Non-Gaussian Tensor Programs

Anonymous Author(s)
Affiliation
Address
email

Abstract

The Tensor Programs framework has produced a series of powerful results by 1) expressing any deep learning computation of concern as a principled composition of elementwise nonlinearities and matrix multiplication, and 2) inductively reasoning about the program behavior as the sizes of the matrices in the program tend to infinity. For example, this framework showed that infinitely wide neural networks of *any architecture* exhibit Gaussian process behavior at initialization and evolve like a kernel model during training in the so-called NTK parametrization [22, 23, 27]. Moreover, this framework yielded a novel parameterization, coined μP [25], that for the first time enabled hyperparameter tuning for enormous networks too expensive to train more than once [26]. However, this framework has so far been limited to Gaussian initialized weights, while uniform or truncated Gaussian distributions are more prevalent in practice. This work extends Tensor Programs to general non-Gaussian initializations, thus recovering all of the above results in all practical settings.

1 Introduction

Just like autograd [18] empirically automates the calculation of chain rule of arbitrary computation graphs, Tensor Programs (TP) [21] has automated the theoretical calculation of infinite-width limits of the same (where width of a computation graph corresponds to the size of matrices). What previously were difficult limits to calculate, now becomes routine via TP. For example, [10] in 1994 showed that randomly initialized wide shallow neural networks are Gaussian Processes (which is called the Neural Network-Gaussian Process Correspondence, or NNGP Correspondence), but only recently this has been extended to deep perceptrons [8, 9] and more advanced architectures such as convolutional neural networks [2, 11], and each such extension requires painstaking calculations and careful applications of Law of Large Numbers and Central Limit Theorem. But with Tensor Programs, one can show that NNGP Correspondence holds for any architecture all at once [22]. Similarly, in a certain parametrization, a wide multi-layer perceptron (MLP) evolves like a linear model during training [6], but showing this for advanced architectures was very difficult. TP [23, 27] again was able to prove this behavior for any architecture. Finally, TP gave rise to the Dynamical Dichotomy Theorem [25], a classification of all natural infinite-width limits of neural networks, and led to the discovery of Maximal Update Parametrization, or μP. These results underlie the hyperparameter transfer technology that for the first time enabled the hyperparameter tuning of enormous neural networks too expensive to train more than once [26].

However, these results were only proven for neural networks randomly initialized with *Gaussian* weights. But in practice, non-Gaussian initializations such as uniform or truncated Gaussian, proliferate. Could such non-Gaussian initializations possess different behavior in wide neural networks?

In this work, we answer NO: under mild conditions, non-Gaussian iid initializations behave the same as Gaussian initializations with the same variance in wide neural networks. From the perspective of universality [14] from statistical mechanics, this is not unexpected; yet proving such results rigorously

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

has always been tricky [19]. In our work, we apply the interpolation technique of [1] to interpolate between a non-Gaussian Tensor Program in question and the corresponding Gaussian Tensor Program, and show that this continuous interpolation in the infinite-width limit has identially zero derivative, and therefore the two programs have the same limits.

This result also makes the proof of the Semicircle and Marchenko-Pastur Laws in [24] automatically valid for non-Gaussian random matrix ensembles. Likewise, it shows that the asymptotic singular value distribution of the input-output Jacobian of a random neural network does not depend on the distribution of random initialization beyond its variance.

2 Setup

Given initial vectors $g^1, \ldots, g^{M_0} \in \mathbb{R}^n$ and initial scalars $c^1, \ldots, c^{M_0} \in \mathbb{R}$, consider the following iteration for $m = M_0 + 1, \ldots, M$ that generates new vectors $g^{M_0+1}, \ldots, g^M \in \mathbb{R}^n$ and scalars $c^{M_0+1}, \ldots, c^M \in \mathbb{R}$:

$$g^m_\alpha = \sum_{\beta=1}^n A^{m,T_m}_{\alpha\beta} x^m_\beta, \quad c^m = \frac{1}{n}\sum_{\beta=1}^n x^m_\beta, \quad \text{where } x^m_\alpha = \phi^m(g^1_\alpha, \ldots, g^{m-1}_\alpha, c^1, \ldots, c^{m-1}). \quad (1)$$

Here ϕ^m is a scalar function with $2m-2$ arguments, A^m is an $n \times n$ matrix, and $T_m \in \{1, \top\}$ marks whether this matrix should be transposed or not. The matrices A^m for different m can possibly be the same. In this work, we shall call any computation of this form a *Tensor Program*, or TP.

This formulation of a Tensor Program is equivalent to NETSOR$\top^+$ in [24], as shown in Appendix D. As such, Eq. (1) can express any computation expressible in a DL framework such as PyTorch [13], including gradient descent iterations of neural networks of any architecture, e.g. [25, 27]. This expressivity allows one to treat a wide range of problems uniformly using just Theorem 1 below.

For example, consider the first forward pass of a simple MLP with scalar input ξ and output $f(\xi)$:

$$f(\xi) = V^\top \sigma(W\sigma(\xi U)),$$

where $\xi \in \mathbb{R}; U, V \in \mathbb{R}^n; W \in \mathbb{R}^{n\times n}$. We can express this in a TP as follows: $g^1 = U, g^2 = nV$ are the initial vectors and $c^1 = \xi, c^2 = 1$ are initial scalars (where c^2 will just be ignored). Then the program computes

$$g^3 = A^3 x^3, \quad \text{where } A^3 = W, x^3 = \sigma(\xi U)$$
$$f(\xi) = c^4 = \frac{1}{n}\sum_{\beta=1}^n x^4_\beta, \quad \text{where } x^4 = (nV) \odot \sigma(g^3)$$

with c^3 and g^4 computed but just discarded. See [22, 23, 27, 25] for more examples.

Gaussian Tensor Programs. [21, 24, 27] most commonly work under the following setup:

Setup 1. *Consider a Tensor Program with M vectors $g^1, \ldots, g^M \in \mathbb{R}^n$ and scalars $c^1, \ldots, c^M$. Suppose 1) all initial vectors $g^1, \ldots, g^{M_0}$ are sampled as $(g^1_\alpha, \ldots, g^{M_0}_\alpha) \sim \mathcal{N}(\mu^{in}, \Sigma^{in})$ iid over $\alpha = 1, \ldots, n$, for some $\mu^{in} \in \mathbb{R}^{M_0}, \Sigma^{in} \in \mathbb{R}^{M_0\times M_0}$; 2) all initial scalars $c^1, \ldots, c^{M_0}$ have almost sure limits as $n \to \infty$; 3) all matrices A^m have iid entries from $\mathcal{N}(0, n^{-1})$; 4) any two matrices A^m, A^l are either equal or independently sampled; 5) all the nonlinearities ϕ^m are pseudo-Lipschitz.*[1]

Under this setup, the following Master theorem holds:

Theorem 1 ([24])**.** *Consider Setup 1. Then, as $n \to \infty$, for any pseudo-Lipschitz ψ,*

$$\frac{1}{n}\sum_{\alpha=1}^n \psi\left(g^1_\alpha, \ldots, g^M_\alpha, c^1, \ldots, c^M\right) \xrightarrow{a.s.} \mathring{\Psi}, \quad (2)$$

where $\mathring{\Psi}$ is a deterministic scalar given by a certain recurrent formula. Moreover, if all nonlinearities ϕ^m are linearly bounded and ψ is quadratically bounded, we get convergence also in mean.

[1] A function $f : \mathbb{R}^k \to \mathbb{R}$ is called *pseudo-Lipschitz* if $|f(x) - f(y)| \le C\|x - y\|(1 + \sum_{i=1}^k |x_i|^d + |y_i|^d)$ for some $C, d > 0$.

Non-Gaussian Tensor Programs. We shall generalize Theorem 1 to non-Gaussian distributions.

Definition 1. *We say $f : \mathbb{R}^k \to \mathbb{R}$ is* polynomially smooth *if it is C^∞ and its partial derivatives of any order are polynomially bounded, i.e. for any multidegree $\boldsymbol{\alpha} \geq 0$, $\left|\frac{\partial^{|\boldsymbol{\alpha}|}}{\partial \mathbf{x}^{\boldsymbol{\alpha}}} f(\mathbf{x})\right| \leq C(1 + |x_1|^p + \cdots + |x_k|^p)$ for some $C, p > 0$ that may depend on α.*

Setup 2. *Consider Setup 1, but replace 3) and 5) with the following: 3*) all matrices A^m have iid entries from a distribution (that could vary with m) with zero mean, variance n^{-1}, and all higher moment existing; 5*) all the nonlinearities ϕ^m are polynomially smooth. We further require 6) all moments of initial scalars $c^1, \ldots, c^{M_0}$ to exist and converge as $n \to \infty$.*

Theorem 2 (Ours)**.** *Consider Setup 2. Then, as $n \to \infty$, for any polynomially smooth ψ,*

$$\frac{1}{n}\sum_{\alpha=1}^{n} \psi(g_\alpha^1, \ldots, g_\alpha^M, c^1, \ldots, c^M) \xrightarrow{prob.} \mathring{\Psi} \tag{3}$$

with the same $\mathring{\Psi}$ as in Theorem 1 (but here convergence is in probability). Moreover, if all the nonlinearities ϕ^m are linearly bounded and ψ is quadratically bounded, then the expectation of the LHS of Eq. (3) *converges to the RHS.*

In short, Theorem 2 relaxes matrix sampling to be non-Gaussian in general, at the cost of a) requiring more smoothness in nonlinearities, and b) weakening almost sure convergence to in-probability. Note that Setup 2 still requires initial vectors to be Gaussian, but this is not essential, as we discuss in Appendix A.

On Tensor Programs with variable dimensions. While Setup 1 and 2 assume all hidden dimensions to be equal, Theorem 1 holds also for Tensor Programs with variable dimensions, see e.g. [24]. Since our proof technique is based on interpolation between Gaussian and non-Gaussian weights, it can be straightforwardly extended to variable dimensions, as long as Theorem 1 holds in this setting.

3 Applications

As mentioned in the introduction, the Tensor Programs series of papers so far has proven a wide range of results, which typically have the characteristic of *architectural universality*, i.e., covering any neural architectures. But all of such results have assumed Gaussian weight initialization. Now, armed with Theorem 2, we show the same results hold with *non-Gaussian* weight initialization as well, under mild assumptions, thus extending them to other prevalent initializations such as uniform and truncated Gaussian. This demonstrates these results to be *distributional universal*, i.e., independent of specifics of weight sampling.[2] Theorem 2 here acts like a drop-in replacement for Theorem 1 in their proofs, except we need to a) add more smoothness assumptions on nonlinearities, and b) any almost sure convergence is weakened to convergence in probability.

Our Theorem 2 can be applied to a very general class of neural networks as given by the following:

Setup 3. *Consider a neural network whose forward pass can be expressed as Eq.* (1) *where $T_m = 1$ for all m. Suppose 1) all the activation functions are polynomially smooth; 2) a bias vector of the output layer is initialized with zeros; 3) bias vectors of all other layers are initialized with elementwise images of iid Gaussian vectors under polynomially smooth nonlinearities; 4) weights of the input layer are initialized with elementwise images of iid Gaussian matrices of appropriate sizes under polynomially smooth nonlinearities; 5) weights of the output layer are initialized with iid Gaussians with zero mean and variance inversely proportional to the number of input neurons of this layer; 6) weights of any other layer are initialized iid from a distribution with zero mean and variance inversely proportional to the number of input neurons of this layer, with all higher moments existing.*

NNGP Correspondence. The following result is an immediate consequence of our Theorem 2 and Proposition G.4 of [22]:

Corollary 1. *Consider Setup 3. At initialization, as width tends to infinity, the neural network function converges weakly to a Gaussian Process (GP). Moreover, the empirical kernel associated to this network converges pointwise in probability to the kernel of this GP.*

[2] As we discuss in Appendix A, having non-Gaussian "vector-like" variables, such as biases, requires modifying the corresponding tensor program. This may alter the limit $\mathring{\Psi}$ in Theorem 2. However, the limit still exists and all the corrolaries we discuss in the present section hold.

NTK correspondence. We can directly plug our Theorem 2 into the proof of [27] and obtain

Corollary 2. *Consider Setup 3 and assume NTK parameterization.*[3] *Then under SGD weight updates, NTK of the network at any optimization step converges pointwise in probability to a finite deterministic limit that does not depend on the timestep. Moreover, the output of the network at any timestep converges pointwise in probability to the output of a kernel method with this limit as a kernel.*

We empirically validate Corollaries 1 and 2 in Appendix I.

In Appendix C, we present a variant of Theorem 2 which works for non-smooth but Lipschitz nonlinearities (like ReLU or MaxPool). It allows to generalize Corollary 1 to ReLU nets, but not Corollary 2 (that's because a Tensor Program expressing the backward pass of a ReLU net has ReLU derivatives, which are not even continuous, as nonlinearities).

Random Matrix Theory. Our Theorem 2 implies the semi-circle law for non-Gaussian Wigner matrices, the Marchenko-Pastur law for $AA^\top$, where A is non-Gaussian, and Free Independence Principle for Tensor Programs with non-Gaussian initial weights, thus generalizing TP3; we discuss these results in Appendix B.

Classification of Infinite-Width Limits. Consider now an L-hidden-layer perceptron trained using stochastic gradient descent (SGD). TP4 [25] proposes the notion of *an abc-parameterization*, which is specified by a set of numbers $\{a_l, b_l\}_l \cup \{c\}$: (1) each weight is parameterized as $W^l = n^{-a_l} w^l$ for the actual trainable parameter w^l, (2) $w^l_{\alpha\beta} \sim \mathcal{N}(0, n^{-2b_l})$ at initialization, and (3) the SGD learning rate is taken as ηn^{-c} for some constant η.

When c is too small (i.e. learning rate is too large), SGD can lead to blowup of logits; such a parameterization is called unstable. When c is too large, the function computed by the network does not change with time; such a parameterization is called trivial. The result below follows from Theorem 1:

Corollary 3 (Dynamical dichotomy; [25])**.** *Any nontrivial stable abc-parameterization yields a (discrete-time) infinite-width limit. This limit either (1) allows the last layer embedding to evolve nontrivially, or (2) is described by kernel gradient descent in function space, but not both.*

The first situation is called feature learning, as opposed to the second one, called kernel learning. Feature learning is argued to be a more sensible learning regime for a neural network.

In TP4, [25] proposed the so-called μ-parameterization, which uniquely exhibits maximal feature learning in the infinite-width limit among all abc-parameterizations. In TP5, [26] generalized this μ-parameterization to the Adam optimizer [7] and a wider class of architectures and used it to facilitate hyperparameter tuning of large models.

Since Dynamical dichotomy (Corollary 3) is a direct consequence of convergence of averages given by Theorem 1, we can readily apply our Theorem 2 for the same purpose:

Corollary 4 (Dynamical dichotomy; ours)**.** *Consider an L-hidden-layer perceptron trained using stochastic gradient descent (SGD) under abc-parameterization, but with w_l sampled independently from a distribution with zero mean, variance n^{-2b_l}, and all higher moments existing, for $l = 2, \ldots, L$. Then the statement of Corollary 3 holds.*

The μ-parameterization in this case is the same, and can be recovered by the same argument as in TP4. We therefore theoretically justify the hyperparameter tuning technique used in TP5 for neural networks with non-Gaussian initializations commonly used in practice.

On potential societal impacts. Our work concerns generic behavior of neural nets in the limit of infinite width and therefore does not provide any foreseen societal impacts. The only direct practical application of our work we are aware of is theoretical justification for the hyperparameter tuning method of TP5 [26] for non-Gaussian weight initializations. However, what our work provides is merely justification for this method, while the method itself existed before our work (and was well-justified for Gaussian weight initializations).

[3] For more context on NTK theory, see the seminal work of [6].

4 Proof of Theorem 2

Denote the matrices, the vectors, and the scalars of our non-Gaussian Tensor Program as $\tilde{A}^{M_0+1}, \dots, \tilde{A}^M$, $\tilde{g}^1, \dots, \tilde{g}^M$, and $\tilde{c}^1, \dots, \tilde{c}^M$, respectively. Consider the same Tensor Program with Gaussian weights with zero mean and the same variance n^{-1} Denote its matrices as $A^{M_0+1}, \dots, A^M$, vectors as $g^1, \dots, g^M$, and scalars as $c^1, \dots, c^M$. Note that initial vectors and scalars of both programs coincide, i.e. $g^m = \tilde{g}^m$ and $c^m = \tilde{c}^m$ $\forall m \in [M_0]$.

Where it will not cause confusion, for the sake of brevity, we will denote all the nonlinearities ϕ^m by the same letter ϕ in the sequel, ignoring the fact that they can be different functions with different number of arguments.

Limitations of the proof technique of Theorem 1. Let us first review the idea of Theorem 1 proof, as given in [24]. For simplicity, let us ignore the scalars in the program and assume that each vector is generated with a fresh new matrix each time (i.e. no weights are shared).

Theorem 1 is then proven by induction on m. For $m = M_0$, for any polynomially bounded ϕ, $\frac{1}{n}\sum_\alpha \phi(g^1_\alpha, \dots, g^{M_0}_\alpha) \xrightarrow{\text{a.s.}} \mathbb{E}_{Z\sim\mathcal{N}(\mu^{in},\Sigma^{in})}\phi(Z)$ by Law of Large Numbers (LLN) since $[g^1_\alpha, \dots, g^{M_0}_\alpha] \sim \mathcal{N}(\mu^{in}, \Sigma^{in})$ independently for each coordinate $\alpha \in [n]$.

Suppose the induction hypothesis holds for $m-1$, i.e. $\frac{1}{n}\sum_\beta \phi(g^1_\beta, \dots, g^{m-1}_\beta)$ has an almost sure limit as $n \to \infty$ for any polynomially bounded ϕ. Given such ϕ, denote this limit as $\mathring{\Psi}_\phi$. Conditioned on all initial vectors and all previous matrices $A^{M_0+1}, \dots, A^{m-1}$, each component of g^m is Gaussian as a linear combination of Gaussians: $g^m_\alpha = \sum_\beta A^m_{\alpha\beta}\phi(g^1_\beta, \dots, g^{m-1}_\beta)$. Variance of this Gaussian is $\frac{1}{n}\sum_\beta [\phi]^2(g^1_\beta, \dots, g^{m-1}_\beta)$, which converges almost surely to a deterministic limit $\mathring{\Psi}_{[\phi]^2}$ by induction hypothesis. Therefore g^m_α converges almost surely to a Gaussian independent of $g^1, \dots, g^{M_0}$ and $A^{M_0+1}, \dots, A^{m-1}$. Therefore $\frac{1}{n}\sum_\beta \phi(g^1_\beta, \dots, g^m_\beta)$ has a deterministic almost sure limit for any polynomially bounded ϕ, which gives the induction step.

If we relax our assumption that no weights are shared, $g^1, \dots, g^{m-1}$ may depend on A^m in the induction step. However by Gaussian conditioning argument, g^m is still Gaussian conditioned on all initial vectors and all previous matrices that do not coincide with A^m. Unfortunately, for non-Gaussian weights $\tilde{A}$, the same argument cannot be applied readily. In this case, even under the same conditioning, we cannot reason about the distribution of g^m.

We therefore apply a different argument. The idea is to interpolate the weights from the Gaussian ones, for which convergence of $\frac{1}{n}\sum_\alpha \psi(g^1_\alpha, \dots, g^M_\alpha)$ is given by Theorem 1, to the non-Gaussian ones, and show that the above average does not change along this interpolation in the limit of large n.

Our proof. Recall we put a "tilde" (e.g. $\tilde{A}$) to denote objects of the program with non-Gaussian weights, while for objects of the same program with Gaussian weights we put nothing (e.g. A).

[1] proved a result similar to our Theorem 2 for approximate message passing; see Appendix E for comparison. Their argument was based on the following weight interpolation: $A^m(s) = \sqrt{1-s}A^m + \sqrt{s}\tilde{A}^m$ for all $m \in [M_0+1 : M]$, where $s \in [0,1]$. Here $s = 0$ corresponds to the Gaussian program, while $s = 1$ corresponds to the "target" non-Gaussian one. We are going to use the same interpolation technique. Denote the corresponding vectors and scalars as $g^1(s), \dots, g^M(s)$ and $c^1(s), \dots, c^M(s)$, respectively.

Recall ψ is a test function in Eq. (3). Denote $\Psi_n = \frac{1}{n}\sum_{\alpha=1}^n \psi\left(g^1_\alpha, \dots, g^M_\alpha, c^1, \dots, c^M\right)$ — the Gaussian neuron average, and $\tilde{\Psi}_n = \frac{1}{n}\sum_{\alpha=1}^n \psi\left(\tilde{g}^1_\alpha, \dots, \tilde{g}^M_\alpha, \tilde{c}^1, \dots, \tilde{c}^M\right)$ — the target non-Gaussian neuron average. Also define

$$\Psi_n(s) = \frac{1}{n}\sum_{\alpha=1}^n \psi\left(g^1_\alpha(s), \dots, g^M_\alpha(s), c^1(s), \dots, c^M(s)\right). \tag{4}$$

Note that $\Psi_n(0) = \Psi_n$ and $\Psi_n(1) = \tilde{\Psi}_n$. We are going to prove the following proposition:

Proposition 1. *For any Lipschitz $\chi \in C^1(\mathbb{R})$,*

$$\sup_{t\in[0,1/\sqrt{2}]} \left|\partial_t \mathbb{E}\,\chi(\Psi_n(t^2))\right| = O(n^{-1/2}), \qquad \sup_{t\in[0,1/\sqrt{2}]} \left|\partial_t \mathbb{E}\,\chi(\Psi_n(1-t^2))\right| = O(n^{-1/2}), \tag{5}$$

where constants in big O's may depend on χ, ψ, and nonlinearities ϕ.

Once it is proven, we get by dominated convergence,

$$\limsup_{n\to\infty}\Big|\mathbb{E}\left[\chi(\Psi_n(1/2))-\chi(\Psi_n(0))\right]\Big| = \\ = \limsup_{n\to\infty}\left|\int_0^{\frac{1}{\sqrt{2}}}\partial_t\mathbb{E}\,\chi(\Psi_n(t^2))\,dt\right| \le \int_0^{\frac{1}{\sqrt{2}}}\lim_{n\to\infty}\left|\partial_t\mathbb{E}\,\chi(\Psi_n(t^2))\right|\,dt = 0. \quad (6)$$

Therefore the limit exists and equals to zero. By a similar argument, we get $\lim_{n\to\infty}\left|\mathbb{E}\left[\chi(\Psi_n(1))-\chi(\Psi_n(1/2))\right]\right| = 0$. Hence $\lim_{n\to\infty}\left|\mathbb{E}\left[\chi(\Psi_n(1))-\chi(\Psi_n(0))\right]\right| = 0$.

