)}\right) , \tag{18}$$

and for every $D \in \mathbb{N}$:

$$\left(\frac{\eta}{r(n)}\right)^{\frac{D}{2}} \left(\nabla^{\times D} F\left(\theta\left(s\right)\right) - \nabla^{\times D} F\left(\theta_0\right)\right) = O\left(\frac{r(n)}{\sqrt{m(n)}}\right) . \tag{19}$$

Explanation. We prove the theorems by considering for a general learning step $s \in \mathbb{N}$, the hypothesis function and its derivatives' Taylor series expansion around the $s-1$ step. Utilizing equations 7,11, we can find that the evolution of the derivatives of F and its derivatives during learning, is governed by a linear combination of the correlations of the form:

$$\Delta \frac{\eta^{\frac{D}{2}}}{D!} \nabla^{\times D} F\left(\theta\right) = \sum_{d=1}^{\infty} \mathfrak{C}^{D,d}\left(\theta\right)\left(-\mathcal{C}'\left(F\left(\theta\right), \hat{y}\right)\right)^{\times d} , \tag{20}$$

for every $D \in \mathbb{N}_0$, where $\Delta \nabla^{\times D} F$ is the change of $\nabla^{\times D} F$. For deterministic functions it is now straightforward to prove the equivalences by employing the arithmetic properties of the big O notation, and that [i] One can choose any $F - \hat{y}$ (as long as its asymptotic behavior is appropriate). [ii] Different components in our sum cannot cancel each other, since we can change η continuously; thus, for the sum to be small, all of the components must be small. The adjustments needed for our case of stochastic functions are minor, as, as we show in appendix C.3, our tensor asymptotic behavior notation satisfies many of the same properties of the deterministic big O notation. The complete proofs are in appendix E. We demonstrate empirically that linearizing systems have weak correlations in appendix A.

□

3.3.2 Relation to Related Results

As shown in theorem 3.2, a rescaling of η, such as $\eta \to r(n)\eta$, can either promote or impede the process of linearization. This observation remains valid for theorem 3.2 as long as $\eta < \eta_{\text{the}}$. This insight offers a deeper understanding of the findings presented by Chizat et al. (2019), specifically elucidating how an alteration of an external scale influences linearization, by affecting the scales of higher-order correlations differently from those of lower-order correlations.

A notable connection to another principal work, Liu et al. (2020), concerns the definition of derivative correlations themselves. In Liu et al. (2020), the authors established that linearization results from a small ratio between the spectral norm of the Hessian, and the norm of the gradient. The derivative correlations can be interpreted as a spectral norm, but concerning solely the gradient when considered as a vector. This interpretation refines the results presented in Liu et al. (2020). Unlike their approach, which required this ratio to be small within a neighborhood (ball), our framework demands its minimization specifically at the initialization point. Consequently, it necessitates the decay of higher-order correlations as well.

Another related work is the work of Huang & Yau (2020). In their work, they characterize the dynamics of wide neural networks using a hierarchy of kernels, where higher-order kernels evolve on slower time scales. Similarly to our paper, these time scales are proportional to $\frac{1}{\sqrt{n}}$, effectively capturing the deviation from the NTK limit. The relation of their work to ours is most evident in the gradient flow case, where their kernels can be expressed as linear combinations of our correlations. However, our result is more general, as it does not rely on the structural assumptions of wide neural networks, and also generalizes to finite learning rate GD. The most immediate benefit of that is that our framework applies to learning systems which are not captured by Huang & Yau (2020)'s framework. More fundamentally, by avoiding restricting ourselves only to neural networks, and instead introducing these new correlations, we obtain not only a sufficient condition for linearization, but an equivalence, providing a universal criterion applicable to generic learning systems.

The connection to Liu et al. (2022) is more abstract. Their argument, that the linear behavior observed in wide neural networks fundamentally emerges from their structural composition as ensembles of numerous weak sub-models is related to our work via the concept that neurons become independent in the infinite-width limit, precisely manifesting the absence of correlation that we emphasize.

The most closely related paper we are aware of is Dyer & Gur-Ari (2019). in this work the authors introduced a method using Feynman diagrams, to analyze wide neural networks. This approach systematically computes the asymptotic behavior of correlation functions, such as the Neural Tangent Kernel, in the large-width limit. The main difference between Our and their approach is that they measure the asymptotic behavior of correlations directly, rather than averaging their values. This distinction significantly restricts their setup, rendering many of their conclusions more conjectural, and less practical, compared to our findings.

3.3.3 The Chicken and the Egg of Linearization and Weak Correlations

The relationship between linearization and weak correlations in over-parameterized systems can be comprehended from two different viewpoints. The first perspective suggests that effective learning in such systems necessitates a form of implicit regularization, which inherently favors simplicity Belkin et al. (2019). This preference can be directly incorporated by imposing a linear (or at least approximately linear), structure in the highly over-parameterized regimes. Notably, in certain scenarios, linearization can facilitate exponential convergence rates, especially with respect to the training datasets, but in some instances, even with respect to the testing datasets Jacot et al. (2018); Lee et al. (2019); Du et al. (2019); Allen-Zhu et al. (2019b); Daniely (2017); Li & Liang (2018); Du et al. (2018); Xu et al. (2020); Allen-Zhu et al. (2019a). Hence, weak derivative correlations can be interpreted as a pragmatic approach for achieving linearization.

An alternative interpretation, aligning more closely with the spirit of this paper, suggests that weak derivative correlations do not primarily serve as a dynamic mechanism for linearization, but rather, as its underlying cause. In this context, persisting derivative correlations may indicate an inherent bias within the system, typically an undesirable one. Therefore, linearization can be viewed as a consequence of our attempt to avoid counterproductive biases, by demanding weak correlations.

This interpretation suggests that the prevalent perception of kernel learning as biased, and neural networks as unbiased, is a result-based fallacy. Had kernel learning empirically outperformed neural networks, it would seem natural to interpret linear learning in the function space, (assigning large eigenvalues to simpler functions and smaller ones to complex functions), as unbiased. In other words, we interpret linear learning as overly unbiased, while finite neural networks (through mechanisms not fully understood) prioritize the inherent bias of realistic data.

Moreover, if we possess some prior knowledge about an inherent biases in our problem, it might be advantageous to allow some non-decaying correlations, counteracting the process of linearization. Furthermore, as certain biases can enhance general learning algorithms (in the form of implicit and explicit regularization), this perspective might provide valuable insights into the "NTK inferiority paradox" introduced in the introduction (1). The reason why linear learning underperforms in comparison to finite neural networks, might be that it lack some beneficial biases, in the form of non vanishing correlations.

We elaborate on this point in appendix H.

4 Properties of Weakly Correlated PGDMLs

4.1 Application: Deviation from Linearity During Learning

Multiple studies have examined the deviation of the hypothesis function F from its linear approximation F_{lin} (8), as a function of n for a fixed learning step (especially in the context of wide neural networks). Yet, it seems that no research has explored the deviation between these functions with respect to the learning step for stochastic GD (7). This aspect is crucial since even if $F - F_{lin}$ vanishes for the initial learning steps, if it deviates too fast during learning, the linearization may not be evident for realistic large n.

To study how learning systems deviate from their linearization during the training process, we examine the case of an exponentially $m(n)$-weakly correlated PGDML, with learning rate satisfying $\eta < \eta_{cor}$. Here, η_{cor} is the standard critical learning rate, ensuring that the system are stable in the

NTK limit, as explained in Appendix F. We consider the problem over the span of $S \in \mathbb{N}$ learning steps, and assume that within this phase the linear solution approaches the true solution exponentially fast for some typical time $0 < T$, such that for every $s = 1 \dots S$:

$$\mathcal{C}'\left(F_{lin}\left(s\right), \hat{y}\right) = O\left(e^{-\frac{s}{T}}\right), \mathcal{C}''\left(F_{lin}\left(s\right), \hat{y}\right) = O\left(1\right) \ , \tag{21}$$

As we show in Appendix F, this is not a restrictive assumption, especially at the beginning of training, where the deviation from linearization matters the most.

Corollary 4.1 (Weakly Correlated PGDML Deviation Over Time)**.** Given the conditions described above, we obtain that for every $s = 1, \dots, S$:

$$F\left(\theta\left(s\right)\right) - F_{lin}\left(s\right) = O\left(\frac{s^0}{m\left(n\right)}\right) \ . \tag{22}$$

where the asymptotic bounds are uniform in s, and s^0 denotes s in the power of zero.

While this result addresses single-input batches stochastic GD, as we explained in appendix I, this it can be greatly generalized. Notably, the analysis for stochastic GD may be more relevant even for deterministic GD, than the conventional approaches that presuppose a training dataset. This is because, while the batch might be fixed, its initial selection is from a stochastic distribution.

Explanation. We prove the corollary by using a similar induction process as in theorems 3.1,3.2. However, here we also consider the dependency in the learning step, as detailed in appendix F. We are able to bound the deviation over time, by leveraging the fact that in the NTK limit, during the initial phases of the learning process, the system converges towards the target function exponentially fast[2] Jacot et al. (2018); Lee et al. (2019); Du et al. (2019); Allen-Zhu et al. (2019b); Daniely (2017); Li & Liang (2018); Du et al. (2018); Xu et al. (2020); Allen-Zhu et al. (2019a). We believe that subsequent research will be able to produce more refined bounds. □

4.2 Example: Wide Neural Networks

Numerous studies have demonstrated that a wide range of neural networks architectures exhibit linearization as they approach the infinite width limit. However, the existing proofs tend to be specific to particular architectures, and are often intricate in nature. The most comprehensive proof we aware of that uniformly encompasses a diverse set of architectures, is presented in Yang & Littwin (2021); Yang (2020). These works employed the tensor product formalism Yang (2019), which can describe most relevant variants of wide neural network architectures, as the composition of global linear operations, and point-wise non linear functions.

1. Relying on the semi-linear structure of FCNNs we were able show explicitly by induction that for appropriate activation functions wide neural networks are $\sqrt{n}$ weakly correlated, and power weakly correlated as shown in appendix G.
2. The framework of low correlations proves effective in discerning how modifications to our network influence its linearization. For instance, it is evident that $\sup_{n\in\mathbb{N}} \frac{\phi^{[n]}}{(n+1)!}$, governs the rate of linearization in FCNNs (G). This observation is why we demand for FCNNs, that over the relevant domain, the activation function satisfy:
$$\phi^{[n]} \leq O\left(\left(n+1\right)!\right) \ , \tag{23}$$
where $\phi^{[n]}$ is the n-th derivative of the network's activation function - ϕ.
3. Our proof for FCNNs can simply be generalised for any wide network, described by the tensor programs formalism (G.5.1). This is because, similarly to FCNNs, all such systems exhibit a wide semi-linear form by definition. Demonstrating that the linearization of these systems arises from weak correlations, allow us to utilize all of the insights we've found for weakly correlated systems in general. We were also been able to conceive linearizing network-based systems, that fall outside the scope of the tensor programs formalism (G.5.2).
Leveraging the notation of the asymptotic tensor behavior, our proof accommodates a broad spectrum of initialization schemes, extending beyond the Gaussian initialization predominantly employed in other studies.

[2]The known bounds for $\mathcal{C}'\left(F_{lin}, \hat{y}\right)$ are typically bounds over the variance. In appendix C.4, we discuss how an average exponential bound can be translated into a uniform probabilistic bound.

5 Discussion and Outlook

The linearization of large and complex learning systems is a widespread phenomenon, but our comprehension of it remains limited. We propose the weak derivatives correlations (3.1), is the underlying structure behind this phenomenon. We demonstrated that this formalism is natural for analyzing this linearization: [i] It allows for the determination of if, and how fast a general system undergoes linearization (3.3.1,4.2). [ii] It aids us in analyzing the deviations from linearization during learning (4.1).

The strength of our approach is that it does not rely on the structure of wide neural networks. This allows us to describe not only a sufficient condition for linearization, but a true equivalence. Furthermore, it enables us to identify precisely the structural properties that generate linearization. With this, we were able to provide new insights into the linearization of wide neural networks, accounting for factors such as training duration and activation function properties. Many of these findings were previously unknown or difficult to derive using existing methods.

These insights carry a practical implications. Effective systems should neither remain too close to their linear limits, nor deviate excessively. A cohesive framework that relates convergence behavior to network width, training dynamics, and activation function characteristics can guide the design of more robust and efficient future models.

Our approach raise a pivotal question (3.3.3): Is the emergence of the weak correlations structure simply a tool to ensure a linear limit for overparameterized systems? Or does weak correlations indicate an absence of inherent biases, leading to linearization? If the latter is true, it suggests that in systems with pre-existing knowledge, specific non-linear learning methodologies reflecting those biases might be beneficial. That could partially explain why the NTK limit falls short in comparison to finite neural networks.

At the core of our weak derivatives correlation framework, is the random tensor asymptotic behavior formalism, outlined in section 2. We have showcased its efficacy in characterizing the asymptotic behavior of random tensors, and we anticipate its utility to extend across disciplines that involve such tensors.

We demonstrate our results empirically in appendix A, and further discuss generalizations and limitations in appendix I.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pp. 242–252. PMLR, 2019b.

Philip W Anderson. More is different: broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047):393–396, 1972.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

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

Yvonne M Bishop, Stephen E Fienberg, and Paul W Holland. *Discrete multivariate analysis: Theory and practice*. Springer Science & Business Media, 2007.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Amit Daniely. Sgd learns the conjugate kernel class of the network. *Advances in Neural Information Processing Systems*, 30, 2017.

Yadolah Dodge. *The Oxford dictionary of statistical terms*. OUP Oxford, 2003.

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*, pp. 1675–1685. PMLR, 2019.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

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

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:5850–5861, 2020.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649. Ieee, 2013.

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*, pp. 770–778, 2016.

Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean. Power of data in quantum machine learning. *Nature communications*, 12(1):2631, 2021.

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

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Erwin Kreyszig. *Introductory functional analysis with applications*, volume 17. John Wiley & Sons, 1991.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

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.

Jaehoon Lee, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. *Advances in Neural Information Processing Systems*, 33:15156–15172, 2020.

Sanjeev Li, Arora, Simon S Du, Zhiyuan Hu, Wei, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31, 2018.

Etai Littwin, Tomer Galanti, and Lior Wolf. On random kernels of residual architectures. In *Uncertainty in Artificial Intelligence*, pp. 897–907. PMLR, 2021.

Chaoyue Liu, Libin Zhu, and Misha Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. *Advances in Neural Information Processing Systems*, 33: 15954–15964, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Transition to linearity of wide neural networks is an emerging property of assembling weak models. *arXiv preprint arXiv:2203.05104*, 2022.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75 (4):667–766, 2022.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. *arXiv preprint arXiv:2009.11848*, 2020.

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

Greg Yang. Tensor programs ii: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.

Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

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

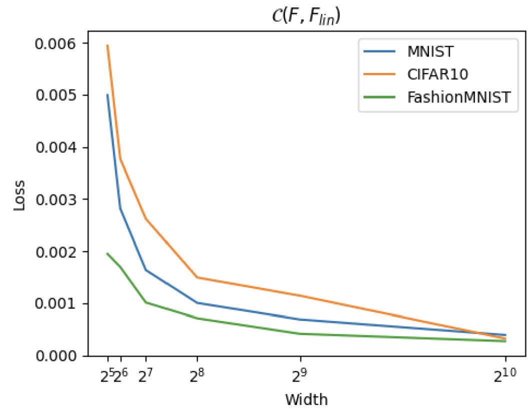

Figure 1: Relative loss between the neural network and its linear approximation versus the width, for three datasets. We used learning rate of 1, with 1160 samples, Relu activation, and 1000 epochs.

A Experimental Results

To support our arguments, we present here empirical numerical experiments. We show the training and testing dynamics of neural network, and its linearized approximation for varying network width. We consider a fully-connected architecture with mini-batch gradient descent, using learning rates according to the NTK normalization Lee et al. (2019), where we chose $\eta_0 = 1$. From computational considerations viewpoint, we focus on 10 classes classification, a total of 160 training samples and 32 test samples. MSE loss function is used for the training, where each class is represented by a different one-hot vector (10 dimensional vector).

We perform the analysis for three datasets: CIFAR10, MNIST and FMNIST, for different activation functions: Relu, Sigmoid, Erf, and for different numbers of layers L: 1, 2, and 3 (in addition to the output layer). For instance, 1 layer, and width of 128 in MNIST means: $784 \to 128 \to 10$.

The simulation were done in JAX packaged, and were based on Lee et al. (2019) work. We share our code in GitHub. All the results were obtained on CPU of a Apple M1 Pro 32GB, the running time is about 1 hour in total.

The difference function between f and f_{lin} was taken to be:

$$C\left(f, f_{\text{lin}}\right) = \frac{1}{10\left|X\right|}\sqrt{\sum_{x\in X}\sum_{i=1}^{10}\left(f_i(x) - f_{i,\text{lin}}(x)\right)^2}\,, \tag{24}$$

where the sub-index represents the output vector index (it depends on the class), and X is the set of the data samples, which consists of 32 samples.