Note that Theorem 1 implies weak convergence of $\Psi(0)$:

$$\lim_{n\to\infty}\mathbb{E}\,\chi(\Psi_n(0)) = \chi(\mathring{\Psi}) \qquad \text{for any bounded Lipschitz } \chi. \quad (7)$$

The above then means that $\lim_{n\to\infty}\mathbb{E}\,\chi(\Psi_n(1)) = \chi(\mathring{\Psi})$ for any bounded Lipschitz $\chi \in C^1(\mathbb{R})$. This is equivalent to weak convergence of $\Psi_n(1) = \tilde{\Psi}_n$ to $\mathring{\Psi}$ by Portmanteau theorem. Since the weak limit is deterministic, weak convergence further upgrades to convergence in probability.

Similarly, if all ϕ are linearly bounded and ψ is quadratically bounded, Theorem 1 gives convergence of expectations: $\lim_{n\to\infty}\mathbb{E}\left[\Psi_n(0)\right] = \mathring{\Psi}$. From the previous, we have $\lim_{n\to\infty}\left|\mathbb{E}\left[\Psi_n(1)-\Psi_n(0)\right]\right| = 0$ by taking χ to be identity. Therefore $\lim_{n\to\infty}\mathbb{E}\left[\tilde{\Psi}_n\right] = \lim_{n\to\infty}\mathbb{E}\left[\Psi_n(1)\right] = \mathring{\Psi}$.

Definitions of s_+, s_-, $s_\pm$, and the "dot" derivative. Recall we use superscripts to enumerate vectors in the program. At the same time, we used t^2 to denote "t squared". In order to avoid notation clutter, we will use s_+ as a substitute for t^2 and s_- as a substitute for $1-t^2$ in the sequel. We will use $s_\pm$ when a given statement holds both for s_+ and s_-. Here and below the "dot" will always mean the t-derivative, i.e. $\dot{\Psi}(s_+) = \partial_t\Psi(t^2)$.

Proof of Proposition 1. For $t \in [0, 1/\sqrt{2}]$, we swap the t-derivative with expectation and apply Cauchy-Schwarz:

$$\left|\partial_t\mathbb{E}\left[\chi(\Psi_n(s_\pm))\right]\right| = \left|\mathbb{E}\left[\chi'(\Psi_n(s_\pm))\dot{\Psi}_n(s_\pm)\right]\right| \le \sqrt{\mathbb{E}\left([\chi'(\Psi_n(s_\pm))]^2\right)\mathbb{E}\left([\dot{\Psi}_n(s_\pm)]^2\right)}; \quad (8)$$

The swap is validated by the following technical lemma:

Lemma 1. *For any Lipschitz $\chi \in C^1(\mathbb{R})$ and $n \ge 1$, $\mathbb{E}\left(\sup_{t\in[0,1/\sqrt{2}]}\left|\partial_t[\chi(\Psi_n(s_\pm))]\right|\right) < \infty$.*

Since χ is Lipschitz and continuously differentiable, $\mathbb{E}\left([\chi'(\Psi_n(s_\pm))]^2\right)$ is uniformly bounded on $t \in [0, 1/\sqrt{2}]$. Uniform bounds for $\mathbb{E}\left([\dot{\Psi}_n(s_\pm)]^2\right)$ is given by the lemma below:

Lemma 2. $\sup_{t\in[0,1/\sqrt{2}]}\mathbb{E}\left([\dot{\Psi}_n(s_\pm)]^2\right) = O(n^{-1})$.

Once it is proven, the statement of Proposition 1 follows. Indeed,

$$\sup_{t\in[0,1/\sqrt{2}]}\left|\partial_t\mathbb{E}\left[\chi(\Psi_n(s_\pm))\right]\right| \le \sup_{t\in[0,1/\sqrt{2}]}\sqrt{\mathbb{E}\left([\chi'(\Psi_n(s_\pm))]^2\right)\mathbb{E}\left([\dot{\Psi}_n(s_\pm)]^2\right)} = O(n^{-1/2}).$$

□

4.1 Proof of Lemmas 1 and 2 in a simple scenario

To make our arguments clear, we illustrate the proofs of Lemmas 1 and 2 on the simplest non-trivial case of $M_0 = 1$, $M = 2$, with no dependence on c^1 or c^2: $\Psi_n(s) = \frac{1}{n}\sum_\alpha \psi(g^1_\alpha, g^2_\alpha(s))$, $g^2(s) = A(s)\phi(g^1)$, and the only initial vector $g^1 \sim \mathcal{N}(0, I)$; since it is initial, it does not depend on s. Recall in our setup, while weights can be non-Gaussian, initial vectors should be Gaussian and hence coincide in the Gaussian and non-Gaussian cases. This does not mean loss of generality as explained in Appendix A. The full proofs of Lemmas 1 and 2 can be found in Appendix F.

Proof of Lemma 1. Consider the exact expression for $\dot{\Psi}_n(s_\pm)$:

$$\dot{\Psi}_n(s_\pm) = \frac{1}{n}\sum_\alpha \partial_2\psi\left(g^1_\alpha, \sum_\beta A_{\alpha\beta}(s_\pm)\phi(g^1_\beta)\right)\sum_\beta \dot{A}_{\alpha\beta}(s_\pm)\phi(g^1_\beta); \tag{9}$$

For any finite n, Lemma 1 follows from the facts that $\sup_{t\in[0,1]}|a(s_\pm)| \leq |a| + |\tilde{a}|$, $\sup_{t\in[0,1/\sqrt{2}]}|\dot{a}(s_\pm)| \leq |a| + |\tilde{a}|$, initial vectors are Gaussian, initial scalars have all moments, and ϕ and ψ are polynomially smooth. We do not present the full proof of Lemma 1 in the main due to space constraints.

Proof of Lemma 2. For the sake of brevity, here and below we omit the argument $s_\pm$ assuming Ψ_n, weight matrices A, and vectors g are functions of $s_\pm$, which is a function of t:

$$[\dot{\Psi}_n]^2 = \frac{1}{n^2}\sum_{\alpha,\beta}\sum_{\hat{\alpha},\hat{\beta}} \partial_2\psi\left(g^1_\alpha, g^2_\alpha\right)\partial_2\psi\left(g^1_{\hat{\alpha}}, g^2_{\hat{\alpha}}\right)\phi(g^1_\beta)\phi(g^1_{\hat{\beta}})\dot{A}_{\alpha\beta}\dot{A}_{\hat{\alpha}\hat{\beta}}. \tag{10}$$

Since $g^2 = A\phi(g^1)$ is a function of weights, we can re-write the above expression as

$$[\dot{\Psi}_n]^2 = \frac{1}{n^2}\sum_{\alpha,\beta}\sum_{\hat{\alpha},\hat{\beta}} f_{\alpha,\beta,\hat{\alpha},\hat{\beta}}(A)\dot{A}_{\alpha\beta}\dot{A}_{\hat{\alpha}\hat{\beta}}, \tag{11}$$

where $f_{\alpha,\beta,\hat{\alpha},\hat{\beta}}(A) = \partial_2\psi\left(g^1_\alpha, (A\phi(g^1))_\alpha\right)\partial_2\psi\left(g^1_{\hat{\alpha}}, (A\phi(g^1))_{\hat{\alpha}}\right)\phi(g^1_\beta)\phi(g^1_{\hat{\beta}})$.

Since $\mathbb{E}\left[\dot{A}_{\alpha\beta}\dot{A}_{\hat{\alpha}\hat{\beta}}\right] = 0$ if $\alpha \neq \hat{\alpha}$ or $\beta \neq \hat{\beta}$ but it is not zero otherwise, it is necessary to study these two cases separately. Let $[\dot{\Psi}_n]^2 = Q_{1,1} + Q_2$, where

$$Q_2 \stackrel{\text{def}}{=} \frac{1}{n^2}\sum_{\alpha,\beta} f_{\alpha,\beta,\alpha,\beta}(A)\left[\dot{A}_{\alpha\beta}\right]^2, \qquad Q_{1,1} \stackrel{\text{def}}{=} \frac{1}{n^2}\sum_{\substack{\alpha,\beta,\hat{\alpha},\hat{\beta}:\\ \alpha\neq\hat{\alpha}\text{ or }\beta\neq\hat{\beta}}} f_{\alpha,\beta,\hat{\alpha},\hat{\beta}}(A)\dot{A}_{\alpha\beta}\dot{A}_{\hat{\alpha}\hat{\beta}}. \tag{12}$$

We can write the sums over indices as sums over sets $\Gamma^{(2)} = \{(\alpha,\beta,\hat{\alpha},\hat{\beta}) :\ \alpha = \hat{\alpha},\ \beta = \hat{\beta}\}$ and $\Gamma^{(1,1)} = \{(\alpha,\beta,\hat{\alpha},\hat{\beta}) :\ \alpha \neq \hat{\alpha} \text{ or } \beta \neq \hat{\beta}\}$. Recall $a_1,\ldots,a_N$ denote the set of all distinct weight entries $\{A_{\alpha\beta}\}_{\alpha,\beta=1}^n$ in arbitrary order. For every configuration $(\alpha,\beta,\hat{\alpha},\hat{\beta}) \in \Gamma^{(1,1)}$, enumerate this set in such a way that $a_1 = A_{\alpha\beta}$ and $a_2 = A_{\hat{\alpha}\hat{\beta}}$; the order of the rest can be arbitrary. Similarly, for every configuration $(\alpha,\beta,\hat{\alpha},\hat{\beta}) \in \Gamma^{(2)}$, enumerate the set in such a way that $a_1 = A_{\alpha\beta} = A_{\hat{\alpha}\hat{\beta}}$. The above expressions take the following form:

$$Q_2 = \frac{1}{n^2}\sum_{u\in\Gamma^{(2)}} f_u(a_1,\ldots,a_N)\dot{a}_1^2, \qquad Q_{1,1} = \frac{1}{n^2}\sum_{u\in\Gamma^{(1,1)}} f_u(a_1,\ldots,a_N)\dot{a}_1\dot{a}_2. \tag{13}$$

As we are interested in $\mathbb{E}\left([\dot{\Psi}_n]^2\right) = \mathbb{E}\,Q_2 + \mathbb{E}\,Q_{1,1}$, we need bounds for the following expectations:

$$|\mathbb{E}\,Q_2| \leq n^{-2}\left|\Gamma^{(2)}\right| \sup_{u\in\Gamma^{(2)}}\left|\mathbb{E}\left(f_u(a_1,\ldots,a_N)\dot{a}_1^2\right)\right|; \tag{14}$$

$$|\mathbb{E}\,Q_{1,1}| \leq n^{-2}\left|\Gamma^{(1,1)}\right| \sup_{u\in\Gamma^{(1,1)}}\left|\mathbb{E}\left(f_u(a_1,\ldots,a_N)\dot{a}_1\dot{a}_2\right)\right|, \tag{15}$$

uniform on $t \in [0, 1/\sqrt{2}]$. In our simple example, cardinalities of Gamma-terms can be easily bounded using their definitions: $\left|\Gamma^{(2)}\right| \leq n^2$ and $\left|\Gamma^{(1,1)}\right| \leq n^4$.

Let us bound $\sup_{u\in\Gamma^{(1,1)}}|\mathbb{E}\left(f_u(a_1,\ldots,a_N)\dot{a}_1\dot{a}_2\right)|$. We start with Taylor-expanding f_u with respect to the first two arguments up to some order K to be specified later:

$$f_u(x_1,\ldots,x_N) = \sum_{k=0}^K \frac{1}{k!}f_u^{[k]}(x_1,\ldots,x_N) + \frac{1}{K!}f_u^{[>K]}(x_1,\ldots,x_N), \tag{16}$$

where the Taylor terms are

$$f_u^{[k]}(x_1,\dots,x_N) = \sum_{\substack{j_1,j_2\geq 0,\\ j_1+j_2=k}} x_1^{j_1}[x_2]^{j_2} \times \partial_1^{j_1}\partial_2^{j_2} f_u(0,0,x_3,\dots,x_N), \tag{17}$$

$$f_u^{[>K]}(x_1,\dots,x_N) = \sum_{\substack{j_1,j_2\geq 0,\\ j_1+j_2=K+1}} x_1^{j_1}[x_2]^{j_2} \times \int_0^1 \partial_1^{j_1}\partial_2^{j_2} f_u(\xi x_1,\xi x_2,x_3,\dots,x_N)\,d\xi. \tag{18}$$

The expectation we are interested in therefore takes the following form:

$$\mathbb{E}\left[f_u(a_1,\dots,a_N)\dot a_1\dot a_2\right] = \sum_{k=0}^{K}\frac{1}{k!}\mathbb{E}\left[f_u^{[k]}(a_1,\dots,a_N)\dot a_1\dot a_2\right] + \frac{1}{K!}\mathbb{E}\left[f_u^{[>K]}(a_1,\dots,a_N)\dot a_1\dot a_2\right].$$

For all terms except the remainder, the expectation splits nicely:

$$\mathbb{E}\left[f_u^{[k]}(a_1,\dots,a_N)\dot a_1\dot a_2\right] = \sum_{\substack{j_1,j_2\geq 0,\\ j_1+j_2=k}} \mathbb{E}\left[\dot a_1\dot a_2 a_1^{j_1}[a_2]^{j_2}\right] \times \mathbb{E}\left[\partial_1^{j_1}\partial_2^{j_2} f_u(0,0,a_3,\dots,a_N)\right].$$

For the remainder, it is given by

$$\mathbb{E}\left[f_u^{[>K]}(a_1,\dots,a_N)\dot a_1\dot a_2\right] = \sum_{\substack{j_1,j_2\geq 0,\\ j_1+j_2=K+1}} \mathbb{E}\left[\dot a_1\dot a_2 a_1^{j_1}a_2^{j_2} \times \int_0^1 \partial_1^{j_1}\partial_2^{j_2} f_u(\xi a_1,\xi a_2,a_{3:N})\,d\xi\right],$$

We split the expectation using Cauchy-Schwarz:

$$\left|\mathbb{E}\left[\dot a_1\dot a_2 a_1^{j_1}[a_2]^{j_2} \times \int_0^1 \partial_1^{j_1}\partial_2^{j_2} f_u(\xi a_1,\xi a_2,a_3,\dots,a_N)\,d\xi\right]\right| \leq$$
$$\leq \sqrt{\mathbb{E}\left[\left(\dot a_1\dot a_2 a_1^{j_1}[a_2]^{j_2}\right)^2\right] \times \mathbb{E}\left[\left(\int_0^1 \partial_1^{j_1}\partial_2^{j_2} f_u(\xi a_1,\xi a_2,a_3,\dots,a_N)\,d\xi\right)^2\right]}. \tag{19}$$

We claim that $\mathbb{E}\left[\partial_1^{j_1}\partial_2^{j_2} f_u(0,0,a_3,\dots,a_N)\right]$ and $\mathbb{E}\left[\left(\int_0^1 \partial_1^{j_1}\partial_2^{j_2} f_u(\xi a_1,\xi a_2,a_3,\dots,a_N)\,d\xi\right)^2\right]$ are bounded uniformly over $u\in\Gamma^{(1,1)}$, $\xi\in[0,1]$, and $t\in[0,1/\sqrt{2}]$. This claim follows from Lemma 9, the main technical lemma of the present work, proven in Appendix H.4.

We are left with bounding $\mathbb{E}\left[\dot a_1\dot a_2 a_1^{j_1}a_2^{j_2}\right]$ and $\mathbb{E}\left[\left(\dot a_1\dot a_2 a_1^{j_1}a_2^{j_2}\right)^2\right]$ uniformly on $t\in[0,1/\sqrt{2}]$. We prove the following Lemma in Appendix G.1:

Lemma 3. *Let a and $\tilde a$ be independent random variables with zero mean, variance n^{-1}, and all higher moments existing. Let $a(s)=\sqrt{1-s}a+\sqrt{s}\tilde a$ for $s\in[0,1]$. Recall $s_+=t^2$, $s_-=1-t^2$, and $\dot a$ could denote either $\partial_t a(s_+(t))$ or $\partial_t a(s_-(t))$. Then $\mathbb{E}[\dot a]=\mathbb{E}[a\dot a]=0\ \forall t\in[0,1)$ and $\sup_{t\in[0,1/\sqrt{2}]}\left(\mathbb{E}\left[\left|a^j\dot a\right|^k\right]\right)=O\left(n^{-(j+1)k/2}\right)\forall j\geq 0\,\forall k\geq 0$.*

This lemma implies $\mathbb{E}\left[\dot a_1\dot a_2 a_1^{j_1}a_2^{j_2}\right]=0$ if $j_1<2$ or $j_2<2$, therefore the first four terms in our Taylor expansion zero-out: $\mathbb{E}\left[f_u^{[k]}(a_1,\dots,a_N)\dot a_1\dot a_2\right]=0$ for $k<4$. Taking $K=3$, we get:

$$\mathbb{E}\left[f_u(a_1,\dots,a_N)\dot a_1\dot a_2\right] = \frac{1}{3!}\mathbb{E}\left[f_u^{[>3]}(a_1,\dots,a_N)\dot a_1\dot a_2\right] \leq$$
$$\leq \frac{1}{3!}\sum_{\substack{j_1,j_2\geq 0,\\ j_1+j_2=4}} \sqrt{\mathbb{E}\left[\left(\dot a_1\dot a_2 a_1^{j_1}a_2^{j_2}\right)^2\right] \times \mathbb{E}\left[\left(\int_0^1 \partial_1^{j_1}\partial_2^{j_2} f_u(\xi a_1,\xi a_2,a_3,\dots,a_N)\,d\xi\right)^2\right]} =$$
$$= O(n^{-(2+4)/2}) = O(n^{-3}) \tag{20}$$

uniformly over $u \in \Gamma^{(1,1)}$ and $t \in [0, 1/\sqrt{2}]$. From Eq. (15), $|\mathbb{E}\, Q_{1,1}| \leq n^{-2} \left|\Gamma^{(1,1)}\right| O(n^{-3}) = O(n^{-1})$ uniformly over $t \in [0, 1/\sqrt{2}]$.

In order to bound $\mathbb{E}\left[f_u(a_1, \ldots, a_N)\dot{a}_1^2\right]$, it suffices to apply Cauchy-Schwarz directly:

$$\mathbb{E}\left[f_u(a_1, \ldots, a_N)\dot{a}_1^2\right] \leq \sqrt{\mathbb{E}\left(\dot{a}_1^4\right) \times \mathbb{E}\left(\left[f_u(a_1, \ldots, a_N)\right]^2\right)} = O(n^{-1}) \tag{21}$$

uniformly over $u \in \Gamma^{(2)}$ and $t \in [0, 1/\sqrt{2}]$. From Eq. (14), $|\mathbb{E}\, Q_2| \leq n^{-2} \left|\Gamma^{(2)}\right| O(n^{-1}) = O(n^{-1})$ uniformly over $t \in [0, 1/\sqrt{2}]$. This gives $\mathbb{E}\left([\dot{\Psi}_n]^2\right) = O(n^{-1})$ uniformly over $t \in [0, 1/\sqrt{2}]$.

5 Related works

The Tensor Programs series discusses different applications of the (Gaussian) Master theorem proven by [21]. They include: Gaussian process behavior at initialization (TP1, [22]), convergence to a kernel method (TP2, [23]), Free Independence Principle (TP3, [24]), dynamical dichotomy and μ-parameterization (TP4, [25]), and finally application of μP to hyperparameter tuning (TP5, [26]).

Neural networks converge to Gaussian processes as their width goes to infinity, as was proven by [8, 9] for fully-connected nets, and by [11, 2] for convolutional nets; see also [5]. Using the Master theorem, [22] showed that this behavior holds for a very wide class of architectures, including not only convolutional, but also graph convolutional and recurrent neural nets, ResNets, networks with batch normalization, and networks with attention. The seminal work of [6] demonstrated that under certain parameterization, the learning dynamics a neural net converges to that of a kernel method. The corresponding kernel was called Neural Tangent Kernel, or NTK, and drawn a lot of attention in recent years. While the result of [6] was proven only for fully-connected nets with smooth activations, the Master theorem allows to generalize this result for a wider class of architectures (the same as mentioned above), see [23]. [17, 15, 20] and others studied trainability of very deep and wide neural networks using random matrix theory. Their analysis crucially relied on the assumption that hidden representations of a neural network at initialization were *freely independent* from the weights. TP3 [24] was among the first works to validate this assumption rigorously; see also [12]. Infinite-width behavior of a neural net depends on scaling of its hyperparameters (like initial weights variance and learning rate) with width. Dynamical dichotomy proposed in TP4 [25] is a classification of scalings that are meaningful in a certain sense. Another classification of scalings with a different notion of meaningfulness was proposed earlier by [3, 4], but only for two-layered networks. A distribution universality property similar to our Theorem 2 was shown by [1] for approximate messaging passing. However, their model does not cover most of possible neural network computations; see Appendix E for discussion.

6 Limitations of our results

First, our Theorem 2 is applicable only to Tensor Programs with smooth nonlinearities, which draws out several popular activation functions like ReLU or MaxPool. Our Theorem 9 (see Appendix C) does not really solve the issue since a Tensor Program expressing the backward pass involves derivatives of the activation functions, which are not even continuous for ReLU. As a workaround, we could consider their smoothed versions (e.g. Softplus instead of ReLU) with a controllable smoothness parameter, and put this parameter very close to zero, thus getting "almost ReLU".

Second, we prove convergence in probability, which is weaker than almost sure convergence that holds in the Gaussian case. Because of this, we are only able to prove a weaker (but more traditional) version of the semi-circle law and Free Independence Principle, compared to TP3 [24]; see Appendix B.

7 Conclusions

We present a generalization of the Master theorem of [21] to non-Gaussian weight initializations. Our generalization allows for the same applications as the original Master theorem, thus broadening the scope of applicability of the Tensor Programs machinery.

References

[1] Wei-Kuo Chen and Wai-Kit Lam. Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26:1–44, 2021.

[2] Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow gaussian processes. *arXiv preprint arXiv:1808.05587*, 2018.

[3] Eugene Golikov. Towards a general theory of infinite-width limits of neural classifiers. In *International Conference on Machine Learning*, pages 3617–3626. PMLR, 2020.

[4] Eugene A Golikov. Dynamically stable infinite-width limits of neural classifiers. *arXiv preprint arXiv:2006.06574*, 2020.

[5] Boris Hanin. Random neural networks in the infinite width limit as gaussian processes. *arXiv preprint arXiv:2107.01562*, 2021.

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

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

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

[10] Radford M Neal. *BAYESIAN LEARNING FOR NEURAL NETWORKS*. PhD thesis, University of Toronto, 1995.

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

[12] Leonid Pastur. On random matrices arising in deep neural networks. gaussian case. *arXiv preprint arXiv:2001.06188*, 2020.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[14] R. K. Pathria and Paul D. Beale. *Statistical mechanics*. Elsevier/Academic Press, Amsterdam ; Boston, 3rd ed edition, 2011.

[15] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in neural information processing systems*, pages 4785–4795, 2017.

[16] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[17] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pages 3360–3368, 2016.

[18] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[19] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[20] Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, pages 5393–5402. PMLR, 2018.

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

[22] Greg Yang. Tensor programs i: Wide feedforward or recurrent neural networks of any architecture are gaussian processes. *arXiv preprint arXiv:1910.12478*, 2019.

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

[24] Greg Yang. Tensor programs iii: Neural matrix laws. *arXiv preprint arXiv:2009.10685*, 2020.

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

[26] Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 2 for scope and Section 2, 3, and 4 for contributions.
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the last paragraph of Section 3.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2 and Setup 2 in particular.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 4 and Appendices C and F to H.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix I.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See plots in Appendix I.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix I.2.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A On Gaussian Initial Vectors

While Setup 2 assumes all initial vectors to be Gaussian, it does not constrain us to have Gaussian initial bias vectors, or Gaussian initial input layer weights. Indeed, as, say, initial bias vectors, we can use images of Gaussian input vectors of the program under polynomially smooth maps. These elementwise maps can be then absorbed into subsequent nonlinearities that use those input vectors.

This trick, however, requires adding a nonlinearity to a program. Hence a program expressing computations in a Gaussian-initialized neural net and a similar program expressing computations of the corresponding non-Gaussian initialized net are not the same. Hence they may have different infinite-width limits, $\mathring{\Psi}$, in Theorem 2. Nevertheless, the limit does not change if we swap only hidden layers (which are expressed with matrices) with their non-Gaussian counterparts.

This means that the NNGP and NTK kernels of Corollaries 1 and 2 are not necessarily distribution-invariant. However both corollaries are still true as they merely state that these kernels exist.

B Additional applications of Theorem 2

TP3 demonstrated the following applications of Theorem 1: the semi-circle law for Gaussian orthogonal ensembles (GOE), the Marchenko-Pastur law for Gaussian Wishart ensembles, and the free independence principle for neural nets with Gaussian initialization. Here we prove results of a similar kind in a general, not necessarily Gaussian, setup. For the Gaussian case, our results are, however, weaker as we are not able to prove almost sure convergence without relying on Gaussianity. Nevertheless, while being weaker, they have more traditional forms, compared to the ones of [24], as we demonstrate below.

Semicircle law for non-Gaussian Wigner ensembles. The semicircle law for GOE proven in [24] is the following result:

Theorem 3. *For each $n \geq 1$, define the random matrix $A = W + W^\top$ for iid Gaussian matrix $W \in \mathbb{R}^{n\times n}$, $W_{\alpha\beta} \sim \mathcal{N}(0, 1/2n)$. Let $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ be the eigenvalues of A and $\mu_n = \frac{1}{n}\sum_{\alpha=1}^n \delta_{\lambda_\alpha}$ be their empirical distribution. Then $\mu_n \xrightarrow{a.s.} \mu_{sc}$, where μ_{sc} is the distribution with density $\propto \sqrt{4-x^2}$.*

Here by almost sure (weak) convergence of measures we mean that for every compactly supported, continuous $\phi : \mathbb{R} \to \mathbb{R}$, as $n \to \infty$, we have

$$\frac{1}{n}\sum_{\alpha=1}^n \phi(\lambda_\alpha) \xrightarrow{\text{a.s.}} \mathbb{E}_{\lambda\sim\mu_{sc}}\phi(\lambda). \tag{22}$$

In order to be able to apply Tensor Programs machinery to prove this result, [24] used the moment method which states that it suffices to prove almost sure convergene of moments of μ_n:

$$\mathbb{E}_{\lambda\sim\mu_n}\lambda^r = \frac{1}{n}\sum_{\alpha=1}^{n}\lambda_\alpha^r = n^{-1}\operatorname{tr}(A^r) \xrightarrow{\text{a.s.}} \mathbb{E}_{\lambda\sim\mu_{sc}}\lambda^r \quad \forall r\in\mathbb{N}. \tag{23}$$

The trace can be expressed as $\operatorname{tr}(A^r) = \mathbb{E}_z[z^\top A^r z]$ for $z_\alpha \sim \mathcal{N}(0,1)$ iid for each $\alpha\in[n]$. For Gaussian z, $A^r z$ can be expressed as a vector in the following Tensor Program:

$$g^1 = z, \quad g^2 = Wz, \quad g^3 = W^\top z, \tag{24}$$

$$g^{2k} = W(g^{2k-1}+g^{2k-2}), \quad g^{2k+1} = W^\top(g^{2k-1}+g^{2k-2}) \quad \forall k\in[2:r]. \tag{25}$$

Here each nonlinearity simply adds two vectors. Note that these nonlinearities are linear polynomials. Then we get

$$\begin{aligned} n^{-1}\operatorname{tr}(A^r) &= \frac{1}{n}\mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^\top A^r z] = \\ &= \frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}_{g_\alpha^1\sim\mathcal{N}(0,1)}\left[g_\alpha^1\left(g_\alpha^{2r}+g_\alpha^{2r+1}\right)\right] = \frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}_{g_\alpha^1\sim\mathcal{N}(0,1)}\psi(g_\alpha^1,\ldots,g_\alpha^{2r+1}), \end{aligned} \tag{26}$$

where $\psi(x_1,\ldots,x_{2r+1}) = x_1(x_{2r}+x_{2r+1})$ — a quadratic polynomial.

The only thing that prevents us from directly applying Theorem 1 to the expression above is conditional expectation. [24] proves the following conditional theorem:

Theorem 4 ([24])**.** *Consider Setup 1 and assume ψ is quadratically bounded and all the nonlinearities are linearly bounded. Let S be a subset of initial vectors. Then, as $n\to\infty$,*

$$\frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}_S\psi\left(g_\alpha^1,\ldots,g_\alpha^M,c^1,\ldots,c^M\right) \xrightarrow{a.s.} \mathring{\Psi}, \tag{27}$$

where $\mathring{\Psi}$ is the same as in Theorem 1.

Therefore $n^{-1}\operatorname{tr}(A^r) \xrightarrow{\text{a.s.}} \mathring{\Psi}$ for certain $\mathring{\Psi}$. Comparing with Eq. (23), what remains to establish Theorem 3, is to prove that $\mathring{\Psi}$ equals to the r-th moment of the semicircle law. One can find the proof in [24]; we do not reproduce it here.

Using the same technique as in the proof of our Theorem 2, we could prove a result similar to Theorem 4, but without Gaussian assumption. A corollary would be

$$\mathbb{E}_{\tilde\lambda\sim\tilde\mu_n}\tilde\lambda^r = \frac{1}{n}\sum_{\alpha=1}^{n}\tilde\lambda_\alpha^r = n^{-1}\operatorname{tr}(\tilde A^r) \xrightarrow{\text{prob.}} \mathbb{E}_{\lambda\sim\mu_{sc}}\lambda^r \quad \forall r\in\mathbb{N}, \tag{28}$$

where $\tilde A = \tilde W + \tilde W^\top$ and $\tilde W$ is an $n\times n$ matrix with iid entries with zero mean, variance $1/(2n)$, and with all higher moments existing. However, since the result above is in-probability, it does not correspond to any commonly-used notion of random measure convergence. Nevertheless, apart from convergence in probability, our Theorem 2 gives also convergence of expectations which implies

$$\mathbb{E}_{\tilde\mu_n}\mathbb{E}_{\tilde\lambda\sim\tilde\mu_n}\tilde\lambda^r = \mathbb{E}_{\tilde\mu_n}\left(\frac{1}{n}\sum_{\alpha=1}^{n}\tilde\lambda_\alpha^r\right) = n^{-1}\mathbb{E}_{\tilde A}\operatorname{tr}(\tilde A^r) \to \mathbb{E}_{\lambda\sim\mu_{sc}}\lambda^r \quad \forall r\in\mathbb{N}. \tag{29}$$

This is equivalent to weak convergence of expected measures, as stated in the following theorem:

Theorem 5 (Ours)**.** *For each $n\geq 1$, define the random matrix $\tilde A = \tilde W + \tilde W^\top$ for $\tilde W$ being an $n\times n$ matrix with iid entries with zero mean, variance $1/(2n)$, and with all higher moments existing. Let $\tilde\lambda_1,\ldots,\tilde\lambda_n\in\mathbb{R}$ be the eigenvalues of $\tilde A$ and $\tilde\mu_n = \frac{1}{n}\sum_{\alpha=1}^{n}\delta_{\tilde\lambda_\alpha}$ be their empirical distribution. Then $\mathbb{E}\,\tilde\mu_n \Rightarrow \mu_{sc}$, where μ_{sc} is the distribution with density $\propto\sqrt{4-x^2}$ and "$\Rightarrow$" denotes weak convergence.*

This result follows from the expression of an expected trace as an expectation of a certain ψ-function as a Tensor Program:

$$n^{-1}\mathbb{E}_{\tilde{A}}\operatorname{tr}(\tilde{A}^r) = \frac{1}{n}\mathbb{E}_{\tilde{A}}\mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^\top \tilde{A}^r z] =$$
$$= \frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}_{\tilde{W}}\mathbb{E}_{\tilde{g}^1_\alpha\sim\mathcal{N}(0,1)}\left[\tilde{g}^1_\alpha\left(\tilde{g}^{2r}_\alpha + \tilde{g}^{2r+1}_\alpha\right)\right] = \frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}\,\psi(\tilde{g}^1_\alpha,\ldots,\tilde{g}^{2r+1}_\alpha), \quad (30)$$

where again $\psi(x_1,\ldots,x_{2r+1}) = x_1(x_{2r}+x_{2r+1})$. Its convergence is given by our Theorem 2.

Since the Tensor Program is the same as the one used in Theorem 3, the limit is the same too. This means we have just proven Eq. (29) and thus Theorem 5.

Marchenko-Pastur law for non-Gaussian Wishart ensembles. Using exactly the same machinery as for the semi-circle law, [24] proves the Marchenko-Pastur law for Wishart ensembles, i.e. for matrices of the form $AA^\top$, where A is an $m\times n$ Gaussian matrix with zero mean and variance n^{-1}, and $m/n \to \rho \in (0,\infty)$ as $n\to\infty$. Our Theorem 2 (with a remark on programs with variable dimensions, see Section 2) gives a similar result with no assumptions on Gaussianity of A:

Theorem 6 (Ours)**.** *For each $n\geq 1$, let $\tilde{A}$ be an $m\times n$ matrix with iid entries with zero mean, variance $1/n$, and with all higher moments existing. Let $\tilde{\lambda}_1,\ldots,\tilde{\lambda}_n\in\mathbb{R}$ be the eigenvalues of $\tilde{A}\tilde{A}^\top$ and $\tilde{\mu}_n = \frac{1}{n}\sum_{\alpha=1}^{n}\delta_{\tilde{\lambda}_\alpha}$ be their empirical distribution. Let m go to infinity as $n\to\infty$ such that $m/n\to\rho\in(0,\infty)$. Then $\mathbb{E}\,\tilde{\mu}_n \Rightarrow \mu_{mp}$, where μ_{mp} has "density" $p_{mp}(x)$ given below:*

$$p_{mp}(x) = \max(0, 1-\rho^{-1})\delta(x) + \frac{1}{\rho 2\pi x}\sqrt{(b-x)(x-a)}1_{x\in[a,b]}, \quad (31)$$

where $\delta(x)$ is the Dirac Delta, $a=(1-\sqrt{\rho})^2$, and $b=(1+\sqrt{\rho})^2$.

Free Indepedence Principle for Tensor Programs with non-Gaussian weights. Using the same machinery again, [24] proves Free Independence Principle for Tensor Programs:

Theorem 7 ([24])**.** *Consider Setup 1. Let $\mathcal{D}$ denote the collection of diagonal matrices formed from bounded coordinatewise images of vectors in the program:*

$$\mathcal{D} = \left\{\operatorname{diag}\left(\psi\left(g^1,\ldots,g^M\right)\right) : \ \psi:\mathbb{R}^M\to\mathbb{R} \text{ is bounded}\right\}. \quad (32)$$

Then $\mathcal{D}$, along with the random matrix collections $\{A, A^\top\}$ for all matrices in the program are almost surely asymptotically free as $n\to\infty$.

Here by almost sure asymptotical freeness we mean the following:

Definition 2. *Fix k. Consider collections of random matrices $\mathcal{W}^1_n,\ldots,\mathcal{W}^k_n\subseteq\mathbb{R}^{n\times n}$ for each $n\geq 1$, of constant cardinalities (with n). We say $\mathcal{W}^1_n,\ldots,\mathcal{W}^k_n$ are almost surely asymptotically freely independent (or just almost surely asymptotically free), if*

$$n^{-1}\operatorname{tr}\left(\prod_{i=1}^{k}\left(P_i(\mathcal{W}^{j_i}_n) - \tau_i I\right)\right) \xrightarrow{a.s.} 0, \quad (33)$$

where $\tau_i = n^{-1}\operatorname{tr}(P_i(\mathcal{W}^{j_i}_n))$, P_i is a non-commutative polynomial in $|\mathcal{W}^{j_i}_n|$ variables, $j_1,\ldots,j_k\in[k]$ are indices with no two adjacent j_i equal, and $\{P_i\}_i$, $\{j_i\}_i$ independent on n.

The proof strategy for Theorem 7 is very similar to the one of Theorem 3. The trace in each τ_i can be expressed as an expectation $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^\top P_i(\mathcal{W}^{j_i}_n)z]$ and $P_i(\mathcal{W}^{j_i}_n)z$ can be expressed as a vector. Therefore each τ_i can be thought as a scalar in a new Tensor Program, and the "outer" trace in Eq. (33) can be expressed as conditional expectation using the same trick. We refer the reader to [24] for details.

For the same reasons as before, we are not able to prove almost sure asymptotical freeness for non-Gaussian Tensor Programs. Nevertheless, we are able to prove more traditional asymptotical freeness in expectation in our non-Gaussian case:

Definition 3. *Fix k. Consider collections of random matrices $\mathcal{W}_n^1, \dots, \mathcal{W}_n^k \subseteq \mathbb{R}^{n\times n}$ for each $n \geq 1$, of constant cardinalities (with n). We say $\mathcal{W}_n^1, \dots, \mathcal{W}_n^k$ are asymptotically freely independent in expectation (or just asymptotically free), if*

$$n^{-1}\mathbb{E}\left[\operatorname{tr}\left(\prod_{i=1}^{k}\left(P_i(\mathcal{W}_n^{j_i}) - \tau_i I\right)\right)\right] \to 0, \tag{34}$$

where $\tau_i = n^{-1}\mathbb{E}\operatorname{tr}(P_i(\mathcal{W}_n^{j_i}))$, P_i is a non-commutative polynomial in $|\mathcal{W}_n^{j_i}|$ variables, $j_1, \dots, j_k \in [k]$ are indices with no two adjacent j_i equal, and $\{P_i\}_i$, $\{j_i\}_i$ independent on n.

Theorem 8 (ours)**.** *Consider Setup 2 and assume all nonlinearities to be linearly bounded. Let $\tilde{\mathcal{D}}$ denote the collection of diagonal matrices formed from bounded polynomially smooth coordinatewise images of vectors in the program:*

$$\tilde{\mathcal{D}} = \left\{\operatorname{diag}\left(\psi\left(\tilde{g}^1, \dots, \tilde{g}^M\right)\right) : \ \psi : \mathbb{R}^M \to \mathbb{R} \text{ is bounded}\right\}. \tag{35}$$

Then $\tilde{\mathcal{D}}$, along with the random matrix collections $\{\tilde{A}, \tilde{A}^\top\}$ for all matrices in the program are asymptotically free in expectation as $n \to \infty$.

Using the same trick as for the semi-circle law, each τ_i can be expressed as an expectation of an empirical average of a certain (quadratically bounded) ψ-function for a Tensor Program with linearly bounded polynomially smooth nonlinearities. Therefore its convergence is given by our Theorem 2 and the limit is the same as in the Gaussian case. Thinking about each τ_i as a scalar in a new Tensor Program, we can express Eq. (34) again as an expectation of an empirical average of a certain (quadratically bounded) ψ-function for this new Tensor Program with linearly bounded polynomially smooth nonlinearities. Its convergence is given by the same theorem and the limit is the same as in the Gaussian case, i.e. zero.

C Non-Gaussian Master theorem for Lipschitz nonlinearities

One of the limitations of our Theorem 2 is smoothness requirement. In the present section, we prove a similar result that assumes nonlinearities to be Lipschitz but not necessarily smooth:

Setup 4. *Consider a Tensor Program of width n with M vectors $g^1, \dots, g^M$ and M scalars $c^1, \dots, c^M$. Suppose 1) all initial vectors $g^1, \dots, g^{M_0}$ are sampled as $(g_\alpha^1, \dots, g_\alpha^{M_0}) \sim \mathcal{N}(\mu^{in}, \Sigma^{in})$ iid over $\alpha = 1, \dots, n$, for some $\mu^{in} \in \mathbb{R}^{M_0}, \Sigma^{in} \in \mathbb{R}^{M_0\times M_0}$; 2) all the initial scalars have a.s. limits as $n \to \infty$ and all their moments exist and converge; 3) elements of all the matrices are sampled independently from a sub-Gaussian distribution with zero mean and variance n^{-1}; 4) any two matrices A^m, A^l are either equal or independently sampled; 5) all the nonlinearities ϕ in the program are Lipschitz.*

Theorem 9 (Ours)**.** *Consider Setup 4. Then, as $n \to \infty$, for any Lipschitz $\psi : \mathbb{R}^M \to \mathbb{R}$,*

$$\frac{1}{n}\sum_{\alpha=1}^{n}\psi(g_\alpha^1, \dots, g_\alpha^M, c^1, \dots, c^M) \xrightarrow{prob.} \mathring{\Psi} \tag{36}$$

with the same $\mathring{\Psi}$ as in Theorem 1. Moreover, the expectation of the LHS of Eq. (36) *converges to the RHS.*

The following result is an immediate consequence of the above theorem:

Corollary 5 (Convergence to GP at initialization)**.** *Consider a neural network under Setup 3 with 1) and 6) replaced with the following: 1*) all activation functions are Lipschitz; 6*) weights of any layer except for the input and output ones are initialized iid from a **sub-Gaussian** distribution with zero mean and variance inversely proportional to the number of input neurons of this layer. Then the (pre-activation) output of any neuron in the network converges in distribution to a Gaussian process.*

Proof. We use the fact that when all matrices in Eq. (1) are not transposed, the limit in Theorem 1 (and hence in our Theorem 9) takes the form $\mathring{\Psi} = \mathbb{E}_{z\sim\mathcal{N}(\mu,\Sigma)}\psi(z)$, where μ and Σ are computed using a certain recurrent formula; see [22].

Let $g^1, \dots, g^B$ be the pre-activation outputs of a given layer on a batch of inputs of size B. Let ψ be Lipschitz and bounded, and depend only on these B vectors. Then for any $\alpha \in [n]$, by symmetry of neurons, $\mathbb{E}\left[\psi(g_\alpha^1, \dots, g_\alpha^B)\right] \to \mathbb{E}_{z \sim \mathcal{N}(\mu, \Sigma)} \psi(z)$ from Theorem 9. This means that on any batch of inputs of size B, outputs of a given neuron $g_\alpha^1, \dots, g_\alpha^B$ converge weakly to a Gaussian vector. This means that the output of this neuron converges weakly to a Gaussian process as a function of network's input. $\square$

The proof of Theorem 9 is based on interpolating Lipschitz nonlinearities with Lipschitz and polynomially smooth ones. The following lemma constructs the required interpolation:

Lemma 4. *Let $\phi : \mathbb{R}^k \to \mathbb{R}$ be Lipschitz with a Lipschitz constant L. Then there exists a family of polynomially smooth functions $(\phi^\delta : \mathbb{R}^k \to \mathbb{R})_{\delta > 0}$ with Lipschitz constants $\leq L$ such that $\sup_{x \in \mathbb{R}^k} |\phi(x) - \phi^\delta(x)| \leq \delta$.*

Proof. Let $\zeta \in C^\infty(\mathbb{R}^k)$ be a non-negative smooth function that sums to one, supported on the unit ball. For example, we can take a standard bump function:

$$\zeta(z) = C e^{-\frac{1}{1 - \|z\|_2^2}} 1_{\|z\|_2 \leq 1} \tag{37}$$

with the constant C chosen appropriately.