Calculating high order derivatives is very costly in terms of computational resources. Therefore, we estimated the high order partial derivatives by a random sampling of set D weights at each layer, and averaged over a batch of samples X. Practically, we set $|D| = 60$ and $|X| = 160$, $d_y = 10$:

$$\mathfrak{C}^{0,2} \approx \frac{1}{S}\sqrt{\frac{1}{d_y}\sum_{i=1}^{10}\sum_{x\in X}\left(\frac{1}{|D|^2}\sum_{\alpha_1,\alpha_2\in D}\partial_{\alpha_1} f_i\left(x\right)\partial_{\alpha_2} f_i\left(x\right)\partial_{\alpha_1}\partial_{\alpha_2} f_i\left(x\right)\right)^2}\,, \tag{25}$$

and the same goes for $\mathfrak{C}^{0,3}$

B Additional Mathematical Background

In this section, we elaborate on several mathematical concepts that form the foundation for the ideas introduced in Section 2. We begin by defining the subordinate tensor norm and its key properties, then introduce a stochastic variant of "Big O" notation to characterize the asymptotic behavior of random tensors.

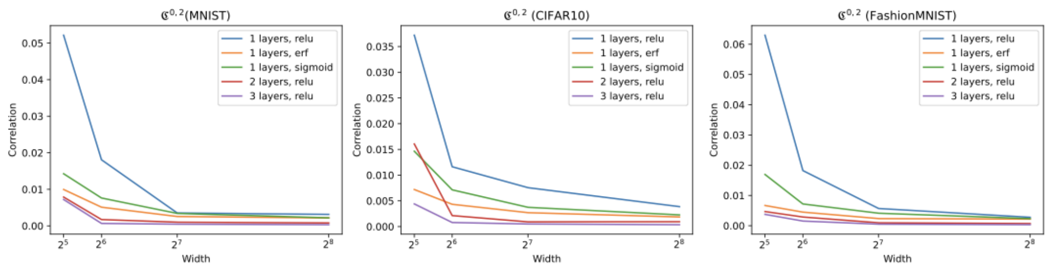

Figure 2: A comparison of the second order correlation approximation versus the width of the network, for different datasets (MNIST, CIFAR10 and FMNIST) and for different activation function (relu, erf, sigmoid) and number of layers (1,2,3). For all of these experiment, we used learning rate of 1, 160 samples, and 1000 epochs comparison of the second order correlation approximation versus the width of the network, for different datasets (MNIST, CIFAR10 and FMNIST), and for different activation function (Relu, Erf, Sigmoid) and number of layers (1,2,3). For all of these experiment, we used learning rate of 1, 1160 samples, and 1000 epochs.

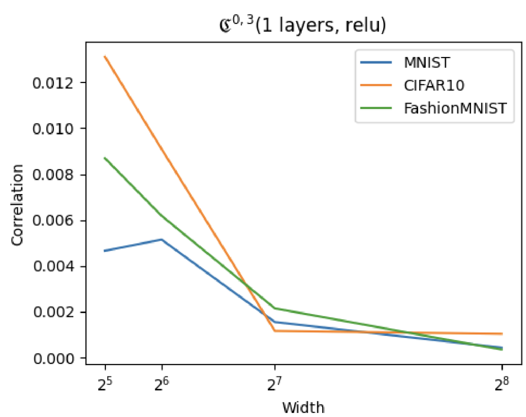

Figure 3: Third order correlation approximation function versus different widths, for three datasets. We used learning rate of 1, 1160 samples, Relu activation and 1000 epochs.

B.1 The Subordinate Tensor Norm

Let M be a tensor of rank $r \in \mathbb{N}_0$. Denote all its indices using the vector $\vec{i}$, such that each i_e for $e = 1...r$ can assume values $i_e = 1...N_e$. Consequently, the tensor comprises a total of $N = N_1 \cdots N_r$ elements.

We will use the *subordinate norm*, defined as Kreyszig (1991):

$$\begin{aligned} \|M\| = \sup \left\{ M \cdot \left(v^1 \times \ldots \times v^r\right) \middle| \, v^{1\ldots r} \in S_{N_{1\ldots r}} \right\} = \\ \sup \left\{ \sum_{i_1 \ldots i_r = 1}^{N_1 \ldots N_r} \left(M_{i_1 \ldots i_r} v^1_{i_1} \cdots v^r_{i_r}\right) \middle| \, v^{1\ldots r} \in S_{N_{1\ldots r}} \right\} \,, \end{aligned} \tag{26}$$

where $S_{N_k} = \left\{ v \in \mathbb{R}^{N_k} : v \cdot v = 1 \right\}$ represents the unit vectors of the appropriate dimensions. This norm satisfies certain algebraic properties outlined in lemma C.1, including: [i] the triangle inequality; [ii] for a tensor M and vectors $v_1 \ldots v_q$ with appropriately defined product, the condition $\left\| M \cdot \left(v^1 \times \ldots \times v^r\right) \right\| \le \|M\| \left\|v^1\right\| \cdots \|v^r\|$ holds; [iii] Given two tensors $M^{(1)}_{\vec{i}_1}, M^{(2)}_{\vec{i}_2}$ defining $M_{\vec{i}_1, \vec{i}_2} = M^{(1)}_{\vec{i}_1} M^{(2)}_{\vec{i}_2}$ then, $\|M\| = \left\|M^{(1)}\right\| \left\|M^{(2)}\right\|$.

Also, one has $\|M\| \le \|M\|_F$ (with equality for vectors) (C.2) where the Frobenius norm is:

$$\|M\|_F^2 = \sum_{\vec{i}} M_{\vec{i}}^2 \,. \tag{27}$$

B.2 Effectiveness of the Stochastic "Big O" Notation

Consider a general random tensor sequence, denoted by $M \equiv \{M_n\}_{n=1}^{\infty}$, which henceforth we will consider as a random tensor that depend on a limiting parameter $n \in \mathbb{N}$[3].

Our objective in this section is to identify a method to describe and bound the asymptotic behavior of such a tensor, which adheres to elementary algebraic properties. Specifically, we aim for the product of multiple bounded random tensors to be constrained by the product of their respective bounds.

Employing our defined norm (26), we can simplify our problem from general random tensors to positive random variables (rank zero tensors), as our norm satisfies the elementary algebraic properties established in Lemma C.1. This reduction is substantial; however, the challenge of addressing the non-deterministic nature of our variable remains.

One might initially consider the expectation value of the tensor's norm as a solution. This approach unfortunately falls short, because that for two positive random variables M_1, M_2 their product variance is not bounded by the product of their variance. In fact, for $M_1 = M_2$, the converse is true:

$$\mathrm{Var}\left(M_1 M_2\right) \ge \mathrm{Var}\left(M_2\right) \mathrm{Var}\left(M_1\right) \tag{28}$$

This issue becomes more pronounced when considering the product of multiple such variables, a frequent occurrence in this work. For instance, even with a basic zero-mean normal distribution with standard deviation σ, the higher moments of this distribution factor as $p!! = p(p-2)(p-4)\cdots$:

$$\forall p \in \mathbb{N} : \langle M^p \rangle = p!!\sigma^p \,. \tag{29}$$

When multiplying multiple such variables, these factors can accumulate in the lower moments, rendering this definition impractical for our purposes. Similarly, any attempt to define asymptotic behavior using the variable's moments will encounter similar difficulties.

To circumvent these challenges, we adopt the stochastic big O notation Dodge (2003); Bishop et al. (2007)[4].

C Random Tensors Asymptotic Behavior

In the following sections, we utilize the results of this section throughout our analyses repeatedly. Due to their intuitive nature, we may not consistently specify when we do so, and which lemma/theorem we are employing.

[3]The results are applicable not only for $\mathbb{N}$, but for any other set possessing an absolute order above it

[4]Our definition slightly differs from the standard definition for big O in probability notation, but it is straightforward to show its equivalence

C.1 PROPERTIES OF OUR NORM

In this subsection, we explore the properties satisfied by the subordinate norm. We omit the proofs as these properties are either well-known, or straightforward to prove, (and also enjoyable to derive).

Lemma C.1 (Algebraic properties of the subordinate norm)**.** The subordinate tensor norm (26) satisfies the following algebraic properties:

1. Given a tensor sequence $\left\{M^{(d)}\right\}_{d=1}^{D}$ where $D \in \mathbb{N} \cup \{\infty\}$, it satisfies the triangle inequality:
$$\left\|\sum_{d=1}^{D} M^{(d)}\right\| \leq \sum_{d=1}^{D}\left\|M^{(d)}\right\| \ , \tag{30}$$
where equality holds when the tensors are positively linearly dependent.

2. Given a tensor $M_{i_1 \ldots i_r}$, $1 \leq i_k \leq N_k$ for $1 \leq k \leq r$, and $q \leq r \in \mathbb{N}$ vectors $v^1_{i_1} \ldots v^q_{i_q}$ (with the same range of indices), then:
$$\left\|M \cdot v^1 \times \cdots \times v^q\right\| \leq \|M\| \left\|v^1\right\| \cdots \left\|v^q\right\| \ . \tag{31}$$

3. Given two tensors $M^{(1)}_{\vec{i}_1}$ and $M^{(2)}_{\vec{i}_2}$, their direct product $M_{\vec{i}_1 \vec{i}_2} = \left(M^{(1)} * M^{(2)}\right)_{\vec{i}_1 \vec{i}_2} = M^{(1)}_{\vec{i}_1} M^{(2)}_{\vec{i}_2}$, satisfies:
$$\|M\| = \left\|M^{(1)}\right\| \left\|M^{(2)}\right\| \ . \tag{32}$$
The generalization an arbitrary finite number of tensors is trivial.

Remark C.1. Parts 1 and 3 are also satisfied by the Frobenius norm.

Lemma C.2 (Relation to the Frobenius Norm)**.** Given a tensor M of rank $r \in \mathbb{N}$, the following holds:

1. For any tensor M:
$$\|M\| \leq \|M\|_F \ , \tag{33}$$
and if $r = 1$ (i.e., the tensor is a vector), then:
$$\|M\| = \|M\|_F = \sqrt{\sum_i M_i^2} \ . \tag{34}$$

2. For every $r' = 1 \ldots r$:
$$\|M\| = \sup\left\{ \left\| M \cdot \begin{pmatrix} v^1 \times \ldots \times v^{r'-1} \times \\ v^{r'+1} \times \ldots \times v^r \end{pmatrix} \right\|_F \middle| \begin{array}{c} v^1 \in S_{N_1} \ldots v^{r'-1} \in S_{N_{r'-1}} \\ v^{r'+1} \in S_{N_{r'+1}} \ldots v^r \in S_{N_r} \end{array} \right\} . \tag{35}$$

The first part of the lemma demonstrates that our norm is always bounded by the Frobenius norm, and the two norms coincide for vectors. The second part generalizes the first, indicating that when reducing any tensor to a vector, the two norms once again agree.

Lemma C.3 (Properties of the Maximizing Vectors)**.** Given a tensor M of rank $r \in \mathbb{N}$, there exist vectors $v^1 \ldots v^r$ of norm 1 such that:
$$\|M\| = M \cdot v^1 \times \cdots \times v^r \ . \tag{36}$$

This result indicates that the supremum is indeed a maximum. The vectors $v^1 \ldots v^{r'-1}, v^{r'+1} \ldots v^r$ are also the ones that maximize the cases demonstrated in the previous lemma.

Moreover, if the tensor is symmetric with respect to the permutation of the indices $i_1, i_2, \ldots, i_q$ and is non-zero, then:
$$v^{i_1} = v^{i_2} = \cdots = v^{i_q} \ . \tag{37}$$

Remark C.2. For $M = 0$, any set of vectors maximizes our result, irrespective of whether the vectors are identical or distinct.

C.2 Existence and Uniqueness of the Tensor Asymptotic Behavior

In this section, we discuss some of the more general properties that the tensor asymptotic behavior notation satisfies, regardless of the norm it is defined with respect to. The first lemma we present is a useful equivalent definition for bounding tensor asymptotic behavior. This equivalent definition will be beneficial for our later discussion:

Lemma C.4 (Equivalent Definitions for Tensor's Asymptotic Bound)**.** For any random tensor M and $f \in \mathcal{N}$, the two definitions for bounding the tensor's asymptotic behavior $O(M) \leq O(f)$ are equivalent (the first is the original definition, (2.1)):

1. $$\forall g \in \mathcal{N}\, s.t\, f = o\left(g\right) : \lim_{n\to\infty} P\left(\|M_n\| \leq g\left(n\right)\right) = 1\ . \tag{38}$$

2. $$\lim_{c\to\infty} \lim_{n\to\infty} P\left(\|M_n\| \leq cf\left(n\right)\right) = 1\ . \tag{39}$$

(The same applies for $O(f) \leq O(M)$).

The order in which we take the limits in equation 39 is crucial, as any random tensor satisfies the equation for any f, if we take first the limit of c.

It is straightforward to show that any random tensor M has lower and upper bounds:

Lemma C.5 (Bounding Tensor Asymptotic Behavior)**.** Given a random tensor M, there exist $h_-, h_+ \in \mathcal{N}$ such that:

$$O\left(h_-\right) \leq O\left(M\right) \leq O\left(h_+\right)\ . \tag{40}$$

To prove the asymptotic tensor behavior has meaning, we need to show that bounds not only always exist, but that **there is always one well-defined "best" upper bound - theorem 2.1**. We prove this theorem after lemma C.5 by using Zorn's lemma.

Remark C.3. It is simple to show that if there exist lower and upper bounds such that $h_+ = h_-$ and the exact asymptotic behavior is well defined, they are the "definite bound" of theorem 2.1.

Proof - Lemma C.4.

We will prove the two directions of the lemma separately.

Assuming the second condition in equation 39 is satisfied:

Given some $0 < p < 1$, we know using equation 39 that there is some $0 < c$ such that for sufficiently large $n \in \mathbb{N}$:

$$p \leq P\left(\|M_n\| \leq cf\left(n\right)\right)\ . \tag{41}$$

Given some $g \in \mathcal{N}$ such that $f = o(g)$, we know that for sufficiently large $n \in \mathbb{N}$:

$$cf\left(n\right) \leq g\left(n\right)\ , \tag{42}$$

which means that for sufficiently large $n \in \mathbb{N}$:

$$p \leq P\left(\|M_n\| \leq cf\left(n\right)\right) \leq P\left(\|M_n\| \leq g\left(n\right)\right)\ . \tag{43}$$

As we proved that for any $0 < p < 1$ we get that:

$$\lim_{n\to\infty} \left(P\left(\|M_n\| \leq g\left(n\right)\right)\right) = 1\ . \tag{44}$$

And as we proved that for any arbitrary $g \in \mathcal{N}$ such that $f = o(g)$, we proved the first part of the lemma.

Assuming the first condition in equation 38, is satisfied:

If we assume in contradiction that equation 39 is not satisfied, we get that there is some $0 < p < 1$ such as:

$$\forall n_0 \in \mathbb{N}\, 0 < c\, \exists n_0 \leq n \in \mathbb{N} : P\left(\|M_n\| \leq cf\left(n\right)\right) < p\ . \tag{45}$$

In particular that means that if we choose the sequence $\{c_i = i\}_{i=1}^{\infty}$, there are $\tilde{n}_1 < \tilde{n}_2 < \tilde{n}_3 ... \in \mathbb{N}$ such as:

$$\forall i \in \mathbb{N} : P\left(\|M_{\tilde{n}_i}\| \leq i f\left(\tilde{n}_i\right)\right) < p \; . \tag{46}$$

The reason that we can require that $\{\tilde{n}_i\}_{i=1}^{\infty}$ is rising, is that we know that we can find such n-s for any sufficiently large n_0 and for any c. So by induction we can require every time that every $\tilde{n}_i$ is bigger than all previous $\tilde{n}$-s.

Assuming the second condition of equation 39 is satisfied:

Suppose, by contradiction, that equation 38 is not satisfied. Then, there exists some $0 < p < 1$ such that:

$$\forall n_0 \in \mathbb{N}, \, 0 < c, \, \exists n_0 \leq n \in \mathbb{N} : P\left(\|M_n\| \leq c f\left(n\right)\right) < p \; . \tag{47}$$

In particular, if we choose the sequence $\forall i \in \mathbb{N} : c_i = i$, there exist $\tilde{n}_1 < \tilde{n}_2 < \tilde{n}_3 \cdots \in \mathbb{N}$ such that:

$$\forall i \in \mathbb{N} : P\left(\|M_{\tilde{n}_i}\| \leq i f\left(\tilde{n}_i\right)\right) < p \; . \tag{48}$$

Since we can find such n-values for any sufficiently large n_0 and any c, and we can require by induction that each $\tilde{n}_i$ is greater than all previous $\tilde{n}$-values.

We can now define the function:

$$\forall n \in \mathbb{N} : g\left(n\right) = \left(\max\left\{i \in \mathbb{N} \mid \tilde{n}_i \leq n\right\}\right) f\left(n\right) \; . \tag{49}$$

Since $\{\tilde{n}_i\}_{i=1}^{\infty}$ is increasing, we know by the Archimedean property that $\max\left\{i \in \mathbb{N} \mid \tilde{n}_i \leq n\right\}$ is also increasing and unbounded, which implies:

$$\lim_{n \to \infty} \frac{g\left(n\right)}{f\left(n\right)} = \lim_{n \to \infty} \max\left\{i \in \mathbb{N} \mid \tilde{n}_i \leq n\right\} = \infty \; . \tag{50}$$

However, by using equations 46 and 49, we also have:

$$\forall n_0 \in \mathbb{N}, \, \exists n_0 \leq n \in \mathbb{N} : P\left(\|M_n\| \leq g\left(n\right)\right) < p \; , \tag{51}$$

which means that:

$$\lim_{n \to \infty} P\left(\|M_n\| \leq g\left(n\right)\right) \neq 1 \; . \tag{52}$$

This contradicts our assumption in equation 38. Therefore, by reductio ad impossibile, equation 39 must be satisfied, completing the proof for the second direction. □

Proof - Lemma C.5.