Let $\zeta^\delta(x) = (L/\delta)^k \zeta(xL/\delta)$. Set

$$\phi^\delta(x) = [\phi * \zeta^\delta](x) = \int_{\mathbb{R}^k} \zeta^\delta(y) \phi(x - y)\, dy = \\ = (L/\delta)^k \int_{\mathbb{R}^k} \zeta(yL/\delta) \phi(x - y)\, dy = \int_{\mathbb{R}^k} \zeta(z) \phi(x - z\delta/L)\, dz. \tag{38}$$

We have for any $x \in \mathbb{R}^k$,

$$|\phi^\delta(x) - \phi(x)| = \left| \int_{\mathbb{R}^k} \zeta(z) \left(\phi(x - z\delta/L) - \phi(x)\right)\, dz \right| \leq \delta \int_{\|z\|_2 \leq 1} \zeta(z) \|z\|_2\, dz \leq \delta. \tag{39}$$

Let us check that ϕ^δ is Lipschitz with the Lipschitz constant L:

$$|\phi^\delta(x_1) - \phi^\delta(x_2)| = \left| \int_{\mathbb{R}^k} \zeta(z) \left(\phi(x_1 - z\delta/L) - \phi(x_2 - z\delta/L)\right)\, dz \right| \leq \\ \leq L|x_1 - x_2| \int_{\mathbb{R}^k} \zeta(z)\, dz = L|x_1 - x_2|. \tag{40}$$

It is left to show that ϕ^δ is polynomially smooth. We are going to check that any of its derivatives is moreover Lipschitz. Indeed, for any $p \geq 0$ and any $j_1, \dots, j_p \in [k]$,

$$\partial_{j_1, \dots, j_p} \phi^\delta(x) = (L/\delta)^{k + \sum_{i=1}^p j_i} \int_{\mathbb{R}^k} \partial_{j_1, \dots, j_p} \zeta((x - y)L/\delta) \phi(y)\, dy = \\ = (L/\delta)^{k + \sum_{i=1}^p j_i} \int_{\|y\|_2 \leq \delta/L} \partial_{j_1, \dots, j_p} \zeta(yL/\delta) \phi(x - y)\, dy. \tag{41}$$

Therefore,

$$\left|\partial_{j_1, \dots, j_p} \phi^\delta(x_1) - \partial_{j_1, \dots, j_p} \phi^\delta(x_2)\right| = \\ = (L/\delta)^{k + \sum_{i=1}^p j_i} \left| \int_{\|y\|_2 \leq \delta/L} \partial_{j_1, \dots, j_p} \zeta(yL/\delta) (\phi(x_1 - y) - \phi(x_2 - y))\, dy \right| \leq \\ \leq (L/\delta)^{k + \sum_{i=1}^p j_i} \left| \int_{\|y\|_2 \leq \delta/L} \partial_{j_1, \dots, j_p} \zeta(yL/\delta)\, dy \right| L|x_1 - x_2| \leq C_p |x_1 - x_2| \tag{42}$$

for some finite C_p depending on p. $\square$

Proof of Theorem 9. Let L be the Lipschitz constant for ϕ and ψ. Let $(\phi^\delta)_\delta$ and $(\psi^\delta)_\delta$ be the corresponding families of Lipschitz and polynomially smooth functions from the lemma above. Define the "smoothed" version of the given Tensor Program:

$$g_\alpha^{m,\delta} = \sum_{\beta=1}^{n} A_{\alpha\beta}^{m,T_m} x_\beta^{m,\delta}, \qquad c^{m,\delta} = \frac{1}{n}\sum_{\beta=1}^{n} x_\beta^{m,\delta}, \tag{43}$$

where

$$x_\alpha^{m,\delta} = \phi^{m,\delta}(g_\alpha^{1,\delta}, \ldots, g_\alpha^{m-1,\delta}, c^{1,\delta}, \ldots, c^{m-1,\delta}) \tag{44}$$

for $m \in [M_0 + 1 : M]$. All input vectors and scalars coincide with the original program: $g^{m,\delta} = g^m$, $c^{m,\delta} = c^m$ for $m \in [M_0]$.

As in the main, we denote the matrices, the vectors, and the scalars of our non-Gaussian (smoothed) Tensor Program as $\tilde{A}^{M_0+1}, \ldots, \tilde{A}^M$, $\tilde{g}^{1,\delta}, \ldots, \tilde{g}^{M,\delta}$, and $\tilde{c}^{1,\delta}, \ldots, \tilde{c}^{M,\delta}$, respectively. We consider the same Tensor Program with Gaussian weights with zero mean and the same variance n^{-1} We denote its matrices as $A^{M_0+1}, \ldots, A^M$, vectors as $g^{1,\delta}, \ldots, g^{M,\delta}$, and scalars as $c^{1,\delta}, \ldots, c^{M,\delta}$. The input vectors and scalars of both programs coincide, i.e. $g^m = \tilde{g}^m$ and $c^m = \tilde{c}^m$ $\forall m \in [M_0]$.

Lemma 5. *Under premise of Theorem 9,* $\forall n \geq 1$ $\forall \delta > 0$

$$\left\|\psi(\tilde{g}^1, \ldots, \tilde{g}^M, \tilde{c}^1, \ldots, \tilde{c}^M) - \psi^\delta(\tilde{g}^{1,\delta}, \ldots, \tilde{g}^{M,\delta}, \tilde{c}^{1,\delta}, \ldots, \tilde{c}^{M,\delta})\right\|_2 \leq \\ \leq \sqrt{n}\delta P\left(L\|\tilde{A}^1\|_2, \ldots, L\|\tilde{A}^M\|_2\right), \tag{45}$$

where P is a polynomial.

Proof. We prove by induction that for all $m \in [M_0 : M-1]$ there exists a polynomial P^m such that

$$\left\|\phi(\tilde{g}^1, \ldots, \tilde{g}^m, \tilde{c}^1, \ldots, \tilde{c}^m) - \phi^\delta(\tilde{g}^{1,\delta}, \ldots, \tilde{g}^{m,\delta}, \tilde{c}^{1,\delta}, \ldots, \tilde{c}^{m,\delta})\right\|_2 \leq \\ \leq \sqrt{n}\delta P^m\left(L\|\tilde{A}^1\|_2, \ldots, L\|\tilde{A}^m\|_2\right). \tag{46}$$

Induction base: since $\tilde{g}^1 = \tilde{g}^{1,\delta}, \ldots, \tilde{g}^{M_0} = \tilde{g}^{M_0,\delta}$,

$$\left\|\phi(\tilde{g}^1, \ldots, \tilde{g}^{M_0}, \tilde{c}^1, \ldots, \tilde{c}^{M_0}) - \phi^\delta(\tilde{g}^{1,\delta}, \ldots, \tilde{g}^{M_0,\delta}, \tilde{c}^{1,\delta}, \ldots, \tilde{c}^{M_0,\delta})\right\|_2 \leq \sqrt{n}\delta. \tag{47}$$

Here $P^{M_0} \equiv 1$.

Suppose the induction hypothesis holds for $m-1$. Then we get the following for m:

$$\begin{aligned}
&\left\|\phi(\tilde{g}^1,\dots,\tilde{g}^m,\tilde{c}^1,\dots,\tilde{c}^m)-\phi^\delta(\tilde{g}^{1,\delta},\dots,\tilde{g}^{m,\delta},\tilde{c}^{1,\delta},\dots,\tilde{c}^{m,\delta})\right\|_2 \le \\
&\qquad \le \left\|\phi(\tilde{g}^1,\dots,\tilde{g}^m,\tilde{c}^1,\dots,\tilde{c}^m)-\phi^\delta(\tilde{g}^1,\dots,\tilde{g}^m,\tilde{c}^1,\dots,\tilde{c}^m)\right\|_2 + \\
&\qquad + \left\|\phi^\delta(\tilde{g}^1,\dots,\tilde{g}^m,\tilde{c}^1,\dots,\tilde{c}^m)-\phi^\delta(\tilde{g}^{1,\delta},\dots,\tilde{g}^{m,\delta},\tilde{c}^{1,\delta},\dots,\tilde{c}^{m,\delta})\right\|_2 \le \\
&\le \sqrt{n}\delta + \sqrt{\sum_{\alpha=1}^{n}\left|\phi^\delta(\tilde{g}^1_\alpha,\dots,\tilde{g}^m_\alpha,\tilde{c}^1,\dots,\tilde{c}^m)-\phi^\delta(\tilde{g}^{1,\delta}_\alpha,\dots,\tilde{g}^{m,\delta}_\alpha,\tilde{c}^{1,\delta},\dots,\tilde{c}^{m,\delta})\right|^2} \le \\
&\le \sqrt{n}\delta + L\sqrt{\sum_{\alpha=1}^{n}\sum_{m'=M_0+1}^{m}\left(\left|\tilde{g}^{m'}_\alpha-\tilde{g}^{m',\delta}_\alpha\right|^2+|\tilde{c}^{m'}-\tilde{c}^{m',\delta}|^2\right)} = \\
&= \sqrt{n}\delta + L\sqrt{\sum_{m'=M_0+1}^{m}\left(\|\tilde{g}^{m'}-\tilde{g}^{m',\delta}\|_2^2+n|\tilde{c}^{m'}-\tilde{c}^{m',\delta}|^2\right)} \le \\
&\le \sqrt{n}\delta + L\sqrt{\sum_{m'=M_0+1}^{m}\left(\|\tilde{A}^{m'}\|_2^2+1\right)\|\phi(\tilde{g}^{1:m'-1},\tilde{c}^{1:m'-1})-\phi^\delta(\tilde{g}^{1:m'-1,\delta},\tilde{c}^{1:m'-1,\delta})\|_2^2} \le \\
&\le \sqrt{n}\delta\left(1+L\sqrt{\sum_{m'=M_0+1}^{m}\left(\|\tilde{A}^{m'}\|_2^2+1\right)\left(P^{m'}(L\|\tilde{A}^1\|_2,\dots,L\|\tilde{A}^{m'-1}\|_2)\right)^2}\right) \le \\
&\le \sqrt{n}\delta\left(1+L\sum_{m'=M_0+1}^{m}\left(\|\tilde{A}^{m'}\|_2+1\right)P^{m'}(L\|\tilde{A}^1\|_2,\dots,L\|\tilde{A}^{m'-1}\|_2)\right). \qquad (48)
\end{aligned}$$

We define the new polynomial P^m as

$$P^m(x_1,\dots,x_m) = 1+\sum_{m'=M_0+1}^{m}(x_{m'}+L)\,P^{m'}(x_1,\dots,x_{m'-1}). \qquad (49)$$

This proves the induction. Doing the same step one more time for ψ gives the result. □

From the lemma above,

$$\begin{aligned}
|\tilde{\Psi}_n-\tilde{\Psi}_n^\delta| &= \frac{1}{n}\left\|\psi(\tilde{g}^1,\dots,\tilde{g}^M,\tilde{c}^1,\dots,\tilde{c}^M)-\psi^\delta(\tilde{g}^{1,\delta},\dots,\tilde{g}^{M,\delta},\tilde{c}^{1,\delta},\dots,\tilde{c}^{M,\delta})\right\|_1 \le \\
&\le \frac{1}{\sqrt{n}}\left\|\psi(\tilde{g}^1,\dots,\tilde{g}^M,\tilde{c}^1,\dots,\tilde{c}^M)-\psi^\delta(\tilde{g}^{1,\delta},\dots,\tilde{g}^{M,\delta},\tilde{c}^{1,\delta},\dots,\tilde{c}^{M,\delta})\right\|_2 \le \\
&\le \delta P(L\|\tilde{A}^1\|_2,\dots,L\|\tilde{A}^M\|_2). \qquad (50)
\end{aligned}$$

Therefore for any Lipschitz χ with Lipschitz constant L_χ,

$$|\chi(\tilde{\Psi}_n)-\chi(\tilde{\Psi}_n^\delta)| \le L_\chi|\tilde{\Psi}_n-\tilde{\Psi}_n^\delta| \le \delta L_\chi P(L\|\tilde{A}^1\|_2,\dots,L\|\tilde{A}^M\|_2). \qquad (51)$$

Since the entries of $\tilde{A}$ are sub-Gaussian, expectation of this difference is bounded as $n\to\infty$. Hence $\mathbb{E}\left(|\chi(\tilde{\Psi}_n)-\chi(\tilde{\Psi}_n^\delta)|\right)$ goes to zero as $\delta\to 0+$ uniformly wrt n. Therefore n and δ limits are interchangeable:

$$\lim_{\delta\to 0+}\lim_{n\to\infty}\mathbb{E}\left(|\chi(\tilde{\Psi}_n)-\chi(\tilde{\Psi}_n^\delta)|\right) = \lim_{n\to\infty}\lim_{\delta\to 0+}\mathbb{E}\left(|\chi(\tilde{\Psi}_n)-\chi(\tilde{\Psi}_n^\delta)|\right). \qquad (52)$$

We can also swap the δ-limit and expectation using dominated convergence. Indeed, for $\delta\in(0,1]$, $|\chi(\tilde{\Psi}_n)-\chi(\tilde{\Psi}_n^\delta)|$ is uniformly bounded with $L_\chi P(L\|\tilde{A}^1\|_2,\dots,L\|\tilde{A}^M\|_2)$ which is integrable for any fixed n. Swapping the limits gives zero: $\mathbb{E}\left(\lim_{\delta\to 0+}|\chi(\tilde{\Psi}_n)-\chi(\tilde{\Psi}_n^\delta)|\right)=0$. Therefore

$$\lim_{\delta\to 0+}\lim_{n\to\infty}\mathbb{E}\left(|\chi(\tilde{\Psi}_n)-\chi(\tilde{\Psi}_n^\delta)|\right) = 0, \qquad (53)$$

which implies

$$\lim_{n\to\infty} \mathbb{E}\left(\chi(\tilde{\Psi}_n)\right) = \lim_{\delta\to 0+} \lim_{n\to\infty} \mathbb{E}\left(\chi(\tilde{\Psi}_n^\delta)\right). \tag{54}$$

For any bounded Lipschitz χ, by Theorem 2,

$$\lim_{\delta\to 0+} \lim_{n\to\infty} \mathbb{E}\left(\chi(\tilde{\Psi}_n^\delta)\right) = \lim_{\delta\to 0+} \chi(\mathring{\Psi}^\delta). \tag{55}$$

Exactly the same logic applies to Ψ_n:

$$\lim_{n\to\infty} \mathbb{E}\left(\chi(\Psi_n)\right) = \lim_{\delta\to 0+} \lim_{n\to\infty} \mathbb{E}\left(\chi(\Psi_n^\delta)\right) = \lim_{\delta\to 0+} \chi(\mathring{\Psi}^\delta). \tag{56}$$

Therefore Ψ_n and $\tilde{\Psi}_n$ have the same weak limit. On the other hand, by Theorem 1, Ψ_n converges weakly to $\mathring{\Psi}$:

$$\lim_{n\to\infty} \mathbb{E}\left(\chi(\tilde{\Psi}_n)\right) = \lim_{n\to\infty} \mathbb{E}\left(\chi(\Psi_n)\right) = \chi(\mathring{\Psi}) \quad \forall \chi \in Lip_b(\mathbb{R}). \tag{57}$$

It means that $\tilde{\Psi}_n$ converges to $\mathring{\Psi}$ in probability.

Since ϕ^δ and ψ^δ are both polynomially smooth and Lipschitz (and hence linearly bounded), Theorem 2 also gives convergence of expectations:

$$\lim_{n\to\infty} \mathbb{E}\left[\tilde{\Psi}_n\right] = \lim_{\delta\to 0+} \lim_{n\to\infty} \mathbb{E}\left[\tilde{\Psi}_n^\delta\right] = \lim_{\delta\to 0+} \mathring{\Psi}^\delta. \tag{58}$$

By applying the same argument to Ψ_n, we conclude that their expectations have the same limit. On the other hand, since ϕ and ψ are Lipschitz, by Theorem 1, $\mathbb{E}\,\Psi_n$ also converges to $\mathring{\Psi}$:

$$\lim_{n\to\infty} \mathbb{E}\left[\tilde{\Psi}_n\right] = \lim_{n\to\infty} \mathbb{E}\left[\Psi_n\right] = \mathring{\Psi}. \tag{59}$$

□

D Tensor Program formulation equivalence

The Tensor Program series [22, 23, 24, 27, 25, 26] defines a Tensor Program as a pair of initial state and a sequence of commands. The initial state consists of variables of three different types: A, G, and C, which correspond to matrices, vectors, and scalars in Eq. (1), respectively. In its body, the program can also generate X-vars (that correspond to a vector x in Eq. (1)), which are size-n vectors but with a different meaning compared to G-vars. The G- and C-vars in the initial state are called input variables.

Each command takes some variables from the state, generates a new variable, and appends it to the state. In the most general version of a Tensor Program, NETSOR$\top^+$, the following commands are available:

- Trsp. Input: A : A. Output $A^\top$: A.
- MatMul. Input: A : A, x : X. Output: Ax : G.
- Nonlin$^+$. Input: g^1 : G, $\ldots, g^k$: G, c^1 : C, $\ldots, c^l$: C. Output: x : X, where $x_\alpha = \phi(g^1_\alpha, \ldots, g^k_\alpha, c^1, \ldots, c^l)\ \forall \alpha \in [n]$.
- Moment. Input: g^1 : G, $\ldots, g^k$: G, c^1 : C, $\ldots, c^l$: C. Output: c : C, where $c = \frac{1}{n}\sum_{\alpha=1}^n \phi(g^1_\alpha, \ldots, g^k_\alpha, c^1, \ldots, c^l)$.

It is easy to see that the iteration of Eq. (1) can be expressed as a NETSOR$\top^+$ program above. Note that in all Master theorems, we care only about vectors of type G and scalars. Let us now show that for any NETSOR$\top^+$ program, we can construct an iteration in the form of Eq. (1) that generates the same set of G- and C-vars.

Note that G-vars can only be generated by MatMul. Each such MatMul uses a single X-var which can be generated only by Nonlin$^+$. For the m-th G-var, we therefore get the following iteration:

$$g^m_\alpha = \sum_{\beta=1}^n A^{m,T_m}_{\alpha\beta} \phi^m(g^1_\beta, \ldots, g^{k_m}_\beta, c^1, \ldots, c^{l_m}),$$

where we put $T_m = 1$ if the corresponding A-var has undergone Trsp even number of times by the moment of generating the above G-var, or $T_m = \top$ otherwise. By generating "placeholder" G- or C-vars (say, zeros), we can assume w.l.o.g. that $k_m = l_m = m - 1$. Moreover, we can generate a G-var and a C-var at the same time using the same function ϕ^m as a subsequent nonlinearity may not depend on one of them if not necessary. This gives us an iteration of the form of Eq. (1).

E Comparison with Chen and Lam [1]

The work of [1] considers the following iteration:

$$\tilde{x}^2_\alpha = \phi^2\left(\sum_\delta \tilde{A}_{\alpha\beta} x^1_\beta\right), \quad \tilde{x}^m_\alpha = \phi^m\left(\sum_\delta \tilde{A}_{\alpha\beta}\tilde{x}^{m-1}_\beta, \tilde{x}^{m-2}_\alpha, \ldots, \tilde{x}^2_\alpha, x^1_\alpha\right) \quad \forall m > 2\ \forall \alpha \in [n], \tag{60}$$

where $\tilde{A}$ is a sum of two **symmetric** $n \times n$ matrices: one has iid sub-Gaussian entries with zero mean and variance n^{-1}, while the other is a deterministic matrix divided by n. Denote the first matrix as $\tilde{Z}$, and the second as X/n, so $\tilde{A} = \tilde{Z} + X/n$. Let $A = Z + X/n$, where Z is a symmetric iid Gaussian matrix with zero mean and variance n^{-1}. Consider the corresponding iteration:

$$x^2_\alpha = \phi^2\left(\sum_\delta A_{\alpha\beta} x^1_\beta\right), \quad x^m_\alpha = \phi^m\left(\sum_\delta A_{\alpha\beta}x^{m-1}_\beta, x^{m-2}_\alpha, \ldots, x^2_\alpha, x^1_\alpha\right) \quad \forall m > 2\ \forall \alpha \in [n], \tag{61}$$

The main result of [1] follows:

Theorem 10 ([1]). *Take $M \geq 2$ and a function ψ with M arguments. Suppose ψ and all ϕ^m for $m \in [2 : M]$ are Lipschitz. Then*

$$\lim_{n\to\infty}\left|\frac{1}{n}\sum_{\alpha=1}^n \left(\psi(x^1_\alpha, \ldots, x^M_\alpha) - \psi(\tilde{x}^1_\alpha, \ldots, \tilde{x}^M_\alpha)\right)\right| = 0 \tag{62}$$

in probability.

In this formulation, $\frac{1}{n}\sum_{\alpha=1}^n \psi(x^1_\alpha, \ldots, x^M_\alpha)$ does not converge to $\mathbb{E}_{z\sim\mathcal{N}(\mu,\Sigma)}\psi(z)$ for some μ and Σ since $x^1_\alpha, \ldots, x^M_\alpha$ are images of nonlinear functions. Therefore the above formulation does not allow for a clear Gaussian process interpretation as in our Corollary 1 or Corollary 5.

One can show equivalence of our iteration (1) and the above iteration in a certain scenario. Namely for the above, let X be and identity matrix and assume x^1 is an elementwise image of a Gaussian vector. For our iteration (1), let $M_0 = 1$, let us generate a "twin" variable for each vector, where the only difference is that one uses A^m, while the other uses $A^{m,\top}$, and let each nonlinearity use a sum of x^j_α, $\bar{x}^j_\alpha$, and c^j for each $j \in [m-1]$, where $\bar{x}^j$ is a twin variable for x^j. Equivalence is then showed by reorganizing nonlinearities of the two variants.

Apart from equivalence in the above scenario, the use of iteration (60) for expressing deep learning computations is very limited. First, iteration (60) assumes weight matrices to be symmetric, while such weight initializations are rarely used in practice. Second, iteration (60) applies the same weight matrix each time. While some layers in neural nets can indeed share weights, assuming that all layers share weights strictly limits us to vanilla recurrent neural nets, drawing out even simplest feedforward nets. Third, the form $A = Z + X/n$ has no clear interpretation in typical neural network computations. And last, iteration (60) assumes only one input variable. In terms of neural network computations, this limits us to a single input, no bias vectors, and no output layer. It automatically draws out the ability to express a backward pass (and therefore learning process). Moreover, we cannot hope even to show Gaussian process behavior (as in Corollary 1) as it requires evaluating the network on a batch of inputs of any finite size.

While our proof is based on the same weight interpolation idea as the proof of Theorem 10, certain crucial details are different. While Theorem 10 requires everything to be Lipschitz, our Theorem 2 allows for certain smooth polynomially bounded nonlinearities and ψ-functions. While restricting oneself to Lipschitz nonlinearities still allows to express a forward pass of a, say, ReLU net, it does not allow for expressing its backward pass since the derivative of a ReLU is not Lipschitz (and not even continuous). Moreover, computing a kernel (NNGP or NTK, see Corollary 1 and Corollary 2) requires a quadratic ψ-function.

F Proof of Lemmas 1 and 2

F.1 Proof of Lemma 1

Recall s_+ denotes t^2, s_- denotes $1-t^2$, and a "dot" always denotes the t-derivative. Observe that $\dot{\Psi}(s_+) = \partial_t \Psi(t^2) = 2t\Psi'(s_+)$ and $\dot{\Psi}(s_-) = \partial_t \Psi(1-t^2) = -2t\Psi'(s_-)$, where Ψ' is a derivative of Ψ wrt its argument. The following lemma gives the exact expression for Ψ':

Lemma 6. *$\Psi'(s)$ can be expressed as a sum of a finite, not depending on n, number of terms of the form*

$$\frac{1}{n}\sum_{\gamma}\sum_{\beta_1}\dots\sum_{\beta_w}\pi_{\beta_{1:w}}(a_1(s),\dots,a_N(s))f_{\gamma,\beta_{1:w}}(a_1(s),\dots,a_N(s))\partial_s A^{m,T_m}_{\beta_w\gamma}(s), \tag{63}$$

where $w \in [1:M-M_0]$, $m \in [M_0+1:M]$, $a_1,\dots,a_N$ is a collection of all weights in the program,

$$f_{\gamma,\beta_{1:w}}(a_1,\dots,a_N) = \partial_{m_1+1}\psi(g^1_{\beta_1},\dots,g^M_{\beta_1},c^1,\dots,c^M)\times$$
$$\times\left[\prod_{j=2}^{w}\partial_{m_j+1}\phi(g^1_{\beta_j},\dots,g^{m_{j-1}}_{\beta_j},c^1,\dots,c^{m_{j-1}})\right]\phi(g^1_\gamma,\dots,g^{m_w}_\gamma,c^1,\dots,c^{m_w}), \tag{64}$$

where $M > m_1 > \dots > m_w = m-1 \in [M_0 : M-1]$, and

$$\pi_{\beta_{1:w}}(a_1,\dots,a_N) = \prod_{k=2}^{w} a'_k, \qquad a'_k \in \left\{A^{m_{k-1}+1,T_{m_{k-1}+1}}_{\beta_{k-1}\beta_k}, n^{-1}\right\}. \tag{65}$$

See Appendix G.2 for the proof. This lemma implies that $\dot{\Psi}(s_\pm)$ can be expressed as a sum of a finite, not depending on n, number of terms of the form

$$\frac{1}{n}\sum_{\gamma}\sum_{\beta_1}\dots\sum_{\beta_w}\pi_{\beta_{1:w}}(a_1(s_\pm),\dots,a_N(s_\pm))f_{\gamma,\beta_{1:w}}(a_1(s_\pm),\dots,a_N(s_\pm))\dot{A}^{m,T_m}_{\beta_w\gamma}(s_\pm) \tag{66}$$

since $2t\partial_s A^{m,T_m}_{\beta_w\gamma}(s)|_{s=s_+} = \dot{A}^{m,T_m}_{\beta_w\gamma}(s_+)$ and $-2t\partial_s A^{m,T_m}_{\beta_w\gamma}(s)|_{s=s_-} = \dot{A}^{m,T_m}_{\beta_w\gamma}(s_-)$.

We are going to use the following definitions in the sequel:

Definition 4. *Let $\sigma_k = \mathrm{MatMul}$ if $a'_k = A^{m_{k-1}+1,T_{m_{k-1}+1}}_{\beta_{k-1}\beta_k}$ and $\sigma_k = \mathrm{Moment}$ otherwise. Let $w_{\mathrm{MatMul}} = \sum_{k=2}^{w} 1_{\sigma_k=\mathrm{MatMul}}$ and let $w_{\mathrm{Moment}} = \sum_{k=2}^{w} 1_{\sigma_k=\mathrm{Moment}}$.*

Note that $w_{\mathrm{MatMul}} + w_{\mathrm{Moment}} = w-1$.

Since derivatives of ϕ and ψ are polynomially bounded, for any fixed finite n, f is polynomially bounded as a function of weights, initial vectors, and initial scalars; π is just a polynomial as a function of weights. Therefore since χ is Lipschitz, for any fixed n, $\chi'(\Psi(s_\pm))\dot{\Psi}(s_\pm)$ is polynomially bounded as a function of weights, t-derivatives of weights, initial vectors, and initial scalars. We conclude using the following lemma:

Lemma 7. *Let h be a polynomially bounded function of weights $a_1(s_\pm),\dots,a_N(s_\pm)$, t-derivatives of weights, initial vectors, and initial scalars. Then*

$$\mathbb{E}\sup_{t\in[0,1/\sqrt{2}]}\left|h\left(a_1(s_\pm),\dots,a_N(s_\pm),\dot{a}_1(s_\pm),\dots,\dot{a}_N(s_\pm),g^1,\dots,g^{M_0},c^1,\dots,c^{M_0}\right)\right| < \infty. \tag{67}$$

Proof. Let P be the corresponding polynomial. Given a polynomial $P(x) = \sum b_k x^k$, define a polynomial $\tilde{P}(x) = \sum |b_k| x^k$. For brevity, we denote the set of arguments of h as $x_1(s_\pm),\dots,x_K(s_\pm)$. We have

$$\sup_{t\in[0,1/\sqrt{2}]}|h(x_1(s_\pm),\dots,x_K(s_\pm))| \le \sup_{t\in[0,1/\sqrt{2}]}|P(x_1(s_\pm),\dots,x_K(s_\pm))| \le$$
$$\le \sup_{t\in[0,1/\sqrt{2}]}\tilde{P}(|x_1(s_\pm)|,\dots,|x_K(s_\pm)|) \le \tilde{P}\left(\sup_{t\in[0,1/\sqrt{2}]}|x_1(s_\pm)|,\dots,\sup_{t\in[0,1/\sqrt{2}]}|x_K(s_\pm)|\right). \tag{68}$$

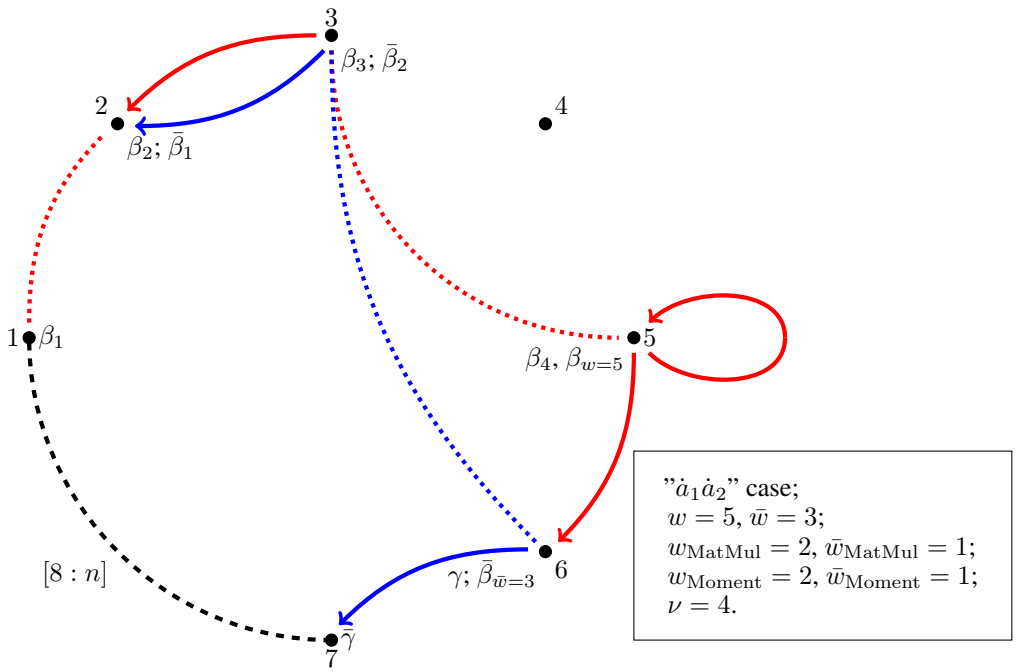

Figure 1: An example of a configuration graph — a subgraph of a union of two paths $(\beta_{1:w}, \gamma)$ and $(\bar\beta_{1:\bar w}, \bar\gamma)$ on an n-node graph, colored red and blue, consisting of w and $\bar w$ edges respectively. Some of the edges can be absent (dotted lines), resulting in $w_{\text{MatMul}} + 1$ edges in one path and $\bar w_{\text{MatMul}} + 1$ in the other. The edge (β_{k-1}, β_k) of the red path is present iff $\sigma_k = \text{MatMul}$, while the edge $(\bar\beta_{k-1}, \bar\beta_k)$ of the blue path is present iff $\bar\sigma_k = \text{MatMul}$. The last edges of both paths, (β_w, γ) and $(\bar\beta_{\bar w}, \bar\gamma)$, are always present; they coincide in the "$\dot a_1^2$" case while they do not in the "$\dot a_1 \dot a_2$" one. Each present edge (β_{k-1}, β_k) of the red path is directed "forward" if $T_k = 1$ and "backward" if $T_k = \top$; same for the blue path. The overall graph consists of ν multi-edges of different multiplicities. In our example, $\nu = 4$, one of the edges has multiplicity two and all others are of multiplicity one.

Observe that $|a(s_+)| = \left|t\tilde a + \sqrt{1-t^2}a\right| \le |a| + |\tilde a|$ for $t \in [0,1]$ and $|\dot a(s_+)| = \frac{1}{2}\left|\tilde a - at/\sqrt{1-t^2}\right| \le |a| + |\tilde a|$ for $t \in [0, 1/\sqrt{2}]$; similarly, $|a(s_-)| \le |a| + |\tilde a|$ and $|\dot a(s_-)| \le |a| + |\tilde a|$ for $t \in [0, 1/\sqrt{2}]$. Therefore there exists a polynomial $\bar P$ with positive finite coefficients such that

$$\sup_{t\in[0,1/\sqrt{2}]} |h\left(x_1(s_\pm), \dots, x_K(s_\pm)\right)| \le \bar P\left(a_1, \dots, a_N, \tilde a_1, \dots, \tilde a_N, g^1, \dots, g^{M_0}, c^1, \dots, c^{M_0}\right). \tag{69}$$

Since all moments of $a_1, \dots, a_N, \tilde a_1, \dots, \tilde a_N, g^1, \dots, g^{M_0}$, and $c^1, \dots, c^{M_0}$ exist, expectation of the above polynomial is finite. □

F.2 Proof of Lemma 2

Due to Lemma 6, $\mathbb{E}\left[[\dot\Psi(s_\pm)]^2\right]$ can be expressed a sum of a finite, not depending on n, number of terms of the form

$$\frac{1}{n^2}\sum_{\gamma,\bar\gamma}\sum_{\beta_1}\cdots\sum_{\beta_w}\sum_{\bar\beta_1}\cdots\sum_{\bar\beta_{\bar w}}\mathbb{E}\left(\left[\pi_{\beta_{1:w}} f_{\gamma,\beta_{1:w}} \dot A^{m,T_m}_{\beta_w\gamma}\right] \times \left[\pi_{\bar\beta_{1:\bar w}} f_{\bar\gamma,\bar\beta_{1:\bar w}} \dot A^{\bar m,\bar T_{\bar m}}_{\bar\beta_{\bar w}\bar\gamma}\right]\right), \tag{70}$$

where $w, \bar w \in [1 : M - M_0]$, and f and π are functions of $a_1(s_\pm), \dots, a_N(s_\pm)$; for brevity, we shall supress the $s_\pm$-argument.We can equivalently express Eq. (70) as a finite, not depending on n, number of terms of one of the two following forms:

$$n^{-2} n^{-w_{\text{Moment}}} n^{-\bar w_{\text{Moment}}} \sum_{(\gamma,\beta_{1:w},\bar\gamma,\bar\beta_{1:\bar w})\in\Gamma^{(2)}_{r_{1:\nu}}} \mathbb{E}\left(f_{\gamma,\beta_{1:w}} f_{\bar\gamma,\bar\beta_{1:\bar w}} \dot a_1^2 a_1^{r_1} \dots a_\nu^{r_\nu}\right), \tag{71}$$

where $r_1 \geq 0$, $r_k \geq 1$ for $k \in [2:\nu]$, $\nu \in [1 : w_{\mathrm{MatMul}} + \bar{w}_{\mathrm{MatMul}} + 1]$, and

$$n^{-2} n^{-w_{\mathrm{Moment}}} n^{-\bar{w}_{\mathrm{Moment}}} \sum_{(\gamma, \beta_{1:w}, \bar{\gamma}, \bar{\beta}_{1:\bar{w}}) \in \Gamma^{(1,1)}_{r_{1:\nu}}} \mathbb{E}\left(f_{\gamma, \beta_{1:w}} f_{\bar{\gamma}, \bar{\beta}_{1:\bar{w}}} \dot{a}_1 \dot{a}_2 a_1^{r_1} \dots a_\nu^{r_\nu} \right), \tag{72}$$

where $r_1 \geq 0$, $r_2 \geq 0$, $r_k \geq 1$ for $k \in [3:\nu]$, $\nu \in [2 : w_{\mathrm{MatMul}} + \bar{w}_{\mathrm{MatMul}} + 2]$; $\sum_{k=1}^{\nu} r_k = w_{\mathrm{MatMul}} + \bar{w}_{\mathrm{MatMul}}$ in both cases.

We define $\Gamma^{(2)}_{r_{1:\nu}}$ as a set of index tuples $(\gamma, \beta_{1:w}, \bar{\gamma}, \bar{\beta}_{1:\bar{w}})$ for each of which there exists a certain enumeration of all distinct weights in the program $a_1, \dots, a_N$ such that

$$\left[\pi_{\beta_{1:w}} \dot{A}^{m, T_m}_{\beta_w \gamma} \right] \left[\pi_{\bar{\beta}_{1:\bar{w}}} \dot{A}^{\bar{m}, \bar{T}_{\bar{m}}}_{\bar{\beta}_{\bar{w}} \bar{\gamma}} \right] = n^{-w_{\mathrm{Moment}}} n^{-\bar{w}_{\mathrm{Moment}}} \dot{a}_1^2 a_1^{r_1} \dots a_\nu^{r_\nu}. \tag{73}$$

Similarly, we define $\Gamma^{(1,1)}_{r_{1:\nu}}$ as a set of index tuples $(\gamma, \beta_{1:w}, \bar{\gamma}, \bar{\beta}_{1:\bar{w}})$ for each of which there exists a certain enumeration of all distinct weights in the program $a_1, \dots, a_N$ such that

$$\left[\pi_{\beta_{1:w}} \dot{A}^{m, T_m}_{\beta_w \gamma} \right] \left[\pi_{\bar{\beta}_{1:\bar{w}}} \dot{A}^{\bar{m}, \bar{T}_{\bar{m}}}_{\bar{\beta}_{\bar{w}} \bar{\gamma}} \right] = n^{-w_{\mathrm{Moment}}} n^{-\bar{w}_{\mathrm{Moment}}} \dot{a}_1 \dot{a}_2 a_1^{r_1} \dots a_\nu^{r_\nu}. \tag{74}$$

We are going to bound only the second term Eq. (72) here; see the bound for the first one in Appendix G.3. We have:

$$n^{-2-w_{\mathrm{Moment}}-\bar{w}_{\mathrm{Moment}}} \left| \sum_{(\gamma, \beta_{1:w}, \bar{\gamma}, \bar{\beta}_{1:\bar{w}}) \in \Gamma^{(1,1)}_{r_{1:\nu}}} \mathbb{E}\left(f_{\gamma, \beta_{1:w}} f_{\bar{\gamma}, \bar{\beta}_{1:\bar{w}}} \dot{a}_1 \dot{a}_2 a_1^{r_1} \dots a_\nu^{r_\nu} \right) \right| \leq$$
$$\leq n^{-2-w_{\mathrm{Moment}}-\bar{w}_{\mathrm{Moment}}} \left| \Gamma^{(1,1)}_{r_{1:\nu}} \right| \sup_{\gamma, \beta_{1:w}, \bar{\gamma}, \bar{\beta}_{1:\bar{w}}} \left| \mathbb{E}\left(f_{\gamma, \beta_{1:w}} f_{\bar{\gamma}, \bar{\beta}_{1:\bar{w}}} \dot{a}_1 \dot{a}_2 a_1^{r_1} \dots a_\nu^{r_\nu} \right) \right|. \tag{75}$$

Cardinality of $\Gamma^{(1,1)}_{r_{1:\nu}}$ is given by the following lemma:

Lemma 8. $\left| \Gamma^{(1,1)}_{r_{1:\nu}} \right| \leq C n^{\nu+2} n^{w_{\mathrm{Moment}}} n^{\bar{w}_{\mathrm{Moment}}}$*, where C depends neither on n, nor on t.*

Proof. Consider a set of n vertices. For a given set of summation indices $(\beta_{1:w}, \gamma)$ associate a set of edges as follows. For each $k \in [2:w]$, take an edge (β_{k-1}, β_k) if $\sigma_k = \mathrm{MatMul}$, and also take an edge (β_w, γ); overall, we get $w_{\mathrm{MatMul}} + 1$ edges. Direct each edge according to T; for example, direct the edge (β_{k-1}, β_k) "forward" if $T_k = 1$ and "backward" if $T_k = \top$. Do the same for $(\bar{\beta}_{1:\bar{w}}, \bar{\gamma})$; obtain a set of $\bar{w}_{\mathrm{MatMul}} + 1$ directed edges. We shall refer the union of these two sets as a configuration graph; see Fig. 1.

A configuration $(\gamma, \beta_{1:w}, \bar{\gamma}, \bar{\beta}_{1:\bar{w}})$ lies in $\Gamma^{(1,1)}_{r_{1:\nu}}$ iff the corresponding configuration graph consists of ν directed multi-edges with multiplicities $r_1 + 1$, $r_2 + 1$, and $r_3, \dots, r_\nu$. We need to compute the number of vertices such a graph traverses. If the graph traverses no more than k vertices then the number of configurations satisfying the condition is no more than Cn^k, where C is the number of non-isomorphic directed graphs with $w_{\mathrm{MatMul}} + \bar{w}_{\mathrm{MatMul}} + 2$ edges and ν multi-edges; obviously, C depends neither on n, nor on t.

Consider a non-directed graph with ν multi-edges. If it has p connected components, it traverses $\nu + p$ vertices. Recall our configuration graph is a union of two paths, having w and $\bar{w}$ edges respectively, with some of the edges erased, resulting in $w_{\mathrm{MatMul}} + 1$ and $\bar{w}_{\mathrm{MatMul}} + 1$ edges respectively. Overall, we had to erase $w_{\mathrm{Moment}} + \bar{w}_{\mathrm{Moment}}$ edges. Given a union of paths, erasing k edges results in introduction of no more than k connected components. If there are only two paths from the beginning, we should have no more than $k + 2$ connected components after erasing k edges.

In our case, $k = w_{\mathrm{Moment}} + \bar{w}_{\mathrm{Moment}}$. Hence our configuration graph traverses no more than $\nu + 2 + w_{\mathrm{Moment}} + \bar{w}_{\mathrm{Moment}}$ vertices. Therefore the number of configurations $(\gamma, \beta_{1:w}, \bar{\gamma}, \bar{\beta}_{1:\bar{w}})$ in $\Gamma^{(1,1)}_{r_{1:\nu}}$ is not greater than $Cn^{\nu+2} n^{w_{\mathrm{Moment}}} n^{\bar{w}_{\mathrm{Moment}}}$ for some constant C independent on both n and t. □

It remains to prove the following proposition:

Proposition 2.

$$\sup_{t\in[0,1/\sqrt{2}]}\sup_{\substack{\gamma,\beta_{1:w},\\ \bar\gamma,\bar\beta_{1:\bar w}}}\left|\mathbb{E}\left(\left[f_{\gamma,\beta_{1:w}}f_{\bar\gamma,\bar\beta_{1:\bar w}}\right](a_1,\dots,a_N)\dot a_1\dot a_2 a_1^{r_1}\dots a_\nu^{r_\nu}\right)\right| = O(n^{-\nu-1}), \tag{76}$$

where $a_1,\dots,a_N$ are functions of $s_\pm$.

Proof. Define $f = f_{\gamma,\beta_{1:w}}\times f_{\bar\gamma,\bar\beta_{1:\bar w}}$. We start with Taylor-expanding f:

$$f(x_1,\dots,x_N) = \sum_{k=0}^{K}\frac{1}{k!}\sum_{\substack{j_{1:\nu}\geq 0,\\ j_1+\dots+j_\nu=k}} x_1^{j_1}\dots x_\nu^{j_\nu}\times\partial_1^{j_1}\dots\partial_\nu^{j_\nu}f(0,\dots,0,x_{\nu+1},\dots,x_N)+$$
$$+\frac{1}{K!}\sum_{\substack{j_{1:\nu}\geq 0,\\ j_1+\dots+j_\nu=K+1}} x_1^{j_1}\dots x_\nu^{j_\nu}\times\int_0^1\partial_1^{j_1}\dots\partial_\nu^{j_\nu}f(\xi x_1,\dots,\xi x_\nu,x_{\nu+1},\dots,x_N)\,d\xi, \tag{77}$$

where K is a fixed number which we are going to choose later in the sequel.

The expectation we are interested in therefore takes the following form:

$$\mathbb{E}\left[f(a_1,\dots,a_N)\dot a_1\dot a_2 a_1^{r_1}\dots a_\nu^{r_\nu}\right] =$$
$$=\sum_{k=0}^{K}\frac{1}{k!}\sum_{\substack{j_{1:\nu}\geq 0,\\ j_1+\dots+j_\nu=k}}\mathbb{E}\left[\dot a_1\dot a_2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\right]\mathbb{E}\left[\partial_1^{j_1}\dots\partial_\nu^{j_\nu}f(0_{1:\nu},a_{\nu+1:N})\right]+$$
$$+\frac{1}{K!}\sum_{\substack{j_{1:\nu}\geq 0,\\ j_1+\dots+j_\nu=K+1}}\mathbb{E}\left[\dot a_1\dot a_2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\int_0^1\partial_1^{j_1}\dots\partial_\nu^{j_\nu}f(\xi a_{1:\nu},a_{\nu+1:N})\,d\xi\right]. \tag{78}$$

By Cauchy-Schwarz,

$$\left|\mathbb{E}\left[\dot a_1\dot a_2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\int_0^1\partial_1^{j_1}\dots\partial_\nu^{j_\nu}f(\xi a_{1:\nu},a_{\nu+1:N})\,d\xi\right]\right|\leq$$
$$\leq\sqrt{\mathbb{E}\left[\left(\dot a_1\dot a_2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\right)^2\right]\mathbb{E}\left[\left(\int_0^1\partial_1^{j_1}\dots\partial_\nu^{j_\nu}f(\xi a_{1:\nu},a_{\nu+1:N})\,d\xi\right)^2\right]}. \tag{79}$$

We need to bound $\mathbb{E}\left[\partial_1^{j_1}\dots\partial_\nu^{j_\nu}f(0_{1:\nu},a_{\nu+1:N}(s_\pm))\right]$ and $\mathbb{E}\left[\partial_1^{j_1}\dots\partial_\nu^{j_\nu}f(\xi a_{1:\nu}(s_\pm),a_{\nu+1:N}(s_\pm))\right]$ uniformly over γ, $\beta_{1:w}$, $\bar\gamma$, $\bar\beta_{1:\bar w}\in[n]$, $t\in[0,1/\sqrt{2}]$, and $\xi\in[0,1]$. Such a bound is given by the following lemma, which we prove in Appendix H.4:

Lemma 9. *Consider a Tensor Program of width n with matrices whose elements are iid and satisfy $\mathbb{E}\,\bar A_{\beta\gamma}=0$ and $\mathbb{E}\,|\bar A_{\beta\gamma}|^k\leq\nu_k n^{-k/2}$, $\nu_k\in\mathbb{R}_+\ \forall k\geq 1$, and with polynomially smooth nonlinearity $\bar\phi$. Denote the corresponding vectors as $\bar g^1,\dots,\bar g^M$ and the corresponding scalars as $\bar c^1,\dots,\bar c^M$.*

Take $u\in\mathbb{N}$ and let $f_1,\dots,f_u$ be polynomially smooth scalar functions; for each $i\in[u]$, f_i has m_i arguments, where $m_i\in[M]$. Then for any collection P of weights in the program and any $n\geq 1$,

$$\sup_{\alpha_1\in[n]}\dots\sup_{\alpha_u\in[n]}\mathbb{E}\left|\partial_P\prod_{i=1}^{u}f_i\left(\bar g^1_{\alpha_i},\dots,\bar g^{m_i}_{\alpha_i},\bar c^1,\dots,\bar c^{m_i}\right)\right|\leq\bar B_{u,|P|}, \tag{80}$$

where $\bar B_{u,|P|}$ is a continuous function of a finite number of (scaled) moments ν.