For a trivial lower bound, we choose h_- such that $\forall n \in \mathbb{N} : h_-(n) = 0$.

We define h_+ as follows:

$$\forall n \in \mathbb{N} : h_+\left(n\right) = \inf\left\{m \in \mathbb{R} \,\middle|\, 1 - \frac{1}{n} \leq P\left(\|M_n\| \leq m\right)\right\} \; . \tag{53}$$

The infimum and the function are well defined because:

1. The set is well defined.
2. The set is non-empty; if it were empty, it would imply that there is some probability that $\|M\|$, which is a positive number, is larger than any real number, which is impossible.
3. The set is defined with a total order " $<$ " and has a lower bound, $m = 0$.

Since for any $0 < p < 1$, there exists some $n_0 \in \mathbb{N}$ such that:

$$\forall n_0 \leq n \in \mathbb{N} : p \leq P\left(\|M_n\| \leq m\right) , \tag{54}$$

we know that for any $h_+ < g \in \mathcal{N}$, this is also true, which implies:

$$O(M) \leq O\left(h_+\right) , \tag{55}$$

completing the proof. □

Proof - Theorem 2.1.

General Idea of the Proof:

The proof proceeds as follows:

- We consider the set of all upper bounds for M, denoted by $\mathcal{Z}$, and use Zorn's lemma to show that every chain[5] in this set has a lower bound within $\mathcal{Z}$.
- Applying Zorn's lemma again, we demonstrate that $\mathcal{Z}$ has a minimum.
- We then show that the limiting behavior of this minimum is unique.

Existence of an Infimum for the Upper Bound Set:

We begin by defining the set:

$$\mathcal{Z} = \{h \in \mathcal{N} \,| O\left(M\right) \leq O\left(h\right)\} \ . \tag{56}$$

This set is:

1. Well defined.
2. Non-empty (as proven in lemma C.5).
3. Defined with a partial order $h_1 < h_2 \leftrightarrow O\left(h_1\right) < O\left(h_2\right)$.

According to Zorn's lemma, if all chains in this set have a lower bound in $\mathcal{Z}$, then $\mathcal{Z}$ has at least one minimum.

Given some chain in the set, $\mathcal{C} \subseteq \mathcal{Z}$, we know it is lower bounded by the function h_-, which means (by using Zorn's lemma) it has at least one infimum (a lower bound without any larger lower bounds). We will choose such an infimum and denote it by $I \in \mathcal{N}$.

Proving that the Infimum is in $\mathcal{Z}$:

We assume, by contradiction, that this infimum is not in $\mathcal{Z}$, which means there exists some $g \in \mathcal{N}$ such that $I = o(g)$ and for every $0 < p < 1, n_0 \in \mathbb{N}$, there exists $n_0 \leq n \in \mathbb{N}$ such that:

$$P\left(\|M_n\| \leq g\left(n\right)\right) < p \ . \tag{57}$$

Since $I = o(g)$, we know that for any $c \in \mathbb{R}$ and sufficiently large $n \in \mathbb{N}$:

$$cI\left(n\right) \leq g\left(n\right) \ . \tag{58}$$

Combining these equations, we obtain:

$$\forall 0 < c,\, n_0 \in \mathbb{N}\, \exists n_0 \leq n \in \mathbb{N} : P\left(\|M_n\| \leq cI\left(n\right)\right) < p \ . \tag{59}$$

In particular, if we choose the sequence $\forall i \in \mathbb{N} : c_i = i^2$, there exist $\tilde{n}_1 < \tilde{n}_2 < \tilde{n}_3 ... \in \mathbb{N}$ such that:

$$\forall i \in \mathbb{N} : P\left(\|M_{\tilde{n}_i}\| \leq i^2 I\left(\tilde{n}_i\right)\right) < p \ . \tag{60}$$

We can require that $\{\tilde{n}_i\}_{i=1}^{\infty}$ is increasing for the same reason as before, as we know that we can find such arbitrarily large n-values for any sufficiently large n_0 and for any c, so we can, by induction, demand that each $\tilde{n}_i$ is greater than all previous $\tilde{n}_1 ... \tilde{n}_{i-1}$.

Now, we define the function:

$$J\left(n\right) = \begin{cases} iI\left(n\right) : \exists i \in \mathbb{N} : n = \tilde{n}_i \\ \quad I\left(n\right) : \text{else} \end{cases} \ . \tag{61}$$

This function is well defined because there is only one i for any n such that $n = \tilde{n}_i$, as it is an increasing sequences.

[5] A chain, as defined in set theory, is a subset for which the given partial order becomes a total order.

Using equations 60,61, we find that the subsequence $\{\tilde{n}_i\}_{i=1}^{\infty}$ satisfies:

$$\forall i \in \mathbb{N} : P\left(\left\|M\left(q\right)_{\tilde{n}_i}\right\| \leq iJ\left(\tilde{n}_i\right)\right) < p\ . \tag{62}$$

Applying lemma C.4 for the equivalency of the asymptotic bound definition, we conclude that above this subsequence $J \notin \mathcal{Z}$, which implies that above this subsequence J is a lower bound of $\mathcal{Z}$ and consequently, also of $\mathcal{C}$. Moreover, for all other n, we have $J = I$, and since I is a lower bound of $\mathcal{C}$, so is J. Since every $n \in \mathbb{N}$ belongs to one of these subsequences, we conclude that J is a lower bound of $\mathcal{C}$ in general.

Furthermore, for every $i \in \mathbb{N}$ as $1 \leq c_i$, we have:

$$\forall n : I(n) \leq J(n) \rightarrow O(I) \leq O(J)\ . \tag{63}$$

However, since $\{\tilde{n}_i\}_{i=1}^{\infty}$ is increasing and unbounded, we know that there exists at least one subsequence such that:

$$\lim_{\tilde{n}_i \to \infty} \frac{J\left(\tilde{n}_i\right)}{I\left(\tilde{n}_i\right)} = \lim_{i \to \infty} c_i = \infty \rightarrow O\left(J\right) \neq O\left(I\right)\ . \tag{64}$$

This implies:

$$O\left(I\right) < O\left(J\right) \rightarrow I < J\ . \tag{65}$$

We have discovered that J is greater than I, but smaller than all functions in $\mathcal{C}$, which implies that it is a larger lower bound than the infimum, which is impossible! and implies by "reductio ad impossibile" that every chain in $\mathcal{Z}$, has a lower bound in $\mathcal{Z}$.

Existence and Uniqueness of the Minimum:

Using Zorn's lemma, we now know that $\mathcal{Z}$ has at least one minimum, denoted by $f \in \mathcal{N}$. Our remaining task is to show that all other minima in $\mathcal{Z}$ exhibit the same limiting behavior as f, which implies the uniqueness of the minimal limiting behavior.

Let $g \in \mathcal{N}$ be another minimum. We define:

$$\forall n \in \mathbb{N} : h\left(n\right) = \min\left\{f\left(n\right), g\left(n\right)\right\}\ . \tag{66}$$

We know that $h \leq f, g$ (as all of its elements are smaller or equal to those of f, g), and we also know that $h \in \mathcal{Z}$ since $f, g \in \mathcal{Z}$ and for every $0 < p < 1$ we can choose the maximal n_0 from f and g. Thus, $h \in \mathcal{Z}$, but $h \leq f, g$ as well, where f, g are minima themselves. This implies:

$$O\left(f\right) = O\left(h\right) = O\left(g\right) \rightarrow O\left(f\right) = O\left(g\right)\ . \tag{67}$$

Therefore, there exists a unique minimal limiting behavior, which implies that the tensor's asymptotic behavior is always well-defined. □

Remark C.4. In our proof, we employed Zorn's lemma twice. First, we used it to demonstrate the existence of an infimum for every chain, and then, after showing that these infima belong to $\mathcal{Z}$, we employed it again to establish that $\mathcal{Z}$ has a minimum. At first glance, it may seem perplexing that we needed to rely on Zorn's lemma, an incredibly abstract and powerful tool equivalent to the somewhat controversial axiom of choice, to prove that the tensor's asymptotic behavior, which has a much more grounded and intuitive meaning, is well-defined.

One possible explanation for this discrepancy is that we may not have actually required the full power of the axiom of choice, and our structures could be simple enough that an alternative approach could have been taken to prove our theorem without using Zorn's lemma. We believe, however, that in the most general case, Zorn's lemma was indeed necessary, but it was only relevant for extreme distributions lacking any tangible "physical meaning." For any well-defined set of distributions with a clear underlying meaning, one could potentially find an alternative method for demonstrating the existence of a tight bound without invoking Zorn's lemma.

In any case, as we demonstrated in Lemma C.5, there is no need for any of these high-level tools to prove the existence of an upper bound.

C.3 Properties of the Asymptotic Behavior Notation

Having established that our notation is meaningful, we now aim to demonstrate its usefulness. First, we need to address our earlier issue and define "uniform asymptotic bound." Once again, we omit the proofs in this (and next) sections.

Definition C.1 (Uniform Tensors Asymptotic Bound)**.** Given a sequence of random tensors $\left\{M^{(d)}\right\}_{d=1}^{D}$, where $D \in \mathbb{N} \cup \{\infty\}$ (or, more precisely, a sequence of random tensor sequence) with a limiting parameter n, we say that it is uniformly asymptotically upper bounded by $f \in \mathcal{N}$ under some rising monotonic function $\mathcal{K}^{1...D} : \mathbb{R} \to \mathbb{R}$:

$$\forall d = 1...D : O\left(M^{(d)}\right) \leq O\left(\mathcal{K}^{d} \circ f\right) \quad \text{Uniformly} \ , \tag{68}$$

if and only if:

$$\forall g \in \mathcal{N} \; s.t \; f = o\left(g\right) : \lim_{n \to \infty} P\left(\forall d = 1...D : \left\| M_{n}^{(d)} \right\| \leq \mathcal{K}^{d} \circ g\left(n\right)\right) = 1 \ . \tag{69}$$

The definition for a uniform lower asymptotic bound is analogous with reversed directions.

Remark C.5. As discussed in definition 2.1, it is clear that if D is finite, then a uniform bound is equivalent to a point-wise bound.

Lemma C.6 (Asymptotic Notation Inherits its Norm Properties)**.** Given a random tensor M and a sequence of jointly distributed random tensors $\left\{M^{(d)}\right\}_{d=1}^{D}$ (with M as well), where $D \in \mathbb{N} \cup \{\infty\}$, such that they are all uniformly bounded:

$$\forall d = 1...D : O\left(M^{(d)}\right) \leq O\left(\mathcal{K}^{d} \circ f\right) \quad \text{Uniformly} \ , \tag{70}$$

then:

1. If some positive linear combination of $M^{(d)}$'s norms satisfies an inequality of the form:

$$\|M\| \leq \sum_{\tilde{d}=1}^{\tilde{D}} \lambda_{\tilde{d}} \prod_{d=D_{\tilde{d}-1}+1}^{D_{\tilde{d}}} \|M_{d}\| \ , \tag{71}$$

   where all of the coefficients are positive: $\forall d = 1...\tilde{D} : 0 \leq \lambda_{\tilde{d}}$ and we divided $1...D$ into a sequence of finite intervals: $0 = D_1 < D_2 < ... < D_{\tilde{D}} = D$. Then the asymptotic behavior of all the tensors satisfies the same inequality as well for every $h \sim f$:

$$O\left(M\right) \leq O\left(\sum_{\tilde{d}=1}^{\tilde{D}} \lambda_{\tilde{d}} \prod_{d=D_{\tilde{d}-1}+1}^{D_{\tilde{d}}} \mathcal{K}^{d} \circ h\right) \ , \tag{72}$$

   and if the inequality is an equality for the norm, it is also an equality for the "large O-s."

2. Our asymptotic notation inherits all of the properties presented in lemma C.1.

Remark C.6. The lemma still holds even if the tensor have additional indices, as we will see in section (G.4), provided the number of additional index possibilities remains finite in n.

C.4 Exploring the Relationship Between Asymptotic Behavior Notation and the Tensors' Moments

The final aspect of the asymptotic behavior notation we wish to explore is the relationship between this notation and the moments of our tensors' norm or variables. This relationship is relatively intuitive and straightforward, and will be useful in Section (G). We first need to introduce a simple notation for every tensor $M_{\vec{i}}$ that will assist in examining tensor moments, the norm expectation value, defined as:

$$[M] = \sqrt{\frac{1}{N}\left\langle \|M\|^{2} \right\rangle} \ , \tag{73}$$

Lemma C.7 (Asymptotic Behavior and Tensor Moments Equivalency)**.** Given a random tensor M and a function $f \in \mathcal{N}$, then:

$$O(M) \leq O(f) \ , \tag{74}$$

if and only if with probability arbitrarily close to 1:

$$[M] = O\left(f\right) \ . \tag{75}$$

The lemma is also applicable for the uniform bound in the case of infinite number random tensors.

In (4.1), we highlighted that most assertions concerning the convergence of $\mathcal{C}'(F - \hat{y})$ relates to its expected value. However, we can now also associate it with its asymptotic behavior throughout the entire training trajectory. This association stems from the understanding that, if our system exhibits a known average decay, the likelihood of significant deviations from this typical variance range must also decrease, and exponentially (at any decaying rate that is slower then our original rate). Given that decaying geometric sums are convergent, we can infer that the overall probability of the system defying our predicted asymptotic behavior is likewise convergent. Given that we can choose the scaling of this probability arbitrarily, we can set conditions such that the cumulative probability of any deviation is arbitrary small. We introduce this notion for the reader's consideration and propose a detailed formulation as a future exercise.

D Additional Definitions

D.1 Derivatives Correlations Asymptotic Behavior

In our main text (3.2.1), we discussed that the definition for the asymptotic behavior of the derivatives correlations is slightly nuanced, due to the many different potential combinations of distinct inputs. Here we define it rigorously.

Definition D.1 (Derivatives Correlations Asymptotic Behavior)**.** For every $D \in \mathbb{N}^0$, $d \in \mathbb{N}$, and $d_1 \leq d_2 \leq ... \leq d_{\tilde{d}} \in \mathbb{N}$ such that $d_1 + \ldots + d_{\tilde{d}} = d$:

$$O_{d_1 \ldots d_{\tilde{d}}}\left(\mathfrak{C}^{D,d}\right) \equiv O_{x_0, x_1 \ldots x_{\tilde{d}} \in \mathcal{P}}\left(\mathfrak{C}^{D,d}\left(x_0, x_1^{\times d_1} \ldots x_{\tilde{d}}^{\times d_{\tilde{d}}}\right)\right) \ . \tag{76}$$

Inputs order doesn't matter as correlations are symmetric concerning their first derivatives. The factor $\frac{d!}{d_1! \cdots d_{\tilde{d}}!}$ accounts for the possible combinations. If $f \in \mathcal{N}$, we say:

$$\mathfrak{C}^{D,d} = O\left(f\right) \ , \tag{77}$$

if and only if all combinations are uniformly bound by f. In the continuous limit (extended training time), only $d_1 = \ldots = d_d = 1$ remains relevant.

D.2 Properly normalised GDML

Our main theorems (3.1,3.2) and corollary (4.1) are applicable for systems that are properly scaled in the initial condition where $n \to \infty$, defined as follows.

Definition D.2 (PGDML)**.** Given a GDML as described in section 3.1, we will say it is properly normalized and denote it as PGDML if and only if:

$$F\left(\theta_0\right) = O\left(n^0\right) \tag{78}$$

$$\Delta F\left(\theta_0\right) = F\left(\theta\left(1\right)\right) - F\left(\theta_0\right) = O\left(n^0\right) \tag{79}$$

$$\mathfrak{C}^1 = \left(N\eta\right) O\left(\nabla F\left(\theta_0\right)\right)^2 \tag{80}$$

$$\forall d \in \mathbb{N} : O\left(\nabla^{\times d} F\left(\theta_0\right)\right) \leq O\left(\nabla F\left(\theta_0\right)\right)^d \quad \text{Uniformly.} \tag{81}$$

Where n^0 symbolizes n in the power of zero.

The first two conditions (78,79) ensure that our system scale remains finite for the initial condition. Condition 80 stipulates that the asymptotic behavior of the kernel is maximal, given the asymptotic behavior of the first derivative. This condition ensures that our system is genuinely learning and not only memorizing. This is because the kernel for different inputs is responsible for extrapolation, while the kernel with the same input twice responsible for memorization[6]. Condition 81 asserts that none of the higher derivatives dominate the first for $n \to \infty$, a property that most realistic scalable GDMLs satisfy, because if it is not satisfied, gradient descent becomes irrelevant. We show that wide neural networks in general satisfy that property in appendix G.5.

E Proof of Theorems 3.1,3.2

We can now proceed with the proofs of theorems 3.1 and 3.2. The general idea has been outlined at the end of section 3.3.1.

E.1 First Direction of Theorems 3.1,3.2

Now that we understand how to work with the asymptotic behavior of random tensors, we can proceed to prove our main theorems and corollary. We will begin with the first direction of the theorems.

Lemma E.1 (Linearization Requires Weak Correlation)**.**

1. In theorem 3.1, if condition 1 is satisfied, then condition 2 is satisfied as well.
2. In theorem 3.2, if condition 1 is satisfied, then condition 2 is satisfied as well.

Proof. We only demonstrate that the $O_1\left(\mathfrak{C}\right)$ are bounded; The proof the rest are bounded is the same, by considering more learning steps after the initial condition.

For the initial condition, we know that any reparameterization $0 < r$ satisfies (8,20):

$$\begin{gathered} F\left(\theta\left(1\right)\right) - F_{lin}\left(1\right) = \\ \sum_{d=1}^{\infty} \tfrac{(r\eta)^d}{d!} \left(\nabla^{\times d} F\left(\theta_0\right)\left(\nabla F\left(\theta_0\right)\left(x_1\right)^T\right)^{\times d}\right)\left(-\mathcal{C}'\left(F\left(\theta_0\right)\left(x_1\right),\hat{y}\left(x_1\right)\right)\right)^{\times d} - \\ \left(-\left(r\eta\right)\nabla F\left(\theta_0\right)\nabla F\left(\theta_0\right)\left(x_1\right)^T\left(-\mathcal{C}'\left(F\left(\theta_0\right)\left(x_1\right),\hat{y}\left(x_1\right)\right)\right)\right) = \\ \sum_{d=2}^{\infty} r^d \left(\tfrac{\eta^d}{d!}\nabla^{\times d} F\left(\theta_0\right)\left(\nabla F\left(\theta_0\right)\left(x_1\right)^T\right)^{\times d}\right)\left(-\mathcal{C}'\left(F\left(\theta_0\right)\left(x_1\right),\hat{y}\left(x_1\right)\right)\right)^{\times d} = \\ \sum_{d=2}^{\infty} r^d \left(\mathfrak{C}^d\right)^{\cdot, x_1^{\times d}}\left(-\mathcal{C}'\left(F\left(\theta_0\right)\left(x_1\right),\hat{y}\left(x_1\right)\right)\right)^{\times d}\ , \end{gathered} \tag{82}$$

and in the same way for every $D \in \mathbb{N}$:

$$\begin{gathered} \tfrac{(r\eta)^{\frac{D}{2}}}{D!}\nabla^{\times D} F\left(\theta\left(1\right)\right) - \tfrac{(r\eta)^{\frac{D}{2}}}{D!}\nabla^{\times D} F_{lin}\left(\theta\left(1\right)\right) = \\ \sum_{d=1}^{\infty} r^{\frac{D}{2}+d}\left(\mathfrak{C}^{D,d}\right)^{\cdot, x_1^{\times d}}\left(-\mathcal{C}'\left(F\left(\theta_0\right)\left(x_1\right),\hat{y}\left(x_1\right)\right)\right)^{\times d}\ . \end{gathered} \tag{83}$$

Utilizing lemma C.6, it becomes evident that for properly normalized gradient descent-based systems:

$$O\left(\mathfrak{C}^{D,d}\mathcal{C}'\left(F\left(\theta_0\right),\hat{y}\right)^{\times d}\right) \leq O\left(\mathfrak{C}^{D,d}\right) O\left(\mathcal{C}'\left(F\left(\theta_0\right),\hat{y}\right)^{\times d}\right) = O\left(\mathfrak{C}^{D,d}\right)\ . \tag{84}$$

However, since our theorem should work for any $\hat{y}$, we can choose $U = F\left(\theta_0\right) + c$, and obtain:

$$O\left(\mathfrak{C}^{D,d}\mathcal{C}'\left(F\left(\theta_0\right),\hat{y}\right)^{\times d}\right) \propto O\left(\mathfrak{C}^{D,d}\mathcal{C}'\left(c\right)^{\times d}\right) = O\left(\mathfrak{C}^{D,d}\right)\ , \tag{85}$$

as we can choose c such that $\mathcal{C}'(c)$ is the vector that maximizes the correlation, as $\mathcal{C}'$ is convex and the correlations are symmetrical.

Given that we can choose an open set of different scalings of r, we know the different elements in the series cannot cancel each other out. Consequently, for $F - F_{lin}$ to decay, all the distinct elements must decay.

<u>Assuming condition 1 in theorem 3.1:</u>

[6]This is a direct consequence of the NTK equation of motion (8). For example, in the case of a single input point, the system behaves like a memorization algorithm for that one input. However, the term $\Theta(x, x')$ governs how the value of the function at x is influenced by its values at other points x'.

Given that $O\left(F\left(\theta\left(1\right)\right)-F_{lin}\left(1\right)\right) = O\left(\frac{1}{m(n)}\right)$ and for every $D \in \mathbb{N}$ we have $O\left(\eta^{\frac{D}{2}}\nabla^{\times D}F\left(\theta\left(1\right)\right)-\eta^{\frac{D}{2}}\nabla^{\times D}F\left(\theta_0\right)\right) = O\left(\frac{1}{\sqrt{m(n)}}\right)$, it follows that each correlation must decay at least like:

$$\forall 2\leq d\in\mathbb{N}:O\left(\mathfrak{C}^{d}\right)\leq O\left(\frac{1}{m\left(n\right)}\right)\quad \text{Uniformly,} \tag{86}$$

and

$$\forall D,d\in\mathbb{N}:O\left(\mathfrak{C}^{D,d}\right)\leq O\left(\frac{1}{\sqrt{m\left(n\right)}}\right)\quad \text{Uniformly.} \tag{87}$$

This completes the first part of the proof.

Assuming condition 1 in theorem 3.2:

By taking $r(n)$ arbitrarily close to $m(n)$, we find that for $F\left(\theta\left(1\right)\right)-F_{lin}\left(1\right)$ to decay, $r^{d}\mathfrak{C}^{d}$ must decay as well, which implies that:

$$\forall d\in\mathbb{N}:O\left(\mathfrak{C}^{d}\right)\leq O\left(\frac{1}{m\left(n\right)}\right)^{d}, \tag{88}$$

and

$$\forall D\in\mathbb{N}^{0},d\in\mathbb{N}:O\left(\mathfrak{C}^{D,d}\right)\leq O\left(\frac{1}{\sqrt{m\left(n\right)}}\right)^{d}. \tag{89}$$

This concludes our proof.

□

E.2 Second Direction of Theorems 3.1,3.2

We will now prove the other direction of the theorems, focusing on theorem 3.1 since the proofs for the other theorems are essentially the same. It should also be noted that the corollary 4.1, which will be proven next, is almost a generalization of this direction, except that it is only applicable for sufficiently small learning rates.

Lemma E.2 (Asymptotic Behavior Normalization for weakly Correlated PGDML)**.** Consider a weakly correlated PGDML as described in theorems 3.1,3.2 then we have:

$$\forall D\in\mathbb{N}:\eta^{D}O\left(\nabla^{\times D}F\left(\theta_0\right)\right)^{2}\leq O\left(1\right)\quad \text{Uniformly.} \tag{90}$$

With Lemma E.2 at hand, we can now demonstrate the second direction of the theorem by proving a slightly stronger version of it.

Lemma E.3 (Weak Correlations Create Linearization - First Theorem)**.** Assuming the conditions of theorem 3.1 part 1, then for every $s=1...S$:

1. $$O\left(F\left(\theta\left(s\right)\right)-F_{lin}\left(s\right)\right)\leq O\left(\frac{1}{m\left(n\right)}\right). \tag{91}$$

2. $$O\left(\eta^{\frac{1}{2}}\nabla F\left(\theta\left(s\right)\right)-\eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)\right)\leq\gamma. \tag{92}$$

3. For every $2\leq D\in\mathbb{N}$

$$O\left(\eta^{\frac{D}{2}}\nabla^{\times D}F\left(\theta\left(s\right)\right)-\eta^{\frac{D}{2}}\nabla^{\times D}F\left(\theta_0\right)\right)\leq O\left(\frac{1}{\sqrt{m\left(n\right)}}\right)\quad \text{uniformly.} \tag{93}$$

Here, γ is an asymptotic notation such that $\gamma = O\left(\frac{1}{\sqrt{m(n)}}\right)$, and when multiplied with a first derivative of the hypothesis function in its initial condition, it exhibits an asymptotic behavior of $O\left(\gamma_t \eta^{\frac{1}{2}} \nabla F\left(\theta_0\right)\right) \leq O\left(\frac{1}{m(n)}\right)$.

From proving lemmas E.1,E.3, we can conclude that theorems 3.1,3.2 have been proven.

Proof of Lemma E.2.

Assume that the lemma is not satisfied, i.e.,

$$\eta O\left(\nabla F\left(\theta_0\right)\right)^2 \not\leq O\left(1\right) \ , \tag{94}$$

then for some probability $0 < p < 1$, we have:

$$O\left(1\right) < \eta O\left(\nabla F\left(\theta_0\right)\right)^2 \ . \tag{95}$$

Utilizing the third property of PGDML systems (80), we conclude that for some relevant probability:

$$O\left(1\right) < O\left(\mathfrak{C}^1\right) \ . \tag{96}$$

However, for the reasons discussed earlier, the different elements in the equation of motion cannot cancel each other out, as η can be chosen from an open set. This implies that the second property of PGDML systems (79) cannot be satisfied, leading to the conclusion that:

$$\eta O\left(\nabla F\left(\theta_0\right)\right)^2 \leq O\left(1\right) \ , \tag{97}$$

must hold.

By employing the fourth property (81) of PGDML systems, we obtain the desired result.

□

Proof of Lemma E.3.

We will prove the lemma using induction over the learning steps (of course). The induction base for the "zero" step, where $\theta = \theta_0$, is trivial. Assuming the lemma holds for $s \in \mathbb{N}^0$, we observe that for every $\left(D \in \mathbb{N}^0, d \in \mathbb{N}\right) \neq (0,1)$, the d, D correlation satisfies the following for sufficiently small learning rate η:

$$\begin{gathered}
\mathfrak{C}^{D,d}\left(\theta\left(s\right)\right) = \eta^{\frac{D}{2}+d} \nabla^{\times D+d} F\left(\theta\left(s\right)\right)^T \nabla F\left(\theta\left(s\right)\right)^{\times d} \\
= \\
\left(\eta^{\frac{D+d}{2}} \nabla^{\times D+d} F\left(\theta_0\right) + \gamma\right)^T \left(\eta^{\frac{1}{2}} \nabla F\left(\theta_0\right) + \gamma\right)^{\times d} \\
= \\
\mathfrak{C}^{D,d} + \gamma^T \left(\eta^{\frac{1}{2}} \nabla F\left(\theta_0\right)\right)^{\times d} + \gamma^T \left(\gamma \times \left(\eta^{\frac{1}{2}} \nabla F\left(\theta_0\right)\right)^{\times d-1}\right) + \\
\eta^{\frac{D+d}{2}} \nabla^{\times D+d} F\left(\theta_0\right) \left(\gamma \times \left(\eta^{\frac{1}{2}} \nabla F\left(\theta_0\right)\right)^{\times d-1}\right) + \text{comb} + O\left(\frac{1}{m(n)}\right) \\
= \\
\mathfrak{C}^{D,d} + O\left(\frac{1}{m(n)}\right) + O\left(\frac{1}{m(n)}\right) + d\mathfrak{C}^{D+1,d-1} \times \gamma + O\left(\frac{1}{m(n)}\right) \\
= \\
\mathfrak{C}^{D,d} + O\left(\frac{1}{m(n)}\right) \ .
\end{gathered} \tag{98}$$

Here, we used the derivatives correlation definition, lemmas, the induction hypothesis, the bound of the correlations from condition 1, and the definition of γ.

By employing the derivative's correlation definition and condition 1, we observe that:

$$\begin{gathered}
\forall 2 \leq d \in \mathbb{N} : O\left(\mathfrak{C}^d\right) = O\left(\frac{1}{m(n)}\right) \ , \\
\forall d \in \mathbb{N} : O\left(\mathfrak{C}^{1,d}\right) = \gamma \ , \\
\forall 2 \leq D \in \mathbb{N}, d \in \mathbb{N} : O\left(\mathfrak{C}^{D,d}\right) = O\left(\frac{1}{\sqrt{m(n)}}\right) \ .
\end{gathered} \tag{99}$$

Furthermore:

$$\mathfrak{C}^{1,d}\eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)=\eta^{\frac{1}{2}+d}\nabla^{\times d+1}F\left(\theta_0\right)^T\left(\eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)\right)^{\times d}= \\ \eta^{d+1}\nabla^{\times d+1}F\left(\theta_0\right)^T\left(\nabla F\left(\theta_0\right)\right)^{\times d+1}=\mathfrak{C}^{d+1}\ . \tag{100}$$

Hence, using this equation, we can deduce that $\mathfrak{C}^{D,d}\left(\theta\left(s+1\right)\right)$ satisfies the given conditions as well. By incorporating this equation into our equation of motion and employing the lemmas, we find that for a sufficiently small learning rate, $F\left(\theta\left(s+1\right)\right)$ also satisfies the lemma. Consequently, by induction, the lemma holds for all $s\in\mathbb{N}$. ☐

F Proof of Corollary 4.1

In this section, we prove corollary 4.1. The general approach for this proof is akin to that of the first direction of theorems 3.1 and 3.2, albeit with an additional focus on the evolution of the deviation throughout the induction process.

Given the complexity of tracking all the derivatives simultaneously, our strategy involves monitoring the difference between the parameters and their linearization, as expressed in Equation (107). A significant challenge arises in solving the equation of motion that these parameters must satisfy.

To circumvent this issue, we establish a link between this deviation and the deviation of the generalization function from its linearization (107) up to the highest order, as outlined in equation 112. By considering only the lowest order terms, we obtain an equation of motion (119). In cases where the cost function decays exponentially, and we are able to bound the deviation of this equation.

F.1 Relations between Different Linearizations

In the main text, we linearised F as F_{lin} (8), by first considering only the linear part of F, and then examining how it changes over time for a given training path. However, there are alternative ways to linearise F that can be useful to consider. One such method involves taking only the linear part of F, without considering the training path:

$$\hat{F}(\theta)=F\left(\theta_0\right)+\nabla F\left(\theta_0\right)^T\left(\theta-\theta_0\right)\ . \tag{101}$$

Another useful definition is to examine how θ would develop over time under the linear approximation for our training path:

$$\theta_{lin}\left(0\right)=\theta_0\quad\forall s\in\mathbb{N}: \\ \theta_{lin}\left(s+1\right)=\theta_{lin}\left(s\right)-\nabla F\left(\theta_0\right)\left(x_s\right)\mathcal{C}'\left(F_{lin}\left(s\right)\left(x_s\right)-\hat{y}\left(x_s\right)\right)\ . \tag{102}$$

It can be observed that $F_{lin},\hat{F},\theta_{lin}$ satisfy the following relation:

$$\forall s\in\mathbb{N}^0: F_{lin}\left(s\right)=\hat{F}\left(\theta_{lin}\left(s\right)\right)\ . \tag{103}$$

A more refined relation is the one between $F\left(\theta_{lin}\right)$ and $F_{lin}\left(\theta\right)$, defined for every $s=0...S$ as follows:

$$O\left(F\left(\theta_{lin}(s)\right)-F_{lin}\left(s\right)\right)\leq O\left(\frac{\varrho^2\left(s\right)}{m\left(n\right)}\right)\ , \tag{104}$$

where ϱ is defined as:

Definition F.1 (Typical Linear Cumulative Deviation)**.** We define the typical linear cumulative deviation as the bound of the cumulative deviation of F_{lin} from $\hat{y}$:

$$O\left(\varrho\left(s\right)\right)=\sum_{s'=0}^{s-1}O\left(\mathcal{C}'\left(F_{lin}\left(s'\right)-\hat{y}\right)\right)\ , \tag{105}$$

and in our case:

$$O\left(\varrho\left(s\right)\right)\leq O\left(\frac{1-e^{-\frac{s}{T}}}{1-e^{-\frac{1}{T}}}\right)\leq O\left(1\right)\ . \tag{106}$$

This implies that $\varrho(s)=o(m(n))$, which is essential for proving (104). We will not provide this proof here, as we will not use it directly in the remainder of this paper, and we will soon prove many similar identities.

F.2 Small Perturbation from the Linear Solution

The initial approach of the proof aimed to demonstrate that F only deviates slightly from F_{lin}, and that also its derivatives deviate slightly at the initial conditions. The intention was to use induction to show that this holds at each time step. This method is effective if the goal is merely to prove that F converges to F_{lin} at a rate of $O\left(\frac{1}{m(n)}\right)$ for a fixed time step. However, it poses challenges when attempting to understand how the two functions deviate from each other over time. This is due to the necessity of simultaneously tracking the evolution of all derivatives and the changes in correlations over time, which is nearly impossible.

To circumvent this issue, rather than tracking all derivatives, we will calculate how $F\left(\theta\left(s\right)\right)$ deviates from $F_{lin}\left(s\right)$ by utilizing a similar relationship to the one we discovered between θ_{lin} and F_{lin}. This will allow us to establish bounds on $F - F_{lin}$. Although the two approaches are equivalent, and the first one is more intuitively clear, the second approach simplifies accurate and simple calculations by focusing on a single object, $F - F_{lin}$.

In the following lemma, we demonstrate how a small perturbation at a given step ($s = 0...S-1$) results in a small perturbation at the subsequent step ($s+1$). Then, we will use these results to inductively show the deviation in time between the hypothesis function and its linear approximation.