Indeed, the set $(0,\dots,0,a_{\nu+1}(s_\pm),\dots,a_N(s_\pm))$ satisfies the conditions on $\bar A$ for any $t\in[0,1]$. The corresponding moments ν are continous functions of s; since the constants in the lemma are also

continuous functions of finite number of moments, we get

$$\sup_{t\in[0,1]} \mathbb{E}\left|\partial_1^{j_1}\dots\partial_\nu^{j_\nu} f(0,\dots,0,a_{\nu+1}(s_\pm),\dots,a_N(s_\pm))\right| \le$$
$$\le \max_{t\in[0,1]} \bar{B}_{u,k}(\nu_1(s_\pm),\dots,\nu_Q(s_\pm)) = O_{n\to\infty}(1), \quad (81)$$

where Q is the maximal index of the scaled moment which $\bar{B}_{u,k}$ depends on. Similarly,

$$\sup_{t\in[0,1]}\sup_{\xi\in[0,1]} \mathbb{E}\left[\left(\int_0^1 \partial_1^{j_1}\dots\partial_\nu^{j_\nu} f(\xi a_1(s_\pm),\dots,\xi a_\nu(s_\pm),a_{\nu+1}(s_\pm),\dots,a_N(s_\pm))\,d\xi\right)^2\right] \le$$
$$\le \max_{t\in[0,1]}\max_{\xi\in[0,1]} \bar{B}_{u,K+1}(\nu_1(\xi,s_\pm),\dots,\nu_Q(\xi,s_\pm)) = O_{n\to\infty}(1), \quad (82)$$

where now we have a different Q and moments depend both on $s\in[0,1]$ and on $\xi\in[0,1]$.

At this point, it suffices to bound $\mathbb{E}\left[\dot{a}_1\dot{a}_2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\right]$ and $\mathbb{E}\left[\left(\dot{a}_1\dot{a}_2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\right)^2\right]$ uniformly over $t\in[0,1/\sqrt{2}]$. There bounds follow from Lemma 3:

$$\sup_{t\in[0,1/\sqrt{2}]} \mathbb{E}\left[\dot{a}_1\dot{a}_2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\right] = O(n^{-1})\prod_{i=1}^{\nu}\left(1_{j_i+r_i\ge 2}O(n^{-(j_i+r_i)/2})\right); \quad (83)$$

$$\sup_{t\in[0,1/\sqrt{2}]} \mathbb{E}\left[\left(\dot{a}_1\dot{a}_2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\right)^2\right] = O(n^{-2})\prod_{i=1}^{\nu} O(n^{-(j_i+r_i)}). \quad (84)$$

This implies

$$\sum_{\substack{j_{1:\nu}\ge 0\\ j_1+\dots+j_\nu=k}} \mathbb{E}\left[\dot{a}_1\dot{a}_2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\right]\mathbb{E}\left[\partial_1^{j_1}\dots\partial_\nu^{j_\nu} f(0_{1:\nu},a_{\nu+1:N})\right] =$$
$$= O\left(n^{-\frac{2+k+\sum_{i=1}^{\nu} r_i}{2}}\right)1_{k+\sum_{i=1}^{\nu} r_i\ge 2s} = O\left(n^{-\frac{2+k+w_{\mathrm{MatMul}}+\bar{w}_{\mathrm{MatMul}}}{2}}\right)1_{k+w_{\mathrm{MatMul}}+\bar{w}_{\mathrm{MatMul}}\ge 2s} \quad (85)$$

uniformly over $t\in[0,1/\sqrt{2}]$, and

$$\sum_{\substack{j_{1:\nu}\ge 0\\ j_1+\dots+j_\nu=K+1}} \sqrt{\mathbb{E}\left[\left(\dot{a}_1\dot{a}_2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\right)^2\right]\mathbb{E}\left[\left(\int_0^1 \partial_1^{j_1}\dots\partial_\nu^{j_\nu} f(\dots)\,d\xi\right)^2\right]}$$
$$= O\left(n^{-\frac{2+(K+1)+\sum_{i=1}^{\nu} r_i}{2}}\right) = O\left(n^{-\frac{3+K+w_{\mathrm{MatMul}}+\bar{w}_{\mathrm{MatMul}}}{2}}\right) \quad (86)$$

uniformly over $t\in[0,1/\sqrt{2}]$.

Take $K=\max(0,2\nu-w_{\mathrm{MatMul}}+\bar{w}_{\mathrm{MatMul}})$; this gives

$$\sup_{t\in[0,1/\sqrt{2}]} \mathbb{E}\left[f(a_1,\dots,a_N)\dot{a}_1\dot{a}_2 a_1^{r_1}\dots a_\nu^{r_\nu}\right] =$$
$$= \begin{cases} O\left(n^{-(w_{\mathrm{MatMul}}+\bar{w}_{\mathrm{MatMul}})/2-1}\right), & w_{\mathrm{MatMul}}+\bar{w}_{\mathrm{MatMul}}\ge 2\nu \\ O\left(n^{-\nu-1}\right), & w_{\mathrm{MatMul}}+\bar{w}_{\mathrm{MatMul}}<2\nu \end{cases} = O(n^{-\nu-1}), \quad (87)$$

which finishes the proof of the proposition. □

See Appendix G.3 for the bound for Eq. (71).

G Technical results

Lemma 10. *Let $\xi_1, \dots, \xi_m$ be real-valued random variables with all of the moments existing. Then*

$$\mathbb{E}\,|\xi_1 \dots \xi_m| \leq \sqrt[m]{\mathbb{E}\,[|\xi_1|^m] \dots \mathbb{E}\,[|\xi_m|^m]}. \tag{88}$$

Proof. By induction on m. Induction base ($m = 1$) is trivial. Induction step follows from Hölder inequality:

$$\mathbb{E}\,|\xi_1 \dots \xi_m| = \mathbb{E}\left[\left(|\xi_1 \dots \xi_{m-1}|^{\frac{m}{m-1}}\right)^{1-\frac{1}{m}} \left(|\xi_m|^m\right)^{\frac{1}{m}}\right] \leq$$
$$\leq \left(\mathbb{E}\left[|\xi_1 \dots \xi_{m-1}|^{\frac{m}{m-1}}\right]\right)^{1-\frac{1}{m}} \left(\mathbb{E}\,[|\xi_m|^m]\right)^{\frac{1}{m}} \leq \left(\mathbb{E}\,[|\xi_1|^m] \dots \mathbb{E}\,[|\xi_{m-1}|^m]\right)^{\frac{1}{m}} \left(\mathbb{E}\,[|\xi_m|^m]\right)^{\frac{1}{m}}. \tag{89}$$

□

G.1 Proof of Lemma 3

Recall $a(s) = \sqrt{1-s}a + \sqrt{s}\tilde{a}$; therefore

$$\dot{a}(t^2) = \tilde{a} - \frac{t}{\sqrt{1-t^2}}a, \qquad \dot{a}(1-t^2) = a - \frac{t}{\sqrt{1-t^2}}\tilde{a}, \tag{90}$$

where "dot" denotes the t-derivative. Hence $\mathbb{E}\,[\dot{a}(t^2)] = \mathbb{E}\,[\dot{a}(1-t^2)] = 0\ \forall t \in [0,1)$ since $\mathbb{E}\,a = \mathbb{E}\,\tilde{a} = 0$.

Swapping t^2 with $1-t^2$ is equivalent to swapping a and $\tilde{a}$. Since a and $\tilde{a}$ have the same moment decay rates and we are only concerned with these rates, we consider only the case t^2 below.

The important property of the above interpolation is the following:

$$a(t^2)\dot{a}(t^2) = \left(t\tilde{a} + \sqrt{1-t^2}a\right)\left(\tilde{a} - \frac{t}{\sqrt{1-t^2}}a\right) = t\tilde{a}^2 - ta^2 + \frac{1-2t^2}{\sqrt{1-t^2}}\tilde{a}a, \tag{91}$$

which implies $\mathbb{E}\,[a(t^2)\dot{a}(t^2)] = 0\ \forall t \in [0,1)$.

Note that $\sup_{t\in[0,1/\sqrt{2}]} |a(t^2)| \leq |a| + |\tilde{a}|$ and $\sup_{t\in[0,1/\sqrt{2}]} |\dot{a}(t^2)| \leq |a| + |\tilde{a}|$. Therefore

$$\sup_{t\in[0,1/\sqrt{2}]} \mathbb{E}\left(\left|[a(t^2)]^j \dot{a}(t^2)\right|^k\right) = \sup_{t\in[0,1/\sqrt{2}]} \mathbb{E}\left(\left|a(t^2)\right|^{jk} \left|\dot{a}(t^2)\right|^k\right) \leq$$
$$\leq \mathbb{E}\left(\sup_{t\in[0,1/\sqrt{2}]} \left|a(t^2)\right|^{jk} \sup_{t\in[0,1/\sqrt{2}]} \left|\dot{a}(t^2)\right|^k\right) \leq \mathbb{E}\left[(|a| + |\tilde{a}|)^{(j+1)k}\right] = O(n^{-(j+1)k/2}). \tag{92}$$

G.2 Proof of Lemma 6

Derivatives wrt g and c. For a nonlinearity ϕ depending elementwise on m vectors and m scalars, denote its derivative wrt j-th vector by $\partial_j^g \phi$ and the derivative wrt j-th scalar by $\partial_j^c \phi$.

A set of equivalence classes of matrices. Since some of the matrices in the program may coincide, define $\mathcal{M}$ as a set of equivalence classes between them, i.e. any $\mu \in \mathcal{M}$ is a subset of $[M_0 + 1 : M]$ and $A^{m_1} = A^{m_2}$ iff $\{m_1, m_2\} \subset \mu$ for some $\mu \in \mathcal{M}$. We define A^μ to be A^m for any $m \in \mu$.

The proof of the lemma. The claim of the lemma follows from the fact that

$$\Psi'(s) = \frac{1}{n}\sum_{\mu}\sum_{\beta_1}\sum_{\alpha,\gamma} \left.\frac{\partial\psi(g_{\beta_1}^1, \dots, g_{\beta_1}^M, c^1, \dots, c^M)}{\partial A_{\alpha\gamma}^\mu}\right|_{\substack{a_1 = a_1(s) \\ \dots \\ a_N = a_N(s)}} \partial_s A_{\alpha\gamma}^\mu(s) \tag{93}$$

and the following lemma:

Lemma 11. $\frac{\partial \psi(g^1_{\beta_1},\ldots,g^M_{\beta_1},c^1,\ldots,c^M)}{\partial A^\mu_{\alpha\gamma}}$ *can be expressed as a sum of a finite, not depending on n, number of terms of the form*

$$\sum_{\beta_2}\cdots\sum_{\beta_w}\pi_{\beta_{1:w}}\tilde{f}_{\alpha,\gamma,\beta_{1:w}},\qquad \tilde{f}_{\alpha,\gamma,\beta_{1:w}}=\begin{cases} f_{\gamma,\beta_{1:w}}, & \text{if } T_m=1 \text{ and } \beta_w=\alpha;\\ f_{\alpha,\beta_{1:w}}, & \text{if } T_m=\top \text{ and } \beta_w=\gamma;\\ 0, & \text{otherwise,}\end{cases} \tag{94}$$

where $w\in[1:M-M_0]$, $m\in[M_0+1:M]$,

$$f_{\gamma,\beta_{1:w}}=\partial_{m_1+1}\psi(g^1_{\beta_1},\ldots,g^M_{\beta_1},c^1,\ldots,c^M)\times \\ \times\left[\prod_{j=2}^{w}\partial^{x_j}_{m_j+1}\phi(g^1_{\beta_j},\ldots,g^{m_{j-1}}_{\beta_j},c^1,\ldots,c^{m_{j-1}})\right]\phi(g^1_\gamma,\ldots,g^{m_w}_\gamma,c^1,\ldots,c^{m_w}) \tag{95}$$

where $M>m_1>\ldots>m_w=m-1\in[M_0:M-1]$, $x_2,\ldots,x_j\in\{g,c\}$ and

$$\pi_{\beta_{1:w}}=\prod_{k=2}^{w}a'_k,\qquad a'_k\in\left\{A^{m_{k-1}+1,T_{m_{k-1}+1}}_{\beta_{k-1}\beta_k},n^{-1}\right\}. \tag{96}$$

Proof. We claim that for any $m_0\in[M_0+1:M-1]$, $\frac{\partial\phi(g^1_{\beta_1},\ldots,g^{m_0}_{\beta_1},c^1,\ldots,c^{m_0})}{\partial A^\mu_{\delta\gamma}}$ can be expressed as a sum of a finite, not depending on n, number of terms of the form

$$\sum_{\beta_2}\cdots\sum_{\beta_w}\pi^{\sigma_{2:w}}_{\beta_{1:w}}\tilde{f}^{m_{0:w}}_{\alpha,\gamma,\beta_{1:w}},\qquad \tilde{f}^{m_{0:w}}_{\alpha,\gamma,\beta_{1:w}}=\begin{cases} f^{m_{0:w}}_{\gamma,\beta_{1:w}}, & \text{if } T_m=1 \text{ and } \beta_w=\alpha;\\ f^{m_{0:w}}_{\alpha,\beta_{1:w}}, & \text{if } T_m=\top \text{ and } \beta_w=\gamma;\\ 0, & \text{otherwise,}\end{cases} \tag{97}$$

where $w\in[1:m_0-M_0]$, $m_0>\ldots>m_w\in[M_0:m_0-1]$,

$$f^{m_{0:w}}_{\gamma,\beta_{1:w}}=\left[\prod_{j=1}^{w}\partial^{x_j}_{m_j+1}\phi(g^1_{\beta_j},\ldots,g^{m_{j-1}}_{\beta_j},c^1,\ldots,c^{m_{j-1}})\right]\phi(g^1_\gamma,\ldots,g^{m_w}_\gamma,c^1,\ldots,c^{m_w}), \tag{98}$$

where $x_1,\ldots,x_j\in\{g,c\}$, and

$$\pi^{\sigma_{2:w}}_{\beta_{1:w}}=\prod_{k=2}^{w}a'_k,\qquad a'_k\in\left\{A^{m_{k-1}+1,T_{m_{k-1}+1}}_{\beta_{k-1}\beta_k},n^{-1}\right\}, \tag{99}$$

where $\sigma_2,\ldots,\sigma_w\in\{\text{Moment},\text{MatMul}\}$. The claim of the lemma will follow by putting $m_0=M$ since ϕ and ψ have the same nature.

Proof by induction. Induction base ($m_0=M_0+1$):

$$\frac{\partial\phi(g^1_{\beta_1},\ldots,g^{M_0+1}_{\beta_1},c^1,\ldots,c^{M_0+1})}{\partial A^\mu_{\alpha\gamma}}= \\ =\partial^g_{M_0+1}\phi(g^1_{\beta_1},\ldots,g^{M_0+1}_{\beta_1},c^1,\ldots,c^{M_0+1})\frac{\partial g^{M_0+1}_{\beta_1}}{\partial A^\mu_{\alpha\gamma}}+ \\ +\partial^c_{M_0+1}\phi(g^1_{\beta_1},\ldots,g^{M_0+1}_{\beta_1},c^1,\ldots,c^{M_0+1})\frac{\partial c^{M_0+1}}{\partial A^\mu_{\alpha\gamma}}= \\ =\partial^g_{M_0+1}\phi(g^1_{\beta_1},\ldots,g^{M_0+1}_{\beta_1},c^1,\ldots,c^{M_0+1})\times \\ \times\frac{\partial}{\partial A^\mu_{\alpha\gamma}}\left(\sum_\delta A^{M_0+1,T_{M_0+1}}_{\beta_1\delta}\phi(g^1_\delta,\ldots,g^{M_0}_\delta,c^1,\ldots,c^{M_0})\right)+ \\ +\partial^c_{M_0+1}\phi(g^1_{\beta_1},\ldots,g^{M_0+1}_{\beta_1},c^1,\ldots,c^{M_0+1})\left[\frac{1}{n}\sum_\delta\frac{\partial\phi(g^1_\delta,\ldots,g^{M_0}_\delta,c^1,\ldots,c^{M_0})}{\partial A^\mu_{\alpha\gamma}}\right]= \\ =\partial^g_{M_0+1}\phi(g^1_{\beta_1},\ldots,g^{M_0+1}_{\beta_1},c^1,\ldots,c^{M_0+1})\times \\ \times\begin{cases}\phi(g^1_\gamma,\ldots,g^{M_0}_\gamma,c^1,\ldots,c^{M_0}), & \text{if } T_{M_0+1}=1,\ \alpha=\beta_1,\ M_0+1\in\mu;\\ \phi(g^1_\alpha,\ldots,g^{M_0}_\alpha,c^1,\ldots,c^{M_0}), & \text{if } T_{M_0+1}=\top,\ \gamma=\beta_1,\ M_0+1\in\mu.\end{cases} \tag{100}$$

It can expressed as a single term of the required form with $w = 1$ and $m_1 = M_0$ if $M_0 + 1 \in \mu$, or as a sum of an empty set of terms otherwise.

Suppose the claim holds for all $m_1 \in [M_0 + 1 : m_0 - 1]$; let us prove it for m_0:

$$\begin{aligned}
&\frac{\partial \phi(g^1_{\beta_1}, \dots, g^{m_0}_{\beta_1}, c^1, \dots, c^{m_0})}{\partial A^\mu_{\alpha\gamma}} = \\
&\qquad = \sum_{m_1=M_0}^{m_0-1} \partial^g_{m_1+1} \phi(g^1_{\beta_1}, \dots, g^{m_0}_{\beta_1}, c^1, \dots, c^{m_0}) \frac{\partial g^{m_1+1}_{\beta_1}}{\partial A^\mu_{\alpha\gamma}} + \\
&\qquad + \sum_{m_1=M_0}^{m_0-1} \partial^c_{m_1+1} \phi(g^1_{\beta_1}, \dots, g^{m_0}_{\beta_1}, c^1, \dots, c^{m_0}) \frac{\partial c^{m_1+1}}{\partial A^\mu_{\alpha\gamma}} = \\
&\qquad = \sum_{m_1=M_0}^{m_0-1} \partial^g_{m_1+1} \phi(g^1_{\beta_1}, \dots, g^{m_0}_{\beta_1}, c^1, \dots, c^{m_0}) \times \\
&\qquad \times \frac{\partial}{\partial A^\mu_{\alpha\gamma}} \left(\sum_{\beta_2} A^{m_1+1, T_{m_1+1}}_{\beta_1\beta_2} \phi(g^1_{\beta_2}, \dots, g^{m_1}_{\beta_2}, c^1, \dots, c^{m_1}) \right) + \\
&+ \sum_{m_1=M_0}^{m_0-1} \partial^c_{m_1+1} \phi(g^1_{\beta_1}, \dots, g^{m_0}_{\beta_1}, c^1, \dots, c^{m_0}) \left[\frac{1}{n} \sum_{\beta_2} \frac{\partial \phi(g^1_{\beta_2}, \dots, g^{m_1}_{\beta_2}, c^1, \dots, c^{m_1})}{\partial A^\mu_{\alpha\gamma}} \right] = \\
&\qquad = \sum_{m_1=M_0}^{m_0-1} \partial_{m_1+1} \phi(g^1_{\beta_1}, \dots, g^{m_0}_{\beta_1}, c^1, \dots, c^{m_0}) \times \\
&\qquad \times \begin{cases} \phi(g^1_\gamma, \dots, g^{m_1}_\gamma, c^1, \dots, c^{m_1}), & \text{if } T_{m_1+1} = 1, \alpha = \beta_1, m_1 + 1 \in \mu \\ \phi(g^1_\alpha, \dots, g^{m_1}_\alpha, c^1, \dots, c^{m_1}), & \text{if } T_{m_1+1} = \top, \gamma = \beta_1, m_1 + 1 \in \mu \end{cases} + \\
&+ \sum_{m_1=M_0}^{m_0-1} \partial^g_{m_1+1} \phi(g^1_{\beta_1}, \dots, g^{m_0}_{\beta_1}, c^1, \dots, c^{m_0}) \sum_{\beta_2} A^{m_1+1, T_{m_1+1}}_{\beta_1\beta_2} \frac{\partial \phi(g^1_{\beta_2}, \dots, g^{m_1}_{\beta_2}, c^1, \dots, c^{m_1})}{\partial A^\mu_{\alpha\gamma}} + \\
&+ \sum_{m_1=M_0}^{m_0-1} \partial^c_{m_1+1} \phi(g^1_{\beta_1}, \dots, g^{m_0}_{\beta_1}, c^1, \dots, c^{m_0}) \left[\frac{1}{n} \sum_{\beta_2} \frac{\partial \phi(g^1_{\beta_2}, \dots, g^{m_1}_{\beta_2}, c^1, \dots, c^{m_1})}{\partial A^\mu_{\alpha\gamma}} \right]. \qquad (101)
\end{aligned}$$

The first sum is a sum of terms of the desired form with $w = 1$. By the induction hypothesis, the second sum is a sum of terms of the form

$$\begin{aligned}
&\partial^g_{m_1+1} \phi(g^1_{\beta_1}, \dots, g^{m_0}_{\beta_1}, c^1, \dots, c^{m_0}) \sum_{\beta_2} A^{m_1+1, T_{m_1+1}}_{\beta_1\beta_2} \sum_{\beta_3} \dots \sum_{\beta_w} \pi^{\sigma_{3:w}}_{\beta_{2:w}} \tilde{f}^{m_{1:w}}_{\alpha,\gamma,\beta_{2:w}} = \\
&\qquad = \sum_{\beta_2} \dots \sum_{\beta_w} \pi^{\sigma_{2:w}}_{\beta_{1:w}} \tilde{f}^{m_{0:w}}_{\alpha,\gamma,\beta_{1:w}}, \qquad (102)
\end{aligned}$$

where we define $\sigma_2 = \mathrm{MatMul}$. As for the third sum, the same holds, but we define $\sigma_2 = \mathrm{Moment}$. □

Due to the lemma above, Ψ' is a sum of $O(1)$ terms of the form

$$\frac{1}{n} \sum_{\beta_1} \dots \sum_{\beta_w} \pi_{\beta_{1:w}} \sum_{\alpha,\gamma} \tilde{f}_{\alpha,\gamma,\beta_{1:w}} \partial_s A^\mu_{\alpha\gamma}, \qquad (103)$$

where $w \in [1 : M - M_0]$ and $\tilde{f}$ and π are functions of $a_1(s), \dots, a_N(s)$. Using the definition of $\tilde{f}$, we can rewrite the expression above equivalently as

$$\frac{1}{n} \sum_{\gamma} \sum_{\beta_1} \dots \sum_{\beta_w} \pi_{\beta_{1:w}} f_{\gamma,\beta_{1:w}} \partial_s A^{\mu, T_m}_{\beta_w \gamma}. \qquad (104)$$

G.3 Proof of Lemma 2: missing results for the "$\dot{a}_1^2$" case

We have to prove analogues of Lemma 8 and Proposition 2. Let us start with the lemma:

Lemma 12. $\left|\Gamma^{(2)}_{r_{1:\nu}}\right| \le Cn^{\nu+1}n^{w_{\text{Moment}}}n^{\tilde{w}_{\text{Moment}}}$, *where C depends neither on n, nor on t.*

Proof. Similarly to Lemma 8, we construct a configuration graph as a union of two path $(\beta_{1:w}, \gamma)$ and $(\tilde{\beta}_{1:\tilde{w}}, \tilde{\gamma})$ with some of the edges erased; we orient the ones that remain according to T_k and $\tilde{T}_k$ respectively. Note that here, since we have $\dot{a}_1^2$ in the definition of $\Gamma^{(2)}$, the last edges of the two paths (which are always present) should coincide.

A configuration $(\gamma, \beta_{1:w}, \tilde{\gamma}, \tilde{\beta}_{1:\tilde{w}})$ lies in $\Gamma^{(2)}_{r_{1:\nu}}$ iff the corresponding configuration graph consists of ν directed multi-edges with multiplicities $r_1 + 2, r_2, \dots, r_\nu$. As before, we need to compute the number of vertices such a graph traverses. If the graph traverses no more than k vertices then the number of configurations satisfying the condition is no more than Cn^k, where C is the number of non-isomorphic directed graphs with $w_{\text{MatMul}} + \tilde{w}_{\text{MatMul}} + 2$ edges and ν multi-edges; obviously, C depends neither on n, nor on t.

We proceed with the same logic as in Lemma 8. Consider a non-directed graph with ν multi-edges. If it has p connected components, it traverses $\nu + p$ vertices. Recall our configuration graph is a union of two paths, having w and $\tilde{w}$ edges respectively, with some of the edges erased, resulting in $w_{\text{MatMul}} + 1$ and $\tilde{w}_{\text{MatMul}} + 1$ edges respectively. Overall, we had to erase $w_{\text{Moment}} + \tilde{w}_{\text{Moment}}$ edges. Given a union of paths, erasing k edges results in introduction of no more than k connected components. Since there are only two paths from the beginning and the last edge of the two is the same, we should have no more than $k + 1$ connected components after erasing k edges.

In our case, $k = w_{\text{Moment}} + \tilde{w}_{\text{Moment}}$. Hence our configuration graph traverses no more than $\nu + 1 + w_{\text{Moment}} + \tilde{w}_{\text{Moment}}$ vertices. Therefore the number of configurations $(\gamma, \beta_{1:w}, \tilde{\gamma}, \tilde{\beta}_{1:\tilde{w}})$ in $\Gamma^{(2)}_{r_{1:\nu}}$ is not greater than $Cn^{\nu+1}n^{w_{\text{Moment}}}n^{\tilde{w}_{\text{Moment}}}$ for some constant C independent on both n and t. □

Let us now turn to the analogue of Proposition 2:

Proposition 3.

$$\sup_{t\in[0,1/\sqrt{2}]} \sup_{\gamma,\beta_{1:w},\tilde{\gamma},\tilde{\beta}_{1:\tilde{w}}} \left|\mathbb{E}\left(\left[f_{\gamma,\beta_{1:w}} f_{\tilde{\gamma},\tilde{\beta}_{1:\tilde{w}}}\right](a_1,\dots,a_N)\dot{a}_1^2 a_1^{r_1}\dots a_\nu^{r_\nu}\right)\right| = O(n^{-\nu}), \tag{105}$$

where $a_1, \dots, a_N$ are functions of $s_\pm$ (recall s_+ denotes t^2 and s_- denotes $1 - t^2$).

Proof. The proof is essentially the same as the proof of Proposition 2. We introduce $f = f_{\gamma,\beta_{1:w}} \times f_{\tilde{\gamma},\tilde{\beta}_{1:\tilde{w}}}$ and Taylor-expand it wrt $a_1, \dots, a_\nu$ at zero up to some order K. Then we bound each of the expansion terms uniformly over $t \in [0, 1/\sqrt{2}]$.

The beginning of proof is the same and we do not reproduce it here. Slight differences start to occur when we bound $\mathbb{E}\left[\dot{a}_1^2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\right]$ and $\mathbb{E}\left[\left(\dot{a}_1^2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\right)^2\right]$ uniformly over $t \in [0, 1/\sqrt{2}]$. By Lemma 3,

$$\sup_{t\in[0,1/\sqrt{2}]} \mathbb{E}\left[\dot{a}_1^2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\right] = O(n^{-1})\left(\prod_{i=1}^{\nu} O(n^{-(j_i+r_i)/2})\right)\prod_{i=2}^{\nu} 1_{j_i+r_i\ge 2}; \tag{106}$$

$$\sup_{t\in[0,1/\sqrt{2}]} \mathbb{E}\left[\left(\dot{a}_1^2 a_1^{j_1+r_1}\dots a_\nu^{j_\nu+r_\nu}\right)^2\right] = O(n^{-2})\prod_{i=1}^{\nu} O(n^{-(j_i+r_i)}), \tag{107}$$

so the asymptotics is a little bit different compared to Proposition 2 due to occurence of the term $\dot{a}_1^2$ instead of $\dot{a}_1\dot{a}_2$: while $\mathbb{E}\left[a_1^{r_1}\dot{a}_1\right] = 0$ when $r_1 \in \{0, 1\}$ by Lemma 3, $\mathbb{E}\left[a_1^{r_1}\dot{a}_1^2\right]$ is never identially

zero. This implies

$$\sum_{\substack{j_{1:\nu}\geq 0\\ j_1+\ldots+j_\nu=k}} \mathbb{E}\left[\dot{a}_1^2 a_1^{j_1+r_1}\ldots a_\nu^{j_\nu+r_\nu}\right]\times\mathbb{E}\left[\partial_1^{j_1}\ldots\partial_\nu^{j_\nu} f(0_{1:\nu},a_{\nu+1:N})\right]=$$

$$=O\left(n^{-\frac{2+k+\sum_{i=1}^{\nu} r_i}{2}}\right)1_{k+\sum_{i=2}^{\nu} r_i\geq 2\nu-2}\leq$$

$$\leq O\left(n^{-\frac{2+k+w_{\mathrm{MatMul}}+\tilde{w}_{\mathrm{MatMul}}}{2}}\right)1_{2+k+w_{\mathrm{MatMul}}+\tilde{w}_{\mathrm{MatMul}}\geq 2\nu} \quad (108)$$

uniformly over $t\in[0,1/\sqrt{2}]$, and

$$\sum_{\substack{j_{1:\nu}\geq 0\\ j_1+\ldots+j_\nu=K+1}} \sqrt{\mathbb{E}\left[\left(\dot{a}_1^2 a_1^{j_1+r_1}\ldots a_\nu^{j_\nu+r_\nu}\right)^2\right]\mathbb{E}\left[\left(\int_0^1 \partial_1^{j_1}\ldots\partial_\nu^{j_\nu} f(\ldots)\,d\xi\right)^2\right]}=$$

$$=O\left(n^{-\frac{2+(K+1)+\sum_{i=1}^{\nu} r_i}{2}}\right)=O\left(n^{-\frac{2+K+w_{\mathrm{MatMul}}+\tilde{w}_{\mathrm{MatMul}}}{2}}\right) \quad (109)$$

uniformly over $t\in[0,1/\sqrt{2}]$.

Recall $f=f_{\gamma,\beta_{1:w}}\times f_{\tilde{\gamma},\tilde{\beta}_{1:\tilde{w}}}$. Take $K=\max(0,2\nu-2-w_{\mathrm{MatMul}}-\tilde{w}_{\mathrm{MatMul}})$; this gives

$$\sup_{t\in[0,1/\sqrt{2}]}\mathbb{E}\left[f(a_{1:N})\dot{a}_1^2 a_1^{r_1}\ldots a_\nu^{r_\nu}\right]=$$

$$=\begin{cases} O\left(n^{-(w_{\mathrm{MatMul}}+\tilde{w}_{\mathrm{MatMul}})/2-1}\right), & w_{\mathrm{MatMul}}+\tilde{w}_{\mathrm{MatMul}}\geq 2\nu-2\\ O\left(n^{-\nu}\right), & w_{\mathrm{MatMul}}+\tilde{w}_{\mathrm{MatMul}}<2\nu-2\end{cases}=O(n^{-\nu}), \quad (110)$$

which finishes the proof of the proposition. □

H Uniform moment controls

Definition 5. *We say f is linearly smooth if it is smooth and any of its derivatives of order $k\geq 1$ is bounded.*

Lemma 13. *Consider a Tensor Program of width n with matrices whose elements are iid and satisfy $\mathbb{E}\,\bar{A}_{\beta\gamma}=0$ and $\mathbb{E}\,|\bar{A}_{\beta\gamma}|^k\leq\nu_k n^{-k/2}$, $\nu_k\in\mathbb{R}_+\ \forall k\geq 1$, and linearly smooth nonlinearity ϕ. Denote the corresponding vectors as $\bar{g}^1,\ldots,\bar{g}^M$ and the corresponding scalars as $\bar{c}^1,\ldots,\bar{c}^M$.*

Then for any $m\in[M_0:M-1]$, any $p\geq 1$, any $n\in\mathbb{N}$, and any collection of weights P,

$$\sup_{\alpha\in[n]}\mathbb{E}\left|\partial_P\phi(\bar{g}^1_\alpha,\ldots,\bar{g}^m_\alpha,\bar{c}^1,\ldots,\bar{c}^m)\right|^p\leq B_{p,|P|}n^{-p|P|/2}, \quad (111)$$

where $B_{p,|P|}$ is a continuous function of a finite number of (scaled) moments ν. Similarly, for a linearly smooth ψ, for any $p\geq 1$, any $n\in\mathbb{N}$, and any collecition of weights P,

$$\sup_{\alpha\in[n]}\mathbb{E}\left|\partial_P\psi(\bar{g}^1_\alpha,\ldots,\bar{g}^M_\alpha,\bar{c}^1,\ldots,\bar{c}^M)\right|^p\leq B_{p,|P|}n^{-p|P|/2}. \quad (112)$$

Proof. Consider the following two claims and induct on $m\in[M_0:M-1]$: for any $p\geq 1$, any $n\in\mathbb{N}$, and any collection of weights P,

C1(m): $\sup_{\alpha\in[n]}\mathbb{E}\left|\sum_\beta \bar{A}^{m+1,T_{m+1}}_{\alpha\beta}\partial_P\phi(\bar{g}^1_\beta,\ldots,\bar{g}^m_\beta,\bar{c}^1,\ldots,\bar{c}^m)\right|^p\leq B_{p,|P|}n^{-p|P|/2}$;

C2(m): $\sup_{\alpha\in[n]}\mathbb{E}\left|\partial_P\phi(\bar{g}^1_\alpha,\ldots,\bar{g}^m_\alpha,\bar{c}^1,\ldots,\bar{c}^m)\right|^p\leq B_{p,|P|}n^{-p|P|/2}$,

where $B_{p,|P|}$ is a continuous function of a finite number of scaled moments ν and C1(m) and C2(m) are short notations for Claim 1(m) and Claim 2(m), respectively. The additional claim of the lemma will follow from Claim 2(M) since ϕ and ψ are of the same nature.

We will make use of the following lemma:

Lemma 14. *For any $n \geq 1$, any $k \geq 1$, any $I \in [n]^k$, there exists $\tilde{B}_k \in C(\mathbb{R}^k)$ such that for any $m \in [M-1]$*

$$\sup_{\alpha \in [n]} \mathbb{E} \left| \prod_{\beta \in I} \bar{A}^{m,T_m}_{\alpha\beta} \right| \leq \tilde{B}_k(\nu_1, \ldots, \nu_k) n^{-k/2}. \tag{113}$$

Proof. Let $\mathbb{I}$ be the set of unique indices in I. Define $p_\beta = |\{j : \ I_j = \beta\}|$. Note that $\sum_{\beta \in \mathbb{I}} p_\beta = k$. Then, since elements of the matrix $\bar{A}^m$ are independent,

$$\mathbb{E} \left| \prod_{\beta \in I} \bar{A}^{m,T_m}_{\alpha\beta} \right| = \prod_{\beta \in \mathbb{I}} \mathbb{E} \left| \bar{A}^{m,T_m}_{\alpha\beta} \right|^{p_\beta} \leq \prod_{\beta \in \mathbb{I}} \nu_{p_\beta} n^{-p_\beta/2} = n^{-k/2} \prod_{\beta \in \mathbb{I}} \nu_{p_\beta}, \tag{114}$$

and it does not depend on α. Therefore the claim is valid for $\tilde{B}_k(\nu_1, \ldots, \nu_k) = \prod_{\beta \in \mathbb{I}} \nu_{p_\beta}$. □

H.1 Induction base

Both claims are trivial for $m = M_0$ and $|P| > 0$ since $\bar{g}^1, \ldots, \bar{g}^{M_0}$ do not depend on any weights.

Claim 2 is trivial for $|P| = 0$ since $\bar{g}^1_\alpha, \ldots, \bar{g}^{M_0}_\alpha \sim \mathcal{N}(\mu^{in}, \Sigma^{in})$ — they do not depend on n. Let us elaborate Claim 1 for $|P| = 0$. For even p,

$$\mathbb{E} \left| \sum_\beta \bar{A}^{M_0+1,T_{M_0+1}}_{\alpha\beta} \phi(\bar{g}^1_\beta, \ldots, \bar{g}^{M_0}_\beta, \bar{c}^1, \ldots, \bar{c}^{M_0}) \right|^p =$$

$$= \sum_{\beta_1} \cdots \sum_{\beta_p} \mathbb{E} \left[\prod_{i=1}^p \bar{A}^{M_0+1,T_{M_0+1}}_{\alpha\beta_i} \phi(\bar{g}^1_{\beta_i}, \ldots, \bar{g}^{M_0}_{\beta_i}, \bar{c}^1, \ldots, \bar{c}^{M_0}) \right]. \tag{115}$$

Since $\bar{g}^1, \ldots, \bar{g}^{M_0}$ do not depend on $\bar{A}^{M_0+1}$ and $\mathbb{E}\, \bar{A}^{M_0+1,T_{M_0+1}}_{\alpha\beta_i} = 0$, only those expectations survive for which the indices $\beta_1, \ldots, \beta_p$ are coupled into groups of even sizes. We get the maximal growth rate when these indices are paired; in this case, only half of the sums survive, which gives the growth rate $O(1)$. The leading constant is finite and depends only on ν_2, in a continuous way, and does not depend on α by Lemma 14. The case of odd p follows from Cauchy-Schwarz.

H.2 Claim 2(m) implies Claim 1(m)

For this section, let us omit the superscript $m+1$ for $\bar{A}$ and the subscript $m+1$ for T, i.e. define $\bar{A}^T = \bar{A}^{m+1,T_{m+1}}$. As one shall further see, proving Claim 1 from Claim 2 never requires dealing with other weight matrices explicitly.

For even p,

$$\mathbb{E} \left| \sum_\beta \bar{A}^T_{\alpha\beta} \partial_P \phi(\bar{g}^1_\beta, \ldots, \bar{g}^m_\beta, \bar{c}^1, \ldots, \bar{c}^m) \right|^p = \mathbb{E} \left[\sum_{\beta_1, \ldots, \beta_p} \prod_{i=1}^p \bar{A}^T_{\alpha\beta_i} \partial_P \phi(\bar{g}^1_{\beta_i}, \ldots, \bar{g}^M_{\beta_i}, \bar{c}^1, \ldots, \bar{c}^m) \right] \tag{116}$$

A naive bound that uses Cauchy-Schwarz and Claim 2(m) results in $O(n^p) O(n^{-p/2}) O(n^{-p|P|/2}) = O(n^{-p(|P|-1)/2})$ — larger than required. Our approach will be to expand $\partial_P \phi(\bar{g}^1_{\beta_i}, \ldots, \bar{g}^M_{\beta_i}, \bar{c}^1, \ldots, \bar{c}^m)$ with respect to $\bar{A}^T_{\alpha\beta_i}$ for $i \in [p]$ at zero. The fastest growing term in this expansion will be the zero-order term. Since it will not depend on $\bar{A}^T_{\alpha\beta_i}$ for $i \in [p]$, the expectation above will become $O(n^{p/2}) O(n^{-p/2}) O(n^{-p|P|/2}) = O(n^{-p|P|/2})$ since weights have zero mean and therefore only a half of the sums will survive.

It is more convenient to expand $\partial_P \phi(\bar{g}^1_{\beta_i}, \ldots, \bar{g}^M_{\beta_i}, \bar{c}^1, \ldots, \bar{c}^m)$ not with respect to $\bar{A}^T_{\alpha\beta_i}$ for $i \in [p]$ directly, but with respect to a single parameter that interpolates $\bar{A}^T_{\alpha\beta_i}$ to zero. More specifically, let $\mathbb{I} = \{(\alpha, \beta_1), \ldots, (\alpha, \beta_p)\}$ if $T = 1$, let $\mathbb{I} = \{(\beta_1, \alpha), \ldots, (\beta_p, \alpha)$ otherwise, let $\mathring{A}^{\mathbb{I}}_{\beta'\gamma'} = \bar{A}_{\beta'\gamma'} 1_{(\beta',\gamma') \notin \mathbb{I}}$, and let $\bar{A}^{\mathbb{I}}(s) = (1-s)\mathring{A}^{\mathbb{I}} + s\bar{A}$. Denote the corresponding $\bar{g}^m$ as $\bar{g}^{m;\mathbb{I}}(s)$ and the

corresponding $\bar{c}^m$ as $\bar{c}^{m;\mathbb{I}}(s)$. For brevity, we will also denote $\phi(\bar{g}_\beta^{1:m;\mathbb{I}}(s), \bar{c}^{1:m;\mathbb{I}}(s))$ as $\phi_\beta^{m;\mathbb{I}}(s)$. For even p, expanding $\partial_P \phi(\bar{g}_{\beta_i}^1, \ldots, \bar{g}_{\beta_i}^M, \bar{c}^1, \ldots, \bar{c}^m) = \partial_P \phi_\beta^{m;\mathbb{I}}(1)$ with respect to s at $s = 0$,

$$\mathbb{E}\left|\sum_\beta \bar{A}_{\alpha\beta}^T \partial_P \phi(\bar{g}_\beta^1, \ldots, \bar{g}_\beta^m, \bar{c}^1, \ldots, \bar{c}^m)\right|^p =$$

$$= \mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p} \prod_{i=1}^p \bar{A}_{\alpha\beta_i}^T \partial_P \phi(\bar{g}_{\beta_i}^1, \ldots, \bar{g}_{\beta_i}^M, \bar{c}^1, \ldots, \bar{c}^m)\right] =$$

$$= \mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p} \prod_{i=1}^p \bar{A}_{\alpha\beta_i}^T \left(\underbrace{\sum_{k=0}^{p-1} \frac{1}{k!} \partial_t^k \partial_P \phi_{\beta_i}^{m;\mathbb{I}}(0)}_{M_i} + \underbrace{\frac{1}{(p-1)!} \int_0^1 (1-s)^{p-1} \partial_t^p \partial_P \phi_{\beta_i}^{m;\mathbb{I}}(s)\, dt}_{N_i}\right)\right] =$$

$$= \mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p} \left(\prod_{i=1}^p \bar{A}_{\alpha\beta_i}^T\right) \sum_{I \subseteq [p]} \left(\prod_{i \in I} M_i\right) \times \left(\prod_{i \in I^c} N_i\right)\right]. \quad (117)$$

We are going to bound this expression separately for each $I \subseteq [p]$. Note the following:

$$\partial_t^p \partial_P \phi_{\beta_i}^{m;\mathbb{I}}(s) = \sum_{j_1,\ldots,j_p=1}^p \partial_{P+\sum_{i'=1}^p [(\alpha,\beta_{j_{i'}})]^T} \phi_{\beta_i}^{m;\mathbb{I}}(s) \prod_{i'=1}^p \bar{A}_{\alpha\beta_{j_{i'}}}^T; \quad (118)$$

$$\partial_t^k \partial_P \phi_{\beta_i}^{m;\mathbb{I}}(0) = \sum_{j_1,\ldots,j_k=1}^p \partial_{P+\sum_{i'=1}^k [(\alpha,\beta_{j_{i'}})]^T} \phi_{\beta_i}^{m;\mathbb{I}}(0) \prod_{i'=1}^k \bar{A}_{\alpha\beta_{j_{i'}}}^T. \quad (119)$$

We see that by Claim 2(m) and Cauchy-Schwarz, the fastest growing term must be $\partial_P \phi_{\beta_i}^{m;\mathbb{I}}(0)$, i.e. the one with $k = 0$. This term is contained in M_i; therefore we suspect the fastest growing subset I must be $[p]$. We start our analysis with that one.