We denote:

$$\delta(s) = F\left(\theta\left(s\right)\right) - F_{lin}\left(s\right)\ ,\ \eta^{\frac{1}{2}}\zeta(s) = \theta(s) - \theta_{lin}(s)\ , \tag{107}$$

and assume that the deviation from linearity is small, hence:

$$O\left(\delta(s)\right) \leq O\left(\frac{f(s)}{m(n)}\right)\ ,\ O\left(\zeta\left(s\right)\right) \leq O\left(g\left(s\right)\right)\gamma\ , \tag{108}$$

where

$$f\left(s\right), g\left(s\right)^2, \varrho\left(s\right)^2 = o\left(m\left(n\right)\right)\ . \tag{109}$$

For some parts of our lemma, it will also be relevant to separate the deviation of the parameters into two components:

$$\zeta\left(s\right) = \zeta_\gamma\left(s\right) + \zeta_m\left(s\right)\ , \tag{110}$$

such that:

$$O\left(\zeta_\gamma\left(s\right)\right) \leq O\left(g_\gamma\left(s\right)\right)\gamma\, ,\ O\left(\zeta_m\left(s\right)\right) \leq O\left(\frac{g_m\left(s\right)}{m\left(n\right)}\right)\ . \tag{111}$$

Remark F.1. Here, we consider the case of a general rate of convergence for $\mathcal{C}'\left(F_{lin}, \hat{y}\right)$, rather than exclusively focusing on an exponential one. This is done to simplify the generalization of our results for reader.

Remark F.2. In the following lemma and its proof, we use the symbol "$\simeq$" to denote higher-order terms of the expressions. This is justified by our assumption that we are working within the framework of analytic functions, where the sum of all higher-order terms still converges.

Lemma F.1 (Deviation of the parameters and of the hypothesis function relations)**.** Given the conditions described above, then up to the leading order:

1. $$\begin{gathered}\delta\left(s\right) = F\left(\theta\left(s\right)\right) - F_{lin}\left(s\right) \simeq \eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)^T\zeta_m\left(s\right) + \eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)^T\zeta_\gamma\left(s\right) + \\ \textstyle\sum_{s_1,s_2=0}^{s-1}\mathfrak{C}^2\mathcal{C}'\left(F_{lin}\left(s_1\right), \hat{y}\right) \times \mathcal{C}'\left(F_{lin}\left(s_2\right), \hat{y}\right) + \\ 2\textstyle\sum_{s'=0}^{s-1}\mathfrak{C}^{1,1}\zeta_\gamma\left(s\right)\mathcal{C}'\left(F_{lin}\left(s'\right), \hat{y}\right) + \eta\nabla^{\times 2}F\left(\theta_0\right)^T\zeta_\gamma\left(s\right)^{\times 2}\ ,\end{gathered} \tag{112}$$

   which means:

   $$\begin{gathered}O\left(\delta\left(s\right)\right) = O\left(F\left(\theta\left(s\right)\right) - F_{lin}\left(s\right)\right) \leq \\ O\left(\frac{g_m(s)}{m(n)}\right) + O\left(\frac{(g_\gamma(s)+\varrho(s))^2}{m(n)}\right) \leq O\left(\frac{(g(s)+\varrho(s))^2}{m(n)}\right)\ .\end{gathered} \tag{113}$$

2. $$O\left(\eta^{\frac{1}{2}}\nabla F\left(\theta\left(s\right)\right)^T - \eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)^T\right) \leq O\left(g\left(s\right) + \varrho\left(s\right)\right)\gamma\, . \tag{114}$$

3.
$$\mathcal{C}'\left(F\left(\theta\left(s\right)\right),\hat{y}\right)-\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\simeq\mathcal{C}''\left(F_{lin}\left(s\right),\hat{y}\right)\delta\left(s\right)\ . \tag{115}$$
where, $\mathcal{C}''\left(F_{\text{lin}}(s),\hat{y}\right)$ denotes a positive random matrix such that, if the asymptotic behavior of $\mathcal{C}'\left(F_{\text{lin}}(s),\hat{y}\right)$ is bounded, then $\mathcal{C}''\left(F_{\text{lin}}(s),\hat{y}\right)$ is bounded as well (as is in our setting).

4.
$$\begin{gathered}\eta^{\frac{1}{2}}\zeta\left(s+1\right)-\eta^{\frac{1}{2}}\zeta\left(s\right)=\theta\left(s+1\right)-\theta_{lin}\left(s+1\right)-\eta^{\frac{1}{2}}\zeta\left(s\right)\simeq\\ -\eta\nabla F\left(\theta_0\right)\mathcal{C}''\left(F_{lin}\left(s'\right),\hat{y}\right)\delta\left(s\right)+O\left(g\left(s\right)+\varrho\left(s\right)\right)\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\eta^{\frac{1}{2}}\gamma\ ,\end{gathered} \tag{116}$$

which means:
$$O\left(\zeta\left(s+1\right)-\zeta\left(s\right)\right)\leq O\left(\frac{f\left(s\right)}{m\left(n\right)}\right)+O\left(\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\right)O\left(g\left(s\right)+\varrho\left(s\right)\right)\gamma\ . \tag{117}$$

5.
$$\begin{gathered}O\left(\delta\left(s+1\right)-\delta\left(s\right)+\Theta_0\mathcal{C}''\left(F_{lin}\left(s\right),\hat{y}\right)\delta\left(s\right)\right)\leq\\ O\left(\tfrac{\left(g(s)+\varrho(s)\right)^2}{m(n)}\right)O\left(\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\right)\ .\end{gathered} \tag{118}$$

Remark F.3. An important note for our proofs is that all of these components can be generalized to the case where $\zeta(s),\delta(s)$ are not the "original" deviations, as long as they satisfy equation 108.

We can now use this result to prove corollary 4.1 by induction. In fact, for the conditions of the corollary at $s=0$, the induction hypothesis is trivially satisfied as $F(\theta)(0)=F_{lin}(0),\theta(0)=\theta_{lin}(0)$. It is straightforward to show that the contributions of the part multiplied by $O\left(\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\right)$ are irrelevant for the possible deviation, as $\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\to 0\ ,\ O\left(\varrho\left(s\right)\right)\leq O\left(1\right)$. Consequently, we are left with equations of motion for the asymptotic behavior of the form:

$$O\left(\zeta\left(s+1\right)-\zeta\left(s\right)\right)\leq O\left(\frac{f\left(s\right)}{m\left(n\right)}\right)\quad\delta\left(s+1\right)-\delta\left(s\right)+\Theta_0\mathcal{C}''\left(F_{lin}\left(s\right),\hat{y}\right)\delta\left(s\right)\simeq 0\ . \tag{119}$$

However, $\Theta_0,\mathcal{C}''$ are positively defined bound matrices, so for a learning rate that is sufficiently small (which would be of the same order of magnitude as the learning rate needed for our system to consistently learning, and for the case where $\mathcal{C}(x)=\frac{1}{2}x^2$, exactly the same), we find that on average this term can only contribute to the shrinkage of $\delta(s)$. This means that neglecting this term for large s would provide an upper bound for the rate of deviation. Thus, we have discovered that the asymptotic behavior of δ (and consequently, ζ) with respect to time is for large s is bounded by:
$$\delta\left(s+1\right)-\delta\left(s\right)\simeq 0\ . \tag{120}$$

This proves our corollary.

***Proof*.**

Part - (1):

$$\begin{gathered}F\left(\theta\left(s\right)\right)=F\left(\theta_{lin}\left(s\right)+\eta^{\frac{1}{2}}\zeta\left(s\right)\right)=_1\\ F\left(\theta_0-\eta\sum_{s'=0}^{s-1}\nabla F\left(\theta_0\right)\mathcal{C}'\left(F_{lin}\left(s'\right),\hat{y}\right)+\eta^{\frac{1}{2}}\zeta\left(s\right)\right)=_2\\ F\left(\theta_0\right)-\sum_{s'=0}^{s-1}\mathfrak{C}^1\mathcal{C}'\left(F_{lin}\left(s'\right),\hat{y}\right)+\eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)^T\zeta\left(s\right)+\\ \sum_{s_1,s_2=0}^{s-1}\mathfrak{C}^2\mathcal{C}'\left(F_{lin}\left(s_1\right),\hat{y}\right)\times\mathcal{C}'\left(F_{lin}\left(s_2\right),\hat{y}\right)+\\ 2\sum_{s'=0}^{s-1}\mathfrak{C}^{1,1}\zeta\left(s\right)\mathcal{C}'\left(F_{lin}\left(s'\right),\hat{y}\right)+\eta\nabla^{\times 2}F\left(\theta_0\right)^T\zeta\left(s\right)^{\times 2}+\ldots\simeq_3\\ F_{lin}\left(s\right)+\eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)^T\zeta\left(s\right)+\sum_{s_1,s_2=0}^{s-1}\mathfrak{C}^2\mathcal{C}'\left(F_{lin}\left(s_1\right),\hat{y}\right)\times\mathcal{C}'\left(F_{lin}\left(s_2\right),\hat{y}\right)+\\ 2\sum_{s'=0}^{s-1}\mathfrak{C}^{1,1}\zeta\left(s\right)\mathcal{C}'\left(F_{lin}\left(s'\right),\hat{y}\right)+\eta\nabla^{\times 2}F\left(\theta_0\right)^T\zeta\left(s\right)^{\times 2}\ ,\end{gathered} \tag{121}$$

where in (1) we used equation (102) the definition of θ_{lin}, in (2) we expanded our generalization function as a Taylor sires, and the definition of the derivatives correlations (3.1). In (3) we used the fact that under our assumptions our system is exponentially weakly correlated. Using this result we get our desired identity.

Subtracting F_{lin} we get using the weak derivatives correlations property that up to the leading order:

$$\begin{gathered}F\left(\theta\left(s\right)\right)-F_{lin}\left(s\right)\simeq\\ \eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)^T\zeta\left(s\right)+\sum_{s_1,s_2=0}^{s-1}\mathfrak{C}^2\mathcal{C}'\left(F_{lin}\left(s_1\right),\hat{y}\right)\times\mathcal{C}'\left(F_{lin}\left(s_2\right),\hat{y}\right)+\\ 2\sum_{s'=0}^{s-1}\mathfrak{C}^{1,1}\zeta\left(s\right)\mathcal{C}'\left(F_{lin}\left(s'\right),\hat{y}\right)+\eta\nabla^{\times2}F\left(\theta_0\right)^T\zeta\left(s\right)^{\times2}\simeq\\ \eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)^T\zeta_m\left(s\right)+\eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)^T\zeta_\gamma\left(s\right)+\\ \sum_{s_1,s_2=0}^{s-1}\mathfrak{C}^2\mathcal{C}'\left(F_{lin}\left(s_1\right),\hat{y}\right)\times\mathcal{C}'\left(F_{lin}\left(s_2\right),\hat{y}\right)+\\ 2\sum_{s'=0}^{s-1}\mathfrak{C}^{1,1}\zeta_\gamma\left(s\right)\mathcal{C}'\left(F_{lin}\left(s'\right),\hat{y}\right)+\eta\nabla^{\times2}F\left(\theta_0\right)^T\zeta_\gamma\left(s\right)^{\times2}=\\ O\left(\frac{g_m(s)+g_\gamma(s)}{m(n)}\right)+O\left(\frac{\varrho(s)^2}{m(n)}\right)+2O\left(\frac{\varrho(s)g_\gamma(s)}{m(n)}\right)+O\left(\frac{g_\gamma^2(s)}{m(n)}\right)=\\ O\left(\frac{g_m(s)}{m(n)}\right)+O\left(\frac{(g_\gamma(s)+\varrho(s))^2}{m(n)}\right)\leq O\left(\frac{(g(s)+\varrho(s))^2}{m(n)}\right)\ ,\end{gathered}\tag{122}$$

which finishes our proof.

Part 2:

Using the same ideas we get:

$$\begin{gathered}\eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)^T=\eta^{\frac{1}{2}}\nabla F\left(\theta_{lin}\left(s\right)+\eta^{\frac{1}{2}}\zeta\left(s\right)\right)^T=\\ \eta^{\frac{1}{2}}\nabla_T F\left(\theta_0-\eta\sum_{s'=0}^{s-1}\nabla F\left(\theta_0\right)\mathcal{C}'\left(F_{lin}\left(s'\right),\hat{y}\right)+\eta^{\frac{1}{2}}\zeta\left(s\right)\right)=\\ \eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)^T-\sum_{s'=0}^{s-1}\mathfrak{C}^{1,1}\mathcal{C}'\left(F_{lin}\left(s'\right),\hat{y}\right)+\eta\nabla^{\times2}F\left(\theta_0\right)^T\zeta\left(s\right)+\ldots=\\ \eta^{\frac{1}{2}}\nabla F\left(\theta_0\right)^T+O\left(\varrho\left(s\right)\right)\gamma_t+O\left(g\left(s\right)\right)\gamma_t\ .\end{gathered}\tag{123}$$

Taking transpose on both sides we get finish our proof.

Part 3:

Using the definition of δ and the fact that $\mathcal{C}$ is analytical we know that up to the highest order:

$$\mathcal{C}'\left(F\left(\theta\left(s\right)\right),\hat{y}\right)=\mathcal{C}'\left(F_{lin}\left(s\right)+\delta\left(s\right),\hat{y}\right)\simeq\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)+\mathcal{C}''\left(F_{lin}\left(s\right),\hat{y}\right)\delta\left(s\right)\tag{124}$$

and as $\mathcal{C}$ is convex (3), we know that it's second derivative is always a positive matrix. And that if the first derivative is bound, so is the second one.

Part 4:

Using the equation of motion for θ (102), and parts 2,3 of this lemma we get that up to leading order:

$$\begin{gathered}\theta\left(s+1\right)=\theta\left(s\right)-\eta\nabla F\left(\theta\left(s\right)\right)\mathcal{C}'\left(F\left(\theta\left(s\right)\right),\hat{y}\right)\simeq\\ \theta\left(s\right)-\eta\begin{pmatrix}\nabla F\left(\theta_0\right)+\\ O\left(g\left(s\right)+\varrho\left(s\right)\right)\eta^{\frac{1}{2}}\gamma\end{pmatrix}\begin{pmatrix}\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)+\\ \mathcal{C}''\left(F_{lin}\left(s'\right),\hat{y}\right)\delta\left(s\right)\end{pmatrix}\simeq\\ \theta\left(s\right)-\eta\nabla F\left(\theta_0\right)\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)-\eta\nabla F\left(\theta_0\right)\mathcal{C}''\left(F_{lin}\left(s'\right),\hat{y}\right)\delta\left(s\right)+\\ O\left(g\left(s\right)+\varrho\left(s\right)\right)\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\eta^{\frac{1}{2}}\gamma\end{gathered}\tag{125}$$

and as:

$$\begin{gathered}\theta\left(s\right)-\eta\nabla F\left(\theta\left(s\right)\right)\mathcal{C}'\left(F\left(\theta\left(s\right)\right),\hat{y}\right)=\\ \theta_{lin}\left(s\right)+\eta^{\frac{1}{2}}\zeta\left(s\right)-\eta\nabla F\left(\theta\left(s\right)\right)\mathcal{C}'\left(F\left(\theta\left(s\right)\right),\hat{y}\right)=\theta_{lin}\left(s+1\right)+\eta^{\frac{1}{2}}\zeta\left(s\right)\ ,\end{gathered}\tag{126}$$

we get the desired result.