Let $I = [p]$:

$$\mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p} \left(\prod_{i=1}^p \bar{A}_{\alpha\beta_i}^T M_i\right)\right] = \mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p} \left(\prod_{i=1}^p \bar{A}_{\alpha\beta_i}^T \sum_{k=0}^{p-1} \frac{1}{k!} \partial_t^k \partial_P \phi_{\beta_i}^{m;\mathbb{I}}(0)\right)\right] =$$

$$= \mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p} \sum_{k_1,\ldots,k_p=0}^{p-1} \prod_{i=1}^p \left(\bar{A}_{\alpha\beta_i}^T \frac{1}{k_i!} \partial_t^{k_i} \partial_P \phi_{\beta_i}^{m;\mathbb{I}}(0)\right)\right] =$$

$$= \mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p} \sum_{k_1,\ldots,k_p=0}^{p-1} \prod_{i=1}^p \left(\frac{1}{k_i!} \sum_{j_{1:k_i}=1}^p \partial_{P+\sum_{i'=1}^{k_i} [(\alpha,\beta_{j_{i'}})]} \phi_{\beta_i}^{m;\mathbb{I}}(0) \bar{A}_{\alpha\beta_i}^T \prod_{i'=1}^{k_i} \bar{A}_{\alpha\beta_{j_{i'}}}^T\right)\right]. \quad (120)$$

The fastest growing term is that with $k_1, \ldots, k_p = 0$; using independency between $\partial_P \phi_{\beta_i}^{m;\mathbb{I}}(0)$ and $\bar{A}_{\alpha\beta_i}^T$ for any $i \in [p]$, we get

$$\mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p} \left(\prod_{i=1}^p \bar{A}_{\alpha\beta_i}^T M_i\right)\right] \leq C\mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p} \prod_{i=1}^p \left(\partial_P \phi_{\beta_i}^{m;\mathbb{I}}(0) \bar{A}_{\alpha\beta_i}^T\right)\right] =$$

$$= C \sum_{\beta_1,\ldots,\beta_p} \mathbb{E}\left[\prod_{i=1}^p \bar{A}_{\alpha\beta_i}^T\right] \mathbb{E}\left[\prod_{i=1}^p \partial_P \phi_{\beta_i}^{m;\mathbb{I}}(0)\right] \leq C\tilde{B}_p(\nu_1, \ldots, \nu_p) \prod_{i=1}^p \sqrt[p]{\mathbb{E}\left[\left(\partial_P \phi_{\beta_i}^{m;\mathbb{I}}(0)\right)^p\right]}, \quad (121)$$

where the last inequality holds due to Lemma 10 and since $\bar{A}^T$ has zero expectation. Using induction hypothesis and redefining the constants appropriately, we get this quantity to be bounded by $B_{p,|P|}n^{-p|P|/2}$.

We proceed our analysis with $I = [p]\setminus\{i^c\}$. While we will not be able to make use of zero expectation of $\bar{A}^T$ as before, we will observe slower growth rate of the corresponding term simply by applying Cauchy-Schwarz. Indeed, intuitively,

$$\mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p}\left(\prod_{i=1}^{p}\bar{A}^T_{\alpha\beta_i}\right)\left(\prod_{i\in[p]\setminus\{i^c\}}M_i\right)\times(N_{i^c})\right] = \\ = O(n^p)O(n^{-p/2})O(n^{-(p-1)|P|/2})O(n^{-(p+|P|)/2})O(n^{-p/2}) = O(n^{-p(|P|+1)/2}), \quad (122)$$

which is even better than needed. Working more formally, we get

$$\mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p}\left(\prod_{i=1}^{p}\bar{A}^T_{\alpha\beta_i}\right)\left(\prod_{i\in[p]\setminus\{i^c\}}M_i\right)\times(N_{i^c})\right] = \\ = \mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p}\left(\prod_{i=1}^{p}\bar{A}^T_{\alpha\beta_i}\right)\left(\prod_{i\in[p]\setminus\{i^c\}}M_i\right)\times\right. \\ \left.\times\int_0^1(1-s)^{p-1}\left(\prod_{i'=1}^{p}\bar{A}^T_{\alpha\beta_{j_{i'}}}\right)\sum_{j_1,\ldots,j_p=1}^{p}\partial_{P+\sum_{i'=1}^{p}[(\alpha,\beta_{j_{i'}})]^T}\phi^{m;\mathbb{I}}_{\beta_{i^c}}(s)\,dt\right] = \\ = \sum_{j_1,\ldots,j_p=1}^{p}\int_0^1(1-s)^{p-1}\times \\ \times\,\mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p}\left(\prod_{i=1}^{p}\bar{A}^T_{\alpha\beta_i}\right)\left(\prod_{i\in[p]\setminus\{i^c\}}M_i\right)\left(\prod_{i'=1}^{p}\bar{A}^T_{\alpha\beta_{j_{i'}}}\right)\partial_{P+\sum_{i'=1}^{p}[(\alpha,\beta_{j_{i'}})]^T}\phi^{m;\mathbb{I}}_{\beta_{i^c}}(s)\right]dt. \quad (123)$$

Applying Cauchy-Schwarz, we get

$$\mathbb{E}\left|\sum_{\beta_1,\ldots,\beta_p}\left(\prod_{i=1}^{p}\bar{A}^T_{\alpha\beta_i}\right)\left(\prod_{i\in[p]\setminus\{i^c\}}M_i\right)\left(\prod_{i'=1}^{p}\bar{A}^T_{\alpha\beta_{j_{i'}}}\right)\partial_{P+\sum_{i'=1}^{p}[(\alpha,\beta_{j_{i'}})]^T}\phi^{m;\mathbb{I}}_{\beta_{i^c}}(s)\right| \leq \\ \leq \sum_{\beta_1,\ldots,\beta_p}\sqrt{\mathbb{E}\left[\left(\prod_{i=1}^{p}\bar{A}^T_{\alpha\beta_i}\right)\left(\prod_{i'=1}^{p}\bar{A}^T_{\alpha\beta_{j_{i'}}}\right)\partial_{P+\sum_{i'=1}^{p}[(\alpha,\beta_{j_{i'}})]^T}\phi^{m;\mathbb{I}}_{\beta_{i^c}}(s)\right]^2\mathbb{E}\left[\prod_{i\in[p]\setminus\{i^c\}}M_i\right]^2}. \quad (124)$$

Let us work out the two multiplicants separately:

$$\mathbb{E}\left[\left(\prod_{i=1}^{p}\bar{A}^T_{\alpha\beta_i}\right)\left(\prod_{i'=1}^{p}\bar{A}^T_{\alpha\beta_{j_{i'}}}\right)\partial_{P+\sum_{i'=1}^{p}[(\alpha,\beta_{j_{i'}})]^T}\phi^{m;\mathbb{I}}_{\beta_{i^c}}(s)\right]^2 \leq \\ \leq \mathbb{E}\left[\left(\prod_{i=1}^{p}\bar{A}^T_{\alpha\beta_i}\right)^2\left(\prod_{i'=1}^{p}\bar{A}^T_{\alpha\beta_{j_{i'}}}\right)^2\right]\mathbb{E}\left[\partial_{P+\sum_{i'=1}^{p}[(\alpha,\beta_{j_{i'}})]^T}\phi^{m;\mathbb{I}}_{\beta_{i^c}}(s)\right]^2 \leq \\ \leq \tilde{B}_p(\nu_1,\ldots,\nu_{4p})n^{-2p}B_{p,|P|}n^{-p-|P|}. \quad (125)$$

The bound is uniform over α due to Lemma 14 and uniform over $t\in[0,1]$ by induction hypothesis. As before, we redefine $B_{p,|P|}$, so the quantity is bounded by $B_{p,|P|}n^{-3p-|P|}$.

For the second multiplicant, we will need the following estimate.

$$\mathbb{E}\left|\partial_t^k \partial_P \phi_{\beta_i}^{m;\mathbb{I}}(0)\right|^p = \mathbb{E}\left|\sum_{j_1,\dots,j_k=1}^{p} \partial_{P+\sum_{i'=1}^{k}[(\alpha,\beta_{j_{i'}})]}\phi_{\beta_i}^{m;\mathbb{I}}(0)\prod_{i'=1}^{k}\bar{A}^T_{\alpha\beta_{j_{i'}}}\right|^p =$$

$$= \mathbb{E}\sum_{j_1^1,\dots,j_k^1=1}^{p}\dots\sum_{j_1^p,\dots,j_k^p=1}^{p}\prod_{l=1}^{p}\left[\partial_{P+\sum_{i'=1}^{k}[(\alpha,\beta_{j_{i'}^l})]}\phi_{\beta_i}^{m;\mathbb{I}}(0)\prod_{i'=1}^{k}\bar{A}^T_{\alpha\beta_{j_{i'}^l}}\right] =$$

$$= \sum_{j_1^1,\dots,j_k^1=1}^{p}\dots\sum_{j_1^p,\dots,j_k^p=1}^{p}\mathbb{E}\left[\prod_{l=1}^{p}\partial_{P+\sum_{i'=1}^{k}[(\alpha,\beta_{j_{i'}^l})]}\phi_{\beta_i}^{m;\mathbb{I}}(0)\right]\mathbb{E}\left[\prod_{l=1}^{p}\prod_{i'=1}^{k}\bar{A}^T_{\alpha\beta_{j_{i'}^l}}\right] \leq$$

$$\leq \sum_{j_1^1,\dots,j_k^1=1}^{p}\dots\sum_{j_1^p,\dots,j_k^p=1}^{p}\prod_{l=1}^{p}\sqrt[p]{\mathbb{E}\left[\left(\partial_{P+\sum_{i'=1}^{k}[(\alpha,\beta_{j_{i'}^l})]}\phi_{\beta_i}^{m;\mathbb{I}}(0)\right)^p\right]}\mathbb{E}\left[\prod_{l=1}^{p}\prod_{i'=1}^{k}\bar{A}^T_{\alpha\beta_{j_{i'}^l}}\right] \leq$$

$$\leq \sum_{j_1^1,\dots,j_k^1=1}^{p}\dots\sum_{j_1^p,\dots,j_k^p=1}^{p} B_{p,|P|+k}n^{-p(|P|+k)/2}\tilde{B}_p n^{-pk/2} \tag{126}$$

by induction hypothesis; again, the bound is uniform over α by Lemma 14. Abusing notation, since $k \in [p]$ and we can redefine the constants appropriately, and say that the quantity above is bounded by $B_{p,|P|}n^{-p(k+|P|/2)}$.

Let us work out the second multiplicant now:

$$\mathbb{E}\left[\prod_{i\in I}\left(\sum_{k=0}^{p-1}\frac{1}{k!}\partial_t^k\partial_P\phi_{\beta_i}^{m;\mathbb{I}}(0)\right)\right]^2 = \mathbb{E}\left[\sum_{k_{i\in I}^{j\in[2]}=0}^{p-1}\prod_{i\in I,j\in[2]}\left(\frac{1}{k_i^j!}\partial_t^{k_i^j}\partial_P\phi_{\beta_i}^{m;\mathbb{I}}(0)\right)\right] \leq$$

$$\leq \sum_{k_{i\in I}^{j\in[2]}=0}^{p-1}\prod_{i\in I,j\in[2]}\frac{1}{k_i^j!}\sqrt[p]{\mathbb{E}\left[\left(\partial_t^{k_i^j}\partial_P\phi_{\beta_i}^{m;\mathbb{I}}(0)\right)^p\right]} \leq \sum_{k_{i\in I}^{j\in[2]}=0}^{p-1}\prod_{i\in I,j\in[2]}\frac{1}{k_i^j!}B_{p,|P|}n^{-k_i^j-|P|/2}, \tag{127}$$

where we applied Lemma 10. The fastest growing term of the sum is the one with all $k_i^j = 0$. Therefore redefining $B_{p,|P|}$ appropriately, we get the quantity to be bounded by $B_{p,|P|}n^{-|P|(p-1)}$.

Going back and using that $|(1-s)^{p-1}| \leq 1$,

$$\mathbb{E}\left|\sum_{\beta_1,\ldots,\beta_p}\left(\prod_{i=1}^{p}\bar{A}^T_{\alpha\beta_i}\right)\left(\prod_{i\in[p]\setminus\{i^c\}}M_i\right)\times(N_{i^c})\right| \leq$$
$$\leq \sum_{j_1,\ldots,j_p=1}^{p}\sup_{t\in[0,1]}\mathbb{E}\left|\sum_{\beta_1,\ldots,\beta_p}\left(\prod_{i=1}^{p}\bar{A}^T_{\alpha\beta_i}\right)\left(\prod_{i\in[p]\setminus\{i^c\}}M_i\right)\times\right.$$
$$\left.\times\left(\prod_{i'=1}^{p}\bar{A}^T_{\alpha\beta_{j_{i'}}}\right)\partial_{P+\sum_{i'=1}^{p}[(\alpha,\beta_{j_{i'}})]^T}\phi^{m;\mathbb{I}}_{\beta_{i^c}}(s)\right| \leq$$
$$\leq \sum_{j_1,\ldots,j_p=1}^{p}\sum_{\beta_1,\ldots,\beta_p}\sqrt{\mathbb{E}\left[\prod_{i\in[p]\setminus\{i^c\}}M_i\right]^2}\times$$
$$\times\sqrt{\sup_{t\in[0,1]}\mathbb{E}\left[\left(\prod_{i=1}^{p}\bar{A}^T_{\alpha\beta_i}\right)\left(\prod_{i'=1}^{p}\bar{A}^T_{\alpha\beta_{j_{i'}}}\right)\partial_{P+\sum_{i'=1}^{p}[(\alpha,\beta_{j_{i'}})]^T}\phi^{m;\mathbb{I}}_{\beta_{i^c}}(s)\right]^2} \leq$$
$$\leq \sum_{\beta_1,\ldots,\beta_p}\sum_{j_1,\ldots,j_p=1}^{p}\sqrt{B^2_{p,|P|}n^{-3p-|P|}n^{-|P|(p-1)}} = \sum_{\beta_1,\ldots,\beta_p}\sum_{k_1,\ldots,k_p=1}^{p}B_{p,|P|}n^{-3p/2}n^{-|P|p/2}. \tag{128}$$

Redefining $B_{p,|P|}$ appropriately, we get the quantity to be bounded by $B_{p,|P|}n^{-p(1+|P|)/2)}$. Hence $\forall I \neq [p]$ the corresponding term is bounded by $B_{p,|P|}n^{-p(1+|P|)/2}$ as well.

As for the case $I \subsetneq [p]\setminus\{i^c\}$, one can follow the same lines as for $I = [p]\setminus\{i^c\}$. An intuitive bound will result in

$$\mathbb{E}\left[\sum_{\beta_1,\ldots,\beta_p}\left(\prod_{i=1}^{p}\bar{A}^T_{\alpha\beta_i}\right)\left(\prod_{i\in I}M_i\right)\times\left(\prod_{i\in I^c}N_i\right)\right] =$$
$$= O(n^p)O(n^{-p/2})O(n^{-|I||P|/2})O(n^{-(p+|P|)(p-|I|)/2})O(n^{-p(p-|I|)/2}) =$$
$$= O(n^{p/2})O(n^{-(p^2+|P|p-|I|p)/2})O(n^{-p(p-|I|)/2}) =$$
$$= O(n^{p/2})O(n^{-p|P|/2})O(n^{-p(p-|I|)}) = O(n^{-p(|P|+1)/2})O(n^{-p(p-|I|-1)}), \tag{129}$$

which is even better than for $I = [p]\setminus\{i^c\}$.

Putting all togeter, we get $\mathbb{E}\left|\sum_\beta \bar{A}^T_{\alpha\beta}\partial_P\phi(\bar{g}^1_\beta,\ldots,\bar{g}^m_\beta,\bar{c}^1,\ldots,\bar{c}^m)\right|^p \leq B_{p,|P|}n^{-p|P|/2}$ $\forall P \in ([n]\times[n])^k$ for any even p. The same bound for odd p follows from Cauchy-Schwarz.

H.3 Claim 1$(M_0 : m-1)$ and Claim 2$(M_0 : m-1)$ imply Claim 2(m)

We start with the case $|P| = 0$:

$$\mathbb{E}\,|\phi(\bar{g}^1_\alpha, \dots, \bar{g}^m_\alpha, \bar{c}^1, \dots, \bar{c}^m)|^p \le \mathbb{E}\Bigg(|\phi(\bar{g}^1_\alpha, \dots, \bar{g}^{M_0}_\alpha, 0, \dots, 0, \bar{c}^1, \dots, \bar{c}^{M_0}, 0, \dots, 0)| +$$
$$+ L \sum_{m'=M_0+1}^{m} \Bigg[\Bigg|\sum_\beta \bar{A}^{m',T_{m'}}_{\alpha\beta} \phi(\bar{g}^1_\beta, \dots, \bar{g}^{m'-1}_\beta, \bar{c}^1, \dots, \bar{c}^{m'-1})\Bigg| +$$
$$+ \Bigg|\frac{1}{n}\sum_\beta \phi(\bar{g}^1_\beta, \dots, \bar{g}^{m'-1}_\beta, \bar{c}^1, \dots, \bar{c}^{m'-1})\Bigg|\Bigg]\Bigg)^p \le$$
$$\le 2^p \mathbb{E}\,|\phi(\bar{g}^1_\alpha, \dots, \bar{g}^{M_0}_\alpha, 0, \dots, 0, \bar{c}^1, \dots, \bar{c}^{M_0}, 0, \dots, 0)|^p +$$
$$+ 2^p L^p (m - M_0)^p \max_{m' \in [M_0+1:m]} \mathbb{E}\left|\sum_\beta \bar{A}^{m',T_{m'}}_{\alpha\beta} \phi(\bar{g}^1_\beta, \dots, \bar{g}^{m'-1}_\beta, \bar{c}^1, \dots, \bar{c}^{m'-1})\right|^p +$$
$$+ 2^p L^p (m - M_0)^p \max_{m' \in [M_0+1:m]} \frac{1}{n}\sum_\beta \mathbb{E}\left|\phi(\bar{g}^1_\beta, \dots, \bar{g}^{m'-1}_\beta, \bar{c}^1, \dots, \bar{c}^{m'-1})\right|^p \le$$
$$\le 2^p \mathbb{E}\,|\phi(\bar{g}^1_\alpha, \dots, \bar{g}^{M_0}_\alpha, 0, \dots, 0, \bar{c}^1, \dots, \bar{c}^{M_0}, 0, \dots, 0)|^p + 2^p L^p (m - M_0)^p B_{p,0}, \quad (130)$$

where L is the Lipschitz constant of ϕ; redefine $B_{p,0}$ as the above quantity. The bound is uniform over α by the induction hypothesis.

Consider now the case $|P| > 0$. We shall think of P as a multi-set of triples (μ, β, γ), where $\mu \in \mathcal{M}$ indicates the "layer", while β and γ indicate the matrix element. Let $(\mu, \beta, \gamma)^\top = (\mu, \gamma, \beta)$. One can express the derivative wrt weights using the following recurrence:

$$\partial_P \phi(\bar{g}^1_\alpha, \dots, \bar{g}^m_\alpha, \bar{c}^1, \dots, \bar{c}^m) =$$
$$= \sum_{j=1}^{|P|} \sum_{m_1, \dots, m_j = M_0+1}^{m} \sum_{x_1, \dots, x_j \in \{g,c\}} \partial^{x_1}_{m_1} \dots \partial^{x_j}_{m_j} \phi(\bar{g}^1_\alpha, \dots, \bar{g}^m_\alpha, \bar{c}^1, \dots, \bar{c}^m) \times$$
$$\times \frac{1}{j!} \sum_{s:\, P \to [j]} \prod_{i=1}^{j} \begin{cases} \partial_{s^{-1}(i)} \bar{g}^{m_i}_\alpha, & x_i = g \\ \partial_{s^{-1}(i)} \bar{c}^{m_i}, & x_i = c. \end{cases} \quad (131)$$

$$\partial_P \bar{g}^m_\alpha = \partial_P \left(\sum_\beta \bar{A}^{m,T_m}_{\alpha\beta} \phi\left(\bar{g}^1_\beta, \dots, \bar{g}^{m-1}_\beta, \bar{c}^1, \dots, \bar{c}^{m-1}\right)\right) =$$
$$= \sum_{(\mu,\beta,\gamma) \in P^{T_m}} 1_{m \in \mu} 1_{\alpha = \beta} \partial_{P \setminus \{(\mu,\beta,\gamma)^{T_m}\}} \phi\left(\bar{g}^1_\gamma, \dots, \bar{g}^{m-1}_\gamma, \bar{c}^1, \dots, \bar{c}^{m-1}\right) +$$
$$+ \sum_\beta \bar{A}^{m,T_m}_{\alpha\beta} \partial_P \phi\left(\bar{g}^1_\beta, \dots, \bar{g}^{m-1}_\beta, \bar{c}^1, \dots, \bar{c}^{m-1}\right). \quad (132)$$

$$\partial_P \bar{c}^m = \partial_P \left(\frac{1}{n}\sum_\beta \phi\left(\bar{g}^1_\beta, \dots, \bar{g}^{m-1}_\beta, \bar{c}^1, \dots, \bar{c}^{m-1}\right)\right) =$$
$$= \frac{1}{n}\sum_\beta \partial_P \phi\left(\bar{g}^1_\beta, \dots, \bar{g}^{m-1}_\beta, \bar{c}^1, \dots, \bar{c}^{m-1}\right). \quad (133)$$

For even p,

$$\mathbb{E}\,|\partial_P\phi(\bar g^1_\alpha,\dots,\bar g^m_\alpha,\bar c^1,\dots,\bar c^m)|^p =$$
$$= \mathbb{E}\left(\sum_{j=1}^{|P|}\sum_{m_1,\dots,m_j=M_0+1}^{m}\sum_{x_1,\dots,x_j\in\{g,c\}}\partial^{x_1}_{m_1}\dots\partial^{x_j}_{m_j}\phi(\bar g^1_\alpha,\dots,\bar g^m_\alpha,\bar c^1,\dots,\bar c^m)\times\right.$$
$$\left.\times\frac{1}{j!}\sum_{s\colon P\twoheadrightarrow[j]}\prod_{i=1}^{j}\begin{cases}\partial_{s^{-1}(i)}\bar g^{m_i}_\alpha, & x_i=g\\ \partial_{s^{-1}(i)}\bar c^{m_i}, & x_i=c\end{cases}\right)^p =$$
$$= \sum_{j_1,\dots,j_p=1}^{|P|}\sum_{m^1_1,\dots,m^1_j=M_0+1}^{m}\dots\sum_{m^p_1,\dots