Part 5:

Using the equation of motion for θ, one can see that:

$$\begin{gathered}
F\left(\theta\left(s+1\right)\right)=F\left(\begin{array}{c}\theta_{lin}\left(s+1\right)-\eta\nabla F\left(\theta_{0}\right)\mathcal{C}''\left(F_{lin}\left(s'\right),\hat{y}\right)\delta\left(s\right)+\\ \eta^{\frac{1}{2}}\zeta\left(s\right)+O\left(g\left(s\right)+\varrho\left(s\right)\right)\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\eta^{\frac{1}{2}}\gamma\end{array}\right)\\
\simeq_{1}\\
F_{lin}\left(s+1\right)-\eta\nabla F\left(\theta_{0}\right)^{T}\nabla F\left(\theta_{0}\right)\mathcal{C}''\left(F_{lin}\left(s'\right),\hat{y}\right)\delta\left(s\right)+\\
\eta^{\frac{1}{2}}\nabla F\left(\theta_{0}\right)^{T}\zeta\left(s\right)+O\left(g\left(s\right)+\varrho\left(s\right)\right)\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\eta^{\frac{1}{2}}\nabla F\left(\theta_{0}\right)^{T}\gamma+\\
\textstyle\sum_{s_{1},s_{2}=0}^{s-1}\mathfrak{C}^{2}\mathcal{C}'\left(F_{lin}\left(s_{1}\right),\hat{y}\right)\times\mathcal{C}'\left(F_{lin}\left(s_{2}\right),\hat{y}\right)+\\
2\textstyle\sum_{s'=0}^{s-1}\mathfrak{C}^{1,1}\zeta\left(s\right)\mathcal{C}'\left(F_{lin}\left(s'\right),\hat{y}\right)+\\
2O\left(g\left(s\right)+\varrho\left(s\right)\right)\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\textstyle\sum_{s'=0}^{s-1}\mathfrak{C}^{1,1}\gamma\mathcal{C}'\left(F_{lin}\left(s'\right),\hat{y}\right)+\\
\eta\nabla^{\times2}F\left(\theta_{0}\right)^{T}\zeta\left(s\right)^{\times2}+O\left(g\left(s\right)+\varrho\left(s\right)\right)^{2}\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)^{2}\eta\nabla^{\times2}F\left(\theta_{0}\right)^{T}\gamma^{\times2}+\\
2O\left(g\left(s\right)+\varrho\left(s\right)\right)\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\eta\nabla^{\times2}F\left(\theta_{0}\right)^{T}\left(\gamma\times\zeta\left(s\right)\right)\\
\simeq_{2}\\
F_{lin}\left(s+1\right)-\Theta_{0}\mathcal{C}''\left(F_{lin}\left(s'\right),\hat{y}\right)\delta\left(s\right)+\delta\left(s\right)+\\
2O\left(\tfrac{g(s)+\varrho(s)}{m(n)}\right)O\left(\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\right)+O\left(\tfrac{(g(s)+\varrho(s))^{2}}{m(n)}\right)O\left(\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\right)^{2}+\\
2O\left(\tfrac{g^{2}(s)+\varrho(s)g(s)}{m(n)}\right)O\left(\mathcal{C}'\left(F_{lin}\left(s\right),\hat{y}\right)\right)\ .
\end{gathered}\tag{127}$$

where in (1) We use part 1 of the lemma, when we remembered that $O(\delta)\leq O\left(\frac{1}{m(n)}\right)$ so it can be consider as ζ_m. In part (2) we use the definition of F_{lin}, Θ_0 and part 1 once again where we gathered all of the components that have only $\zeta(s)$ to get $\delta(s)$. Then we just used the asymptotic behavior of all of the components and took the "worst case scenario" to get equation 118. □

G WIDE NEURAL NETWORKS ARE WEAKLY CORRELATED PGDML SYSTEMS

G.1 GENERAL IDEA

We start with fully connected neural networks. Although the proof is technically intricate, its underlying concept is straightforward: For the first layer, we observe that all higher correlations exhibit the appropriate asymptotic behavior. We then proceed to prove by induction that all layers manifest the same asymptotic behavior. Consider the second correlation, for instance, which we analyze as follows:

For any general layer $l=1,\ldots,L$, defining ∇_{-l} as the derivatives with respect to parameters from layers 1 to $l-1$ (G.2), we employ the equation for fully connected neural networks (133):

$$\begin{gathered}
l=0,\ldots,L:F^{(l)}=\theta^{(l,l-1)}\phi\left(F^{(l-1)}\right)+\theta^{(l)}\ ,\\
\forall x\in X:F(\theta)\left(x\right)=F^{(L)}\left(x\right),\quad F^{(0)}\left(x\right)=a\ ,
\end{gathered}\tag{128}$$

to demonstrate that:

$$\begin{gathered}
\nabla_{(-l)}^{\times2}F^{(l)}=\nabla_{(-l)}^{\times2}\left(\theta^{(l,l-1)}\phi\left(F^{(l-1)}\right)+\theta^{(l)}\right)=\\
\nabla_{(-l)}\times\nabla_{(-l)}\left(\theta^{(l,l-1)}\phi\left(F^{(l-1)}\right)+\theta^{(l)}\right)=\nabla_{(-l)}\times\left(\theta^{(l,l-1)}\nabla_{(-l)}\phi\left(F^{(l-1)}\right)\right)=\\
\nabla_{(-l)}\times\left(\theta^{(l,l-1)}\phi'\left(F^{(l-1)}\right)\nabla_{(-l)}F^{(l-1)}\right)=\\
\theta^{(l,l-1)}\phi''\left(F^{(l-1)}\right)\nabla_{(-l)}F^{(l-1)}\times\nabla_{(-l)}F^{(l-1)}+\theta^{(l,l-1)}\phi'\left(F^{(l-1)}\right)\nabla_{(-l)}^{\times2}F^{(l-1)}
\end{gathered}\tag{129}$$

Consequently, the contribution to the l-th correlation (10) from this part is proportional to:

$$\theta^{(l,l-1)}\phi''\left(F^{(l-1)}\right)\mathfrak{C}_{(l-1)}^{1}\times\mathfrak{C}_{(l-1)}^{1}+\theta^{(l,l-1)}\phi'\left(F^{(l-1)}\right)\mathfrak{C}_{(l-1)}^{2}\ .\tag{130}$$

Here we have two terms. We can show the right-hand term is small simply by induction. The proof that the left-hand term is also small is more complex, involving the demonstration that for all hidden layers, the relevant contribution from the first correlation originates from its diagonal terms, i.e., $(\mathfrak{C}^1_{(-l)})_{ii}$.

We can now show that in the term, the left index is identical for both correlations, which follows that for most indices, the relevant terms are offset by the irrelevant ones, keeping our expression small.

For the case that one of the derivative does not belong to layers $l = 1$ to $l - 1$, we explicitly show this term to be negligible, as for most indices it simply resets:

$$\nabla_{i^l i^{l-1}} F_i^{(l)} \propto \delta_{i^l i} \tag{131}$$

In the general case of the D-th correlation, while there is some complexity in tracing the combinatorial terms from various combinations of derivatives, the fundamental principle remains consistent.

The generalization of this approach for other architectures is discussed in Section G.5.

G.2 Asymptotic Behavior of Wide FCN at Initialisation

Remark G.1. Throughout this paper we considered $\|M\|$ or $O(M)$ as our way to evaluate the size of our random tensors. But here we mainly consider the normalised terms instead:

$$\frac{1}{\sqrt{N}} \|M\| \quad \text{and} \quad \frac{1}{\sqrt{N}} O\left(M\right) \ . \tag{132}$$

This is because, in practice, what we are interested of is the average asymptotic behavior of a tensor, and not the accumulative one.

Fully connected neural networks of depth $2 \leq L \in \mathbb{N}$, characterized by L parameter vectors (the biases $\theta^{(1)}, \ldots, \theta^{(L)}$), and L parameter matrices (the weights $\theta^{(L,L-1)}, \ldots, \theta^{(1,0)}$), such as:

$$\begin{aligned} l = 0, ..., L : F^{(l)} = \theta^{(l,l-1)} \phi\left(F^{(l-1)}\right) + \theta^{(l)} \ , \\ \forall x \in X : F(\theta)\left(x\right) = F^{(L)}\left(x\right), \quad F^{(0)}\left(x\right) = a \ . \end{aligned} \tag{133}$$

In this representation, $F^{(0)}, F^{(1)}, \ldots, F^{(L-1)}$, and $F^{(L)}$ constitute the input, inner, and output layers, respectively. The activation function ϕ is analytical, and all of its derivatives are bounded as described in (23).

Remark G.2. Generally when working with FCNN we do not operate the activation function over the zero layer, the input. But to make the induction slightly easier, we will simplify our expression such as ϕ operates over all layers. It makes no real difference

We focus on "wide" neural networks where the depth L is fixed. As long as $L = O(\log(n))$, we can expect an NTK-like behavior for large n, but for simplicity, we focus on the scenario where L remains constant in n. We introduce a limiting parameter $n \in \mathbb{N}$ such that the width of all the hidden layers satisfies $n \leq n_1, \ldots, n_{L-1}$. To simplify our work, we will amend this assumption by postulating that all layers exhibit the same asymptotic behavior of n - $n_1, \ldots, n_{L-1} \sim n$. This modification does not affect our theorems and lemmas, as it merely establishes a lower bound of our original assumption. As the sizes of the zeroth and last layer are constant (the dimensions of the input and output layers stay fixed in n of course), we arrive at:

$$n_1, \ldots, n_{L-1} \sim n \quad \text{and} \quad n_0, n_L \sim 1 \ . \tag{134}$$

Back in the 1960s, it was demonstrated that with Gaussian initialization, we can keep our layers normalised by selecting initial parameters as follows:

$$\forall l = 1, \ldots, L : \theta_0^{(l,l-1)} \sim \mathcal{N}\left(0, \frac{1}{n_l}\right), \, \theta_0^{(l)} \sim \mathcal{N}\left(0, 1\right) \ . \tag{135}$$

Despite the specificity of this initialization algorithm, it contradicts the broader spirit of this paper. It's not only overly restrictive but also complicates our work by colliding with our framework of tensor's asymptotic behavior. Rather than focusing on a particular initialization scheme like the normal distribution, we will identify and utilize the relevant properties inherent in the distribution.

Definition G.1 (Appropriate Initialization scheme for Wide Neural Networks)**.** Given a wide neural network as defined above, we characterize the distribution for the initial condition θ as appropriate if and only if for every probability arbitrarily close to 1, the following properties hold:

1. Different elements of θ are independent. And for each layer $l = 1, ..., L$, $\theta^{(l,l-1)}$'s and $\theta^{(l)}$'s elements share the same distribution.
2. θ is symmetric around 0 (implying that all odd moments are nullified):
$$\forall D \in \mathbb{N} \setminus 2\mathbb{N} : \left\langle \theta^{\cdot D} \right\rangle = 0 \ . \tag{136}$$
3. For every layer $l = 1, ..., L$, all moments of θ are uniformly normalized:
$$\forall D \in \mathbb{N} : \begin{matrix} O\left(1\right)^{D} \leq \frac{1}{\sqrt{n_l}} O\left(\left(\theta^{(l)}\right)^{\cdot D}\right) \leq D! O\left(1\right)^{D} \ , \\ O\left(\frac{1}{\sqrt{n_{l-1}}}\right)^{D} \leq \frac{1}{\sqrt{N_l}} O\left(\left(\theta^{(l,l-1)}\right)^{\cdot D}\right) \leq D! O\left(\frac{1}{\sqrt{n_{l-1}}}\right)^{D} \ , \end{matrix} \quad \text{Uniformly} \tag{137}$$
where $N_l = n_l n_{l-1}$ is the total number of parameters in the l-th layer.

where the elemental tensor power defined such as:
$$\forall D \in \mathbb{N} : \left(M^{\cdot D}\right)_{\vec{i}} = M_{\vec{i}}^{D} \ . \tag{138}$$

The first two conditions ensure that our system is unbiased, while the third condition guarantees that our system will not be dominated by a disproportionate probabilistic "tail."

We delegate to the reader the verification that Gaussian initialization qualifies as an appropriate initialization.

Remark G.3. Conditions 1,2 can be generalized to be fulfilled in the limit of large n, provided this convergence occurs rapidly enough. Nevertheless, any complexities arising from this generalization are technical and do not affect our analysis.

For the remainder of this section, we will omit the biases from our discussion, as they do not add any substantial insights or implications for the points under consideration and won't change any of our results.

Lemma G.1 (Normalization of Layers in Proper Wide Neural Networks)**.** Given a wide neural network, if the initial condition is appropriately set, then all the moments across every layer $l = 1...L$ are well normalized:
$$\frac{1}{\sqrt{n_l}} O\left(F^{(l)}\right) = O\left(1\right) \ . \tag{139}$$

The final parameter that we need to normalize in our system is the dynamic one - the learning rate, denoted by η. In an attempt to generalize Gaussian initialization, we will adopt the standard method of normalization for η:
$$\eta \sim \frac{1}{n} \ . \tag{140}$$

This condition, coupled with the demand for an appropriate initialization strategy, is sufficient to demonstrate that wide neural networks are exponentially weakly correlated PGDML-s.

In the remainder of this section, we will proceed under the assumption that our parameters are initialized appropriately and that $\eta \sim \frac{1}{n}$.

We can now use this result to find the asymptotic behavior of the layers derivatives:

Lemma G.2 (Asymptotic Behavior of Layer's Derivatives)**.** Given our established conditions and initialisation, all derivatives are uniformly bound for each natural number D and layer $l = 1...L$. Specifically, we have:
$$\frac{\eta^{\frac{D}{2}}}{\sqrt{N_D}} O\left(\nabla^{\times D} F^{(l)}\right) \leq O\left(1\right) \quad \text{Uniformly} \tag{141}$$

Here, $N_D = n_l n_{l-1}^{D} n^{D}$ represents the asymptotic behavior of the number of elements in the derivatives.

Proof of lemma G.1.

We approach the proof by induction, across the entire proof we use lemma C.7 to show equivalence between the asymptotic behavior of the system and its tensorial average (73). It is known that the base case, the zeroth layer, naturally satisfies the lemma. By inductive assumption, let us presume that the $l-1$ layer adheres to the lemma. Our task is to establish the lemma's validity for the l-th layer for all $l = 1...L$:

$$\begin{gathered}\left[F^{(l)}\right]^2 =_1 \tfrac{1}{n_l}\sum_i \left\langle \left(\sum_j \theta_{ij}^{(l,l-1)} F_j^{(l-1)}\right)\left(\sum_k \theta_{ik}^{(l,l-1)} F_k^{(l-1)}\right)\right\rangle = \\ \tfrac{1}{n_l}\sum_i \left\langle \left(\sum_{j,k} \theta_{ij}^{(l,l-1)}\theta_{ik}^{(l,l-1)} F_j^{(l-1)} F_k^{(l-1)}\right)\right\rangle =_2 \\ \tfrac{1}{n_l}\sum_i \sum_{j,k}\left\langle \theta_{ij}^{(l,l-1)}\theta_{ik}^{(l,l-1)}\right\rangle \left\langle F_j^{(l-1)} F_k^{(l-1)}\right\rangle = \\ \tfrac{1}{n_l}\sum_i \sum_{j\neq k}\left\langle \theta_{ij}^{(l,l-1)}\theta_{ik}^{(l,l-1)}\right\rangle \left\langle F_j^{(l-1)} F_k^{(l-1)}\right\rangle + \tfrac{1}{n_l}\sum_i\sum_j \left\langle\left(\theta_{ij}^{(l,l-1)}\right)^2\right\rangle\left\langle\left(F_j^{(l-1)}\right)^2\right\rangle =_3 \\ \sum_i\sum_j \tfrac{1}{n_l}\left\langle\left(\theta_{ij}^{(l,l-1)}\right)^2\right\rangle\left\langle\left(F_j^{(l-1)}\right)^2\right\rangle =_4 \sum_{i,j}\tfrac{1}{n_l}\left\langle\left(\theta_{ij}^{(l,l-1)}\right)^2\right\rangle \sum_k \tfrac{1}{n_{l-1}}\left\langle\left(F_k^{(l-1)}\right)^2\right\rangle =_5 \\ n_{l-1}\left[\theta^{(l,l-1)}\right]^2\left[F^{(l-1)}\right]^2 =_6 O\left(1\right)O\left(1\right) = O\left(1\right)\ .\end{gathered} \tag{142}$$

Throughout these equalities, we rely on the premise of a proper initialization. Specifically:

- In "1" and "5", we employ the structure of neural networks and the definition of the moment's norm.
- In "2" and "4", we note that $F^{(l-1)}$ is dependent only on the inner parameters of l, which are independent of $\theta^{(l,l-1)}$. This is enabled by the proper initialization ensuring $\theta^{(l,l-1)}$ is uniformly distributed.
- In "3", we invoke the fact that different elements of $\theta^{(l,l-1)}$ are independent and symmetric. Hence, for every $i, j \neq k$:

$$\left\langle \theta_{ij}^{(l,l-1)}\theta_{ik}^{(l,l-1)}\right\rangle = \left\langle \theta_{ij}^{(l,l-1)}\right\rangle\left\langle \theta_{ik}^{(l,l-1)}\right\rangle = 0\ . \tag{143}$$

- In "6", we apply the induction hypothesis and observe that for a proper initialization (G.1-3):

$$\forall l = 1...L : \left[\theta^{(l,l-1)}\right] = O\left(\frac{1}{\sqrt{n_{l-1}}}\right)\ . \tag{144}$$

Through the application of the principle of mathematical induction, we conclude the lemma holds for all $l = 1...L$.

Using lemma C.7 again, we get that $O\left(F^l\right) \leq O\left(1\right)$, but as we know that even if we neglect a small part of the probability distribution the proof should still hold, we get that:

$$O\left(F^l\right) = O\left(1\right)\ . \tag{145}$$

exactly.

□

Proof of lemma G.2. Given ω, drawn from another proper initialisation, we can observe that $\theta+\omega$ is also properly initialised or sub-properly initialised. Hence, assuming we initialise $F^{(l)}$ accordingly, we find:

$$\frac{1}{\sqrt{n_l}}O\left(F^{(l)}\right) \leq O\left(1\right)\ . \tag{146}$$

Since $F^{(l)}$ is analytical, we can apply its Taylor expansion around θ to get:

$$\frac{1}{\sqrt{n_l}}O\left(\sum_{D=0}^{\infty}\nabla^{\times D}F^{(l)}\left(\theta\right)\omega^{\times D}\right) \leq O\left(1\right)\ . \tag{147}$$

By continuously rescaling ω without violating the proper property, we see that all of the component of the expression must be uniformly bounded:

$$\forall D \in \mathbb{N} : \frac{1}{\sqrt{n_l}} O\left(\nabla^{\times D} F^{(l)}(\theta)\,\omega^{\times D}\right) \leq O(1) \quad \text{Uniformly.} \tag{148}$$

This is because, all the terms are scaled differently by $omega$, meaning that the only way to ensure the expression remains bounded under any finite scaling of ω is to bound each of its terms separately and uniformly.

Considering the symmetry of the derivative in its components, and by invoking lemma C.3, we can identify a vector of size 1 that maximises it, yielding a vector with a size equal to its norm. By setting ω as this vector and rescaling it to be proper, we obtain using lemma C.7 that:

$$\forall D \in \mathbb{N} : \frac{1}{\sqrt{n_l}} O\left(\nabla^{\times D} F^{(l)}(\theta)\right) = \frac{1}{\sqrt{n_l}} \frac{1}{\sqrt{n_{l-1}^D}} O\left(\nabla^{\times D} F^{(l)}(\theta)\,\omega^{\times D}\right) \leq O(1) \quad \text{Uniformly.} \tag{149}$$

Given that:

$$\frac{1}{\sqrt{n_l}} \frac{1}{\sqrt{n_{l-1}^D}} = \frac{1}{\sqrt{n_l n_{l-1}^D}} \sim \frac{\eta^{\frac{D}{2}}}{\sqrt{n_l n_{l-1}^D n^D}} = \frac{\eta^{\frac{D}{2}}}{\sqrt{N_D}}\,, \tag{150}$$

we arrive at the desired result. □

G.3 Representation of the Network's Layers as a Composition of Previous Layer Components

In this part we use the semilinear structure of wide neural network to establish a linear relation between the correlations of the l-th layer to the one of the $l-1$ layer. We will then use this relation next part to show by induction the correlations are weak. For that will define the following useful notation:

Definition G.2 (Inner and Outer Derivatives)**.** Given a layer $l = 1...L$. We denote the l-th layer's outer parameters, which includes its weights (and biases), as follows:

$$\theta^{(l,l-1)}_{i^l,i^{l-1}}\,. \tag{151}$$

Meanwhile, the inner parameters are defined as any of the weights (and biases) from the layers spanning $1...l-1$, and are denoted by:

$$\theta \in \theta^{(-l)}\,. \tag{152}$$

Following the same notation, we denote the gradient of the outer parameters as $\nabla_{(l)}$, and the gradient of the inner parameters as $\nabla_{(-l)}$. The same applies for the correlations, denoted as $\mathfrak{C}_{(l)}, \mathfrak{C}_{(-l)}$.

Remark G.4. It is important to note that, as $F^{(l-1)}$ depends only in the inner parameters of the l-th layer, the following relationship holds:

$$\nabla_{(-l)} F^{(l-1)} = \nabla F^{(l-1)}\,. \tag{153}$$

This notation can be employed to express the derivative of the l-th layer as a combination of derivatives from the $l-1$-th layer.

Lemma G.3 (Representation of the l-th layer derivative, as a combination of its previous layer's derivatives)**.** Given a fully connected wide neural network as specified above, for each $l = 1...L$ layer, the $D \in \mathbb{N}$-th derivative can be presented as follows:

1. When all the derivatives are inner, the expression is:

$$\left(\nabla^{(-l)}\right)^{\times D} F^{(l)} = \theta^{(l,l-1)} \tilde{\nabla}^{\times D} F^{(l-1)}\,. \tag{154}$$

2. When one derivative is outer, and the rest are inner, the expression becomes:

$$\nabla^{(l)}_{i_l i_{l-1}} \times \left(\nabla^{(-l)}\right)^{\times D-1} F^{(l)}_i = \delta_{i i_l} \tilde{\nabla}^{\times D-1} F^{(l-1)}_{i_{l-1}}\,. \tag{155}$$

3. When $2 \leq D$, and for $2 \leq d \in \mathbb{N} \leq D$ where the derivatives are outer, the expression simplifies to:
$$\left(\nabla^{(l)}\right)^{\times d} \times \left(\nabla^{(-l)}\right)^{\times D-d} F^{(l)} = 0\ . \tag{156}$$

Here, $\tilde{\nabla}^{\times D} F^{(l-1)}$ is the compound derivative, defined such as for $D \in \mathbb{N}$:
$$\tilde{\nabla}^{\times D} F^{(l)} = \sum_{d=1}^{D} \sum_{d_1 \ldots d_d \in \mathbb{N}}^{d_1+\ldots+d_d=D} \phi^{[d]}\left(F^{(l)}\right)\left(\nabla^{\times d_1} F^{(l)} \times \cdots \times \nabla^{\times d_d} F^{(l)}\right) + \text{comb} \tag{157}$$

and for $D = 0$:
$$\tilde{\nabla}^{\times 0}_{ij} F^{(l)}_k = \delta_{ik} \phi\left(F_j\right)\ . \tag{158}$$

The "comb" term refers to all possible combinations of the derivatives' indices. For instance, if we consider one term of the third derivative as follows:
$$\theta^{(l,l-1)}\left(\phi^{[2]}\left(F^{(l-1)}\right)\left(\nabla F^{(l-1)} \times \nabla^{\times 2} F^{(l-1)}\right)\right) \tag{159}$$

then, for every three distinct derivative indices $\alpha_1, \alpha_2, \alpha_3$, there are three unique ways to arrange the indices, disregarding irrelevant parts:
$$\nabla_{\alpha_1} F^{(l-1)} \times \nabla^{\times 2}_{\alpha_2\alpha_3} F^{(l-1)}, \nabla_{\alpha_2} F^{(l-1)} \times \nabla^{\times 2}_{\alpha_1\alpha_3} F^{(l-1)}, \nabla_{\alpha_3} F^{(l-1)} \times \nabla^{\times 2}_{\alpha_1\alpha_2} F^{(l-1)}\ . \tag{160}$$

While the first combination naturally arises from our expression, the "comb" term accounts for the other two.

It should be mentioned that only unique terms are counted, even if they originate from different orders of the derivatives. Therefore, for another component of the third derivative, $\theta^{(l,l-1)}\left(\phi^{[3]}\left(F^{(l-1)}\right) \nabla F^{(l-1)} \times \nabla F^{(l-1)} \times \nabla F^{(l-1)}\right)$, and distinct $\alpha_1, \alpha_2, \alpha_3$:
$$\nabla_{\alpha_1} F^{(l-1)} \nabla_{\alpha_2} F^{(l-1)} \nabla_{\alpha_3} F^{(l-1)}, \nabla_{\alpha_1} F^{(l-1)} \nabla_{\alpha_3} F^{(l-1)} \nabla_{\alpha_2} F^{(l-1)} \ldots \tag{161}$$

are identical, hence should only be counted once.

We can use this result to construct the l-th layer correlations using the correlations from the $l-1$ layer:

Lemma G.4 (Representation of the l-th layer correlations, as a combination of its previous layer's correlations)**.** Given the same condition as in lemma G.3, then:
$$\begin{gathered}\mathfrak{C}^{D,d}_{(l)} = \theta^{(l,l-1)} \times \left(\tilde{\theta}^{(l,l-1)}\right)^{\times d} \tilde{\mathfrak{C}}^{D,d}_{(l-1)} + \\ \eta^{\frac{1}{2}} I \times \eta^{\frac{1}{2}} \phi\left(F^{(l-1)}\right) \times \left(\tilde{\theta}^{(l,l-1)}\right)^{\times d-1} \tilde{\mathfrak{C}}^{D,d-1}_{(l-1)} + \text{comb} + \\ \left(\tilde{\theta}^{(l,l-1)}\right)^{\times d} \hat{\mathfrak{C}}^{D-1,d}_{(l-1)} + \text{comb}\ .\end{gathered} \tag{162}$$

or when showing the indices explicitly, using Einstein's notation for summation:
$$\begin{gathered}\left(\mathfrak{C}^{D,d}_{(l)}\right)_{i_0 i_1 \ldots i_d} = \theta^{(l,l-1)}_{i_0 j_0} \tilde{\theta}^{(l,l-1)}_{i_1 j_1} \cdots \tilde{\theta}^{(l,l-1)}_{i_d j_d} \left(\tilde{\mathfrak{C}}^{D,d}_{(l-1)}\right)_{j_0, j_1 \ldots j_d} + \\ \eta^{\frac{1}{2}} \delta_{i_0 i_1} \eta^{\frac{1}{2}} \phi\left(F^{(l-1)}_{j_0}\right) \tilde{\theta}^{(l,l-1)}_{i_2 j_2} \cdots \tilde{\theta}^{(l,l-1)}_{i_d j_d} \left(\tilde{\mathfrak{C}}^{D,d-1}_{(l-1)}\right)_{j_0, j_2 \ldots j_d} + \text{comb} + \\ \tilde{\theta}^{(l,l-1)}_{i_1 j_1} \cdots \tilde{\theta}^{(l,l-1)}_{i_d j_d} \left(\hat{\mathfrak{C}}^{D-1,d}_{(l-1)}\right)_{i_0, j_1 \ldots j_d}\ ,\end{gathered} \tag{163}$$

where the "comb" term includes all index pairings with the zero index, i.e., $(i_0, i_2) \ldots (i_0, i_D)$, and the θ defined as:
$$\tilde{\theta}^{(l,l-1)}_{ij} = \theta^{(l,l-1)}_{ij} \phi'\left(F^{(l-1)}_j\right)\ . \tag{164}$$

The first compound derivative defined such as for $D \in \mathbb{N}_0, d \in \mathbb{N}$:
$$\tilde{\mathfrak{C}}^{D,d}_{(l)} = \sum_{d'=1}^{D+d} \left\{ C_{\vec{d},\vec{D}} \phi^{[d']}\left(F^{(l)}\right) \mathfrak{C}^{D_1,d_1}_{(l)} \times \cdots \times \mathfrak{C}^{D_{d'},d_{d'}}_{(l)} \;\middle|\; \begin{array}{c} d_1 + \ldots + d_{d'} = d \\ D_1 + \ldots + D_{d'} = D \end{array} \right\} + \text{Comb} \tag{165}$$

where:

$$C_{\vec{d},\vec{D}} = \frac{(D_1! \cdots D_{d'}!)\,(d_1! \cdots d_{d'}!)}{D!d!}\ . \tag{166}$$

Also for $D \in \mathbb{N}_0, d = 0$:

$$\tilde{\mathfrak{C}}^{D,0}_{(l)} = \eta^{\frac{D}{2}} \tilde{\nabla}_t^{\times D} F^{(l)}\ . \tag{167}$$

The second compound derivative defined such as for $D \in \mathbb{N}, d \in \mathbb{N}$:

$$\left(\hat{\mathfrak{C}}^{D-1,d}_{(l)}\right)^{\alpha_{d+1}\ldots\alpha_{d+D}}_{i_0,j_1\ldots j_d} = \eta^{\frac{1}{2}} \delta^{\alpha_{d+1}}_{(i_0 j_0)} \left(\tilde{\mathfrak{C}}^{D-1,d}_{(l)}\right)^{\alpha_{d+2}\ldots\alpha_{d+D}}_{j_0,j_1\ldots j_d} + \text{comb}\ , \tag{168}$$

where the "comb" term is defined as before. For $D = 0$ this compound derivative vanishes.

Remark G.5. For the following lemma and the subsequent section, we make the assumption that $D << n$. This assumption is permissible even though, in considering the limit, the limit of D should technically be taken prior to that over n. This is because higher order derivatives typically exert a decreasing influence over system behavior, leading us to essentially consider them negligible beyond a certain point.

It is important to note that this assumption is not strictly necessary. We could directly address the intricate combinatorial factors without it. Despite this, we prefer to make this assumption to avoid introducing unnecessary complications into our analysis.

Lemma G.5 (Counting combinations of the derivatives and correlations)**.**

1. For the conditions of lemma G.3, for every $d_1 \ldots d_d$, the number of combinations of the derivatives indices is:
$$\frac{1}{d!} \frac{D!}{d_1! \cdots d_d!}\ , \tag{169}$$
and the total number of combinations above all possible $d = 1 \ldots D$-s is the D-th "bell number" (which is very close to $D!$).

2. For the conditions of lemma G.4, for every $d_1 \ldots d_{d'}$ and $D_1 \ldots D_{d'}$, the number of combinations of the compound correlations is:
$$\frac{1}{d'!} \frac{d!}{d_1! \cdots d_{d'}!} \frac{D!}{D_1! \cdots D_{d'}!}\ . \tag{170}$$

We assume for this lemma the indices are different, as $D \ll n$.

Proof - lemmas G.3,G.4.

We will prove the lemma by induction for a general layer $l = 1 \ldots L-1$ starting with $l = 1$.

The induction base is simple, a this is a direct consequence of taking a derivative over our equation for neural networks (133). This calculation hinges on the concept that, by definition, the inner derivatives are independent of the outer parameters.

$$\begin{gathered}\nabla_{(l)} F^{(l)} = \nabla_{(l)} \theta^{(l,l-1)} \phi\left(F^{(l-1)}\right) = \theta^{(l,l-1)} \nabla_{(l)} \phi\left(F^{(l-1)}\right) = \\ \theta^{(l,l-1)} \left(\phi^{[1]}\left(F^{(l-1)}\right) \nabla_{(l)} F^{(l-1)}\right)\ ,\end{gathered} \tag{171}$$

which gives us the induction base.

Assuming by induction our lemma is satisfied for some $D - 1 \in \mathbb{N}$: the inner D-th derivative satisfies:

$$\begin{gathered}\nabla^{\times D}_{(-l)} F^{(l)} = \nabla_{(-l)} \times \nabla^{\times D-1}_{(-l)} F^{(l)} = \\ \nabla_{(-l)} \times \theta^{(l,l-1)} \sum_{d=1}^{D-1} \sum_{d_1 \ldots d_d \in \mathbb{N}}^{d_1+\ldots+d_d=D-1} \phi^{[d]}\left(F^{(l-1)}\right)\left(\nabla^{\times d_1} F^{(l-1)} \times \cdots \times \nabla^{\times d_d} F^{(l-1)}\right) \\ +\text{comb} \\ = \\ \theta^{(l,l-1)} \sum_{d=1}^{D-1} \sum_{d_1 \ldots d_d \in \mathbb{N}}^{d_1+\ldots+d_d=D-1} \nabla \times \phi^{[d]}\left(F^{(l-1)}\right)\left(\nabla^{\times d_1} F^{(l-1)} \times \cdots \times \nabla^{\times d_d} F^{(l-1)}\right) \\ + \\ \theta^{(l,l-1)} \sum_{d=1}^{D-1} \sum_{d_1 \ldots d_d \in \mathbb{N}}^{d_1+\ldots+d_d=D-1} \phi^{[d]}\left(F^{(l-1)}\right)\left(\nabla \times \nabla^{\times d_1} F^{(l-1)} \times \cdots \times \nabla^{\times d_d} F^{(l-1)}\right) \\ + \\ \text{comb}\end{gathered} \tag{172}$$

We have here a sum of two two different summations, we will analyse each one separably:

Starting from the first one:

$$\begin{gathered}\sum_{d=1}^{D-1}\sum_{d_1\ldots d_d\in\mathbb{N}}^{d_1+\ldots+d_d=D-1}\nabla\times\phi^{[d]}\left(F^{(l-1)}\right)\left(\nabla^{\times d_1}F^{(l-1)}\times\cdots\times\nabla^{\times d_d}F^{(l-1)}\right)\\ =\\ \sum_{d=1}^{D-1}\sum_{d_1\ldots d_d\in\mathbb{N}}^{d_1+\ldots+d_d=D-1}\phi^{[d+1]}\left(F^{(l-1)}\right)\left(\nabla F^{(l-1)}\times\nabla^{\times d_1}F^{(l-1)}\times\cdots\times\nabla^{\times d_d}F^{(l-1)}\right)\\ =\\ \sum_{d=1}^{D-1}\sum_{d_1=1,d_2\ldots d_{d+1}\in\mathbb{N}}^{d_1+d_2+\ldots+d_{d+1}=D}\phi^{[d+1]}\left(F^{(l-1)}\right)\left(\nabla^{\times d_1}F^{(l-1)}\times\nabla^{\times d_2}F^{(l-1)}\times\cdots\times\nabla^{\times d_{d+1}}F^{(l-1)}\right)\\ =\\ \sum_{d=2}^{D}\sum_{d_1=1,d_2\ldots d_d\in\mathbb{N}}^{d_1+d_2+\ldots+d_d=D}\phi^{[d]}\left(F^{(l-1)}\right)\left(\nabla^{\times d_1}F^{(l-1)}\times\nabla^{\times d_2}F^{(l-1)}\times\cdots\times\nabla^{\times d_d}F^{(l-1)}\right)\ .\end{gathered}\tag{173}$$

The second term can be represented as:

$$\begin{gathered}\sum_{d=1}^{D-1}\sum_{d_1\ldots d_d\in\mathbb{N}}^{d_1+\ldots+d_d=D-1}\phi^{[d]}\left(F^{(l-1)}\right)\left(\nabla\times\nabla^{\times d_1}F^{(l-1)}\times\cdots\times\nabla^{\times d_d}F^{(l-1)}\right)\\ =\\ \sum_{d=1}^{D-1}\sum_{d_1\ldots d_d\in\mathbb{N}}^{(d_1+1)+\ldots+d_d=D}\phi^{[d]}\left(F^{(l-1)}\right)\left(\nabla^{\times d_1+1}F^{(l-1)}\times\cdots\times\nabla^{\times d_d}F^{(l-1)}\right)\\ =\\ \sum_{d=1}^{D-1}\sum_{1<d_1\in\mathbb{N},d_2\ldots d_d\in\mathbb{N}}^{d_1+\ldots+d_d=D}\phi^{[d]}\left(F^{(l-1)}\right)\left(\nabla^{\times d_1}F^{(l-1)}\times\cdots\times\nabla^{\times d_d}F^{(l-1)}\right)\ .\end{gathered}\tag{174}$$

Combining the two sums we get exactly the form that we were searching for, which finishes the proof of the lemma's first case.

Lemma (G.4) is a direct result. □

Proof - lemma (G.5).

Proving the lemma's first part:

The number of way to sort into d distinct sets with $d_1 \ldots d_d$ objects is:

$$\frac{(d_1+\cdots d_d)!}{d_1!\cdots d_d!}=\frac{D!}{d_1!\cdots d_d!}\ ,\tag{175}$$

but our sets are not distinct, so we need to divide by the appropriate coefficient. But if the sets are not the same, they repeat in different arrangements, so we get the $\frac{1}{d!}$. summing over all of these options we get the definition of the D-th bell number.

The second part is the same. □

G.4 Wide FCNNs are Weakly Correlated PGDML Systems

Here we will show a detailed heuristic proof of why wide neural networks are weakly correlated PGDML as described in lemma G.2.

Remark G.6. For this section we assume that the width of the last layer, i.e the L-th layer is exactly $L = 1$. That won't impact any of our results of the system asymptotic behavior as L is fixed in n as discussed in remark C.6.

Remark G.7. In the entire section we will use Einstein's summation notation (liberally).

We initiate our exploration of wide neural network correlations (and derivatives) by focusing on the most critical one - the kernel - $\mathfrak{C}^1$.

For the final layer $l = L$, the kernel norm is simply expressed as:

$$\left\|\mathfrak{C}_{(L)}\right\|=\left|\mathfrak{C}_{(L)}\right|\ .\tag{176}$$

Given that $n_L = 1$, the kernel is merely a scalar.

Leveraging lemma G.4, we can construct the L-th layer kernel from the components of the preceding layer:

$$\mathfrak{C}^1_{(L)}=\theta_i^{(L,L-1)}\theta_j^{(L,L-1)}\left(\mathfrak{C}^1_{(L-1)}\right)_{ij}+\eta\phi\left(F_j^{(L-1)}\right)^2\ .\tag{177}$$

Applying lemma (G.1) and the Lipschitz property of ϕ, we discern that the right term has the asymptotic behavior of $\eta\phi\left(F_j\right)^2 \sim O\left(1\right)$. Concerning the left term, lemma (G.4) once again provides:

$$\left(\mathfrak{C}^1_{(L-1)}\right)_{ij} = \theta^{(L,L-1)}_{ip}\theta^{(L,L-1)}_{jq}\left(\mathfrak{C}^1_{(L-2)}\right)_{pq} + \delta_{ij}\eta\phi\left(F^{(L-2)}_k\right)^2 . \tag{178}$$

This means we have an $O(1)$ term and another that depends on the previous term. Continuing this process by induction and employing the fact that everything is symmetric, hence positive, we conclude that the kernel's asymptotic behavior is precisely $O(1)$. **In combination with (G.2), we find that our system satisfies the criteria of a PGDML (D.2)!**

Let's now consider a general $D \in \mathbb{N}_0, d \in \mathbb{N}$ final correlation. By invoking lemma C.3, we know that there exists a vector $v \in S_N$ achieving the norm:

$$\left\|\mathfrak{C}^{D,d}_{(L)}\right\| = \left|\mathfrak{C}^{D,d}_{(L)} \cdot v^{\times D}\right| . \tag{179}$$

Applying lemma G.4, we find that this expression can be constructed from $D-1$ correlations. Considering only the first term among the three in the equation, (the treatment for others would be the same), and focusing solely on the first correlations, we obtain (up to $\frac{1}{d!}$ when omitting the $\frac{1}{D!}$ as we do not consider the different combinations):

$$\left(\phi^{[d+D]}\left(F^{(L-1)}\right)\left(\theta^{(L,L-1)}\right)\right) \times \left(\tilde{\theta}^{(L,L-1)}\right)^{\times d} \cdot \left(\left(\mathfrak{C}_{(L-1)}\right)^{\times d} \times \left(\eta^{\frac{1}{2}}\nabla F^{(L-1)}\right)^{\times D} \cdot v^{\times D}\right) . \tag{180}$$

Using (23), and that the $L-1$ layer and L are independent at initialization, we can dismiss the ϕ-s, leaving the asymptotic behavior unchanged (we would discuss the $d!$ later):

$$\left(\theta^{(L,L-1)}\right)^{\times d+1} \cdot \left(\left(\mathfrak{C}_{(L-1)}\right)^{\times d} \times \left(\eta^{\frac{1}{2}}\nabla F^{(L-1)}\right)^{\times D} \cdot v^{\times D}\right) . \tag{181}$$

When constructing the kernels from the preceding layer, as each one consists of two terms (177), resulting in 2^d terms in total. This factor of 2^d does not alter the system's asymptotic behavior, so instead, we can consider only the maximal terms, which are the ones with only one kind of first correlation terms. We will choose the first kind of terms, dealing with the others via induction:

$$\begin{gathered}\theta^{(L,L-1)}_{i_0}\theta^{(L,L-1)}_{i_1}\cdots\theta^{(L,L-1)}_{i_d}\left(\delta_{i_0 i_1}\eta\phi\left(F^{(L-2)}_k\right)^2\right)\cdots\left(\delta_{i_0 i_d}\eta\phi\left(F^{(L-2)}_k\right)^2\right)\cdot \\ \left(\left(\eta^{\frac{1}{2}}\nabla F^{(L-2)}_{i_0}\right)^{\times D} \cdot v^{\times D}\right) .\end{gathered} \tag{182}$$

As $\eta\phi\left(F^{(L-2)}_k\right)^2 \sim O(1)$, after reducing the deltas, we obtain an asymptotic behavior of at most:

$$\left(\theta^{(L,L-1)}_i\right)^{d+1}\left(\left(\eta^{\frac{1}{2}}\nabla F^{(L-2)}_{i_0}\right)^{\times D} \cdot v^{\times D}\right) . \tag{183}$$

Now, as we already have that $O\left(\eta^{\frac{1}{2}}\nabla F^{(L-2)}_{i_0}\right) \leq O\left(1\right)$, if $D \in \mathbb{N}$ we find multiplied by a vector of at most size $O(1)$. In the worst case, this object will have an asymptotic behavior of:

$$\left(\theta^{(L,L-1)}_i\right)^{d+2} . \tag{184}$$

We know from our proper initialization that it is uniformly bounded for all d-s by:

$$d! O\left(\frac{1}{\sqrt{n}}\right)^d . \tag{185}$$

which means that by reintroducing the $\frac{1}{d!}$ we get:

$$O\left(\frac{1}{\sqrt{n}}\right)^d . \tag{186}$$

If $D = 0$ however, the $\left(\eta^{\frac{1}{2}} \nabla F_{i_0}^{(L-2)}\right)^{\times D} \cdot v^{\times D}$ term disappears and we are left with:

$$\left(\theta_i^{(L,L-1)}\right)^{d+1} . \tag{187}$$

For odd d-s, we still have $O\left(\frac{1}{\sqrt{n}}\right)^d$ as θ is symmetric. However, for even ones, we find:

$$O\left(\frac{1}{\sqrt{n}}\right)^{d-1} . \tag{188}$$

This explains why, while our system is $\sqrt{n}$ weakly and power correlated. Nonetheless, for the time deviation, one can easily confirm that this term remains negligible as $n \to \infty$.

Of course, there are many other terms rather then the first derivatives ones. But they can be treated similarly.

Assuming that for $l-1$ layer:

$$\phi^{[d']}\left(F^{(l-1)}\right) \mathfrak{C}_{(l-1)}^{D_1,d_1} \times \cdots \times \mathfrak{C}_{(l-1)}^{D_{d'},d_{d'}} . \tag{189}$$

contributes at most:

$$O\left(\frac{1}{\sqrt{n}}\right)^{d \text{ or } d-1} . \tag{190}$$

We get utilizing lemma G.5, and replacing $\phi^{[d']}\left(F^{(l-1)}\right) \to d'!$ (as warranted by equation 23), we find that the total contribution is bounded by:

$$\begin{gathered} \sum_{d'=1}^{D+d} \sum \sum \frac{1}{d'!} \frac{d!}{d_1! \cdots d_{d'}!} \frac{D!}{D_1! \cdots D_{d'}!} d'! \frac{d_1! \cdots d_{d'}!}{d!} \frac{D_1! \cdots D_{d'}!}{D!} O\left(\frac{1}{\sqrt{n}}\right)^{d \text{ or } d-1} \\ \sim 2^{D+d} O\left(\frac{1}{\sqrt{n}}\right)^{d \text{ or } d-1} \sim O\left(\frac{1}{\sqrt{n}}\right)^{d \text{ or } d-1} . \end{gathered} \tag{191}$$

In a similar vein, it can be demonstrated that multiple correlations taken together exhibit the same behavior at the l-th layer. Which means that we can prove by induction in the same way we did for the first correlations, that all of them behave the same, **thereby concluding our (heuristic) proof**.

G.5 Generalization beyond FCNNs

G.5.1 Tensor Programs

While FCNNs are the prototypical network architecture, numerous other architectures are utilized practice as we discussed in section 4.2. The tensor programs formalism, as detailed in Yang & Littwin (2021), offers a unified language to encapsulate most relevant neural network architectures, by viewing them as a composites of global linear operations and pointwise nonlinear functions. This formalism encompasses an extensive array of neural network architectures, including recurrent neural networks and attention-based networks. In their work they demonstrated that any wide network described by this formalism exhibit linearization.

Our weak correlation approach naturally aligns with the tensor programs framework, simplifying the proof that such networks not only exhibit linearization, but also are low correlated PGDMLs. This comes with all of the additional implications that, like deviations over learning and the influence of network augmentation on the linearization rate.

Our proof for FCNNs can be simply generalised for any wide network described by this formalism, because, similarly to FCNNs, all such systems exhibit a wide semi-linear form by definition.

G.5.2 Beyond Tensor Programs

Given the broad generality of the tensor programs formalism, it's challenging to devise linearizing networks that fall outside its scope. However, here we suggest two network-based architectures that demonstrate linearization and, to our belief, stand outside this formalism.

The first is FCNN as outlined in equation 133, but where each neuron possesses a unique activation function:

$$F_i^{(l)} = \sum_{j=1}^{n_{l-1}} \theta_{ij}^{(l,l-1)} \phi_j \left(F_j^{(l-1)} \right) + \theta_i^{(l)} \ . \tag{192}$$

The proof of the linearization of this system, assuming ϕ_i satisfies condition 23, simply parallels our proof for FCNNs.

Not all such systems are outside the random tensor formalism's purview, if we can represent ϕ_i as a function of two distinct inputs - F_i and another external input given by the index $j \in \mathbb{N}$, such as:

$$\forall j = 1...n_{l-1} : \phi_j \left(F_j^{(l-1)} \right) = \phi \left(F_j^{(l-1)}, j \right) \ . \tag{193}$$

However, since ϕ and all its derivatives must remain bounded by some polynomial to fit within the theorems of Yang & Littwin (2021); Yang (2020) for wide neural networks, if ϕ_i is exceedingly diverse, pinpointing a suitable ϕ could be very challenging or even impossible.

A more definite (albeit synthetic) example of a linearizing network-based system outside the tensor programs realm can be formulated as:

$$z(x) = \sum_{i=1}^{n} \theta_i f_i(x) + \sum_{i,j=1}^{n} \theta_i \theta_j g_i(x) g_j(x) \quad g = Af \ , \tag{194}$$

initialized by $\theta = 0$, where A is a $90°$ rotating matrix across the relevant axis as $n \to \infty$, and f_i are chosen as the eigenfunctions of some external kernel.

This system can be viewed just an NTK approximation, but with a non-trivial second derivative that is perpendicular to the first. Hence, our system will still behave linearly as $n \to \infty$. It's also not evident how this system can be derived from the tensor programs framework.

While one might contend that this example seems artificially contrived to the point of limiting its significance, it underscores the existence of weakly correlated, network-based systems that are not encapsulated by the tensor programs formalism.

Furthermore, in line with our discourse in section 3.3.3, if we manage to discern the types of effective correlations that could prove advantageous, such systems might find practical applications.

H The Chicken and The Egg - Elaborated

In this section, we aim to elaborate on the points made in section 3.3.3. We begin by discussing why we argue that the derivatives correlations represent a form of bias in the system.

The simplest way to see the equivalence between weak derivative correlations and an inherent bias within the system, is by considering the case of wide neural networks. In our demonstration that wide neural networks exhibit weak derivative correlations (appendix G), we assumed the absence of correlations in the initial distribution of θ in the infinite width limit. If we introduce such correlations in θ, then these correlations contribute to the derivative correlations, such that they do not vanish. This relation also holds true the other way around, meaning persisting derivative correlations are equivalent to correlations in the initial distribution of θ in the large width limit. Such correlations in the initial distribution of θ, indicate an inherent bias in the initial hypothesis function, since they imply a predisposition towards specific regions in the parameter space. Therefor, weak derivatives correlations are a manifestation of an inherent bias in our initial hypothesis function.

Finite neural networks, by their very nature of having a finite number of parameters, are limited to a small subset of the parameter space. This is because they can be considered as infinite neural networks, with multiple parameters set to zero (and not allowed to change during learning). This explains why even when drawing the initial parameters from an iid distribution, finite neural networks still exhibit non-vanishing derivative correlations, which are minimized when expanding the width.

The equivalence between weak derivative correlations and inherent bias in the system is also manifested in the process of gradient descent (equation 7). When considering this equation, one observes that there are two objects that the optimisation process attempts to minimize: the first is the derivative of the cost function, $\mathcal{C}'(F(\theta), \hat{y})$, and the second is the gradient of the hypothesis function, $\nabla F(\theta)$. Minimizing the norm of the first term signifies the learning of the data, as this term is minimized when the hypothesis function most effectively fit the target function. On the other hand, minimizing the gradient of the hypothesis function's norm, signifies the system's learning its own structure, independently of the data, hence a bias. To be able to minimize the second term via gradient descent, we need the higher derivative correlations to have the same asymptotic behaviour as the gradient, as seen in equation 11 for $1 \leq D$. Thus, weak derivative correlations impede the system's ability to learn its own structure instead of learning the data, which effectively minimizes the bias.

Furthermore, we argue that this interpretation explains both why linear learning is so common, and why linear systems are generally outperformed by their non-linear counterparts. We argue that the derivative correlations represent an inherent bias in the system, and that linear learning should be understood as a consequence of our attempt to minimize this bias. However, in some contexts certain bias can facilitate learning, as exemplified by explicit and implicit regularization. Thus, having weak but nonzero derivative correlations can be beneficial, which explains why near-linear learning is generally better than linear learning. In other words, strict linear learning takes the weak correlations principle to an unproductive extreme.

I Limitations, Further dissection and Generalization

In this section, we enumerate the key assumptions that underpin our analysis and propose potential extensions to our findings beyond these stipulated preconditions. Additionally, we identify potential avenues for related further research.

I.1 Section 2

Our analysis here did not rely on any hidden or nontrivial assumptions, except for those explicitly stated during the tensor definition. Our findings are generalizable and applicable to any random tensor or variable that is dependent on some limiting parameter $n \in \mathbb{N}$. Extending our results to any set with a total order is straightforward.

We anticipate this analytical tool to be beneficial not only for the investigation of wide neural networks but also for the learning of random tensors and variables in general, particularly when focusing on their limiting behavior, for reasons delineated in this paper. It upholds several useful algebraic properties C.3, provides a well-defined, optimal asymptotic bound for any tensor 2.1, and harmonizes naturally with the notion of "convergence in distribution". Further, owing to its inherent generality, it offers widespread applicability. We recommend further exploration into the utilization of this tool in solving other problems.

I.2 Sections 3,4

I.2.1 Assumptions

1. We presuppose that F, $\mathcal{C}$, and ϕ are analytical in their parameters, that is, they are smooth, and their Taylor series converges.
2. All of ϕ derivatives are bounded such as in equation 23.
3. Our analysis is constrained to the case of single-batch stochastic gradient descent, and we assume that our training and testing distributions coincide.
4. We assume that $\mathcal{C}$ is convex, that is, $\mathcal{C}''$ is positive definite.
5. Our theorems 3.1,3.2 and corollary 4.1 are exclusively applicable to PGDML systems, as defined in D.2.
6. theorem 3.1 and corollary 4.1 are valid only for sufficiently small η that is of the same order of magnitude as the η necessary for effective linear studies.
7. Corollary 4.1 stipulates that the first derivative of $\mathcal{C}$ decays exponentially, and the second derivative remains bounded over time for the linear solution.
8. The equivalence showed in theorems 3.1,3.2 demand that all of the derivatives stay fixed. But one can describe a more nuanced equivalence, where the derivatives do significantly

change, but the network itself do behaves linearly, if this change is perpendicular to $\nabla F(\theta_0)$. However, given the fact that neural networks satisfy our simpler conditions we will remain with the above stated version of the equivalence.

I.2.2 Generalizations of the Assumptions

For condition 1, while we typically deal with smooth analytical functions, non-continuous hypothesis functions are common, as with the "ReLU" activation function in neural networks. If our system can be represented as a linear approximation plus a function that is analytical over patches, with the understanding that non-smooth points are of zero measure, then the techniques presented herein can be applied.

Regarding the bound imposed on the derivatives of ϕ, 2, this bound is relatively non-restrictive. Especially considering that ϕ should be analytic and this condition only needs to hold over an arbitrarily large probability set, not the entire probability space.

Extending the single-input batch gradient descent case 3 to other batch schemes, such as multiple-input batches or deterministic single batch GD, is straightforward. This extension simply involves replicating our work while adjusting the specifics of the optimization algorithm of interest. The generalization for more complex gradient-based algorithms follows similar lines, albeit with more nuances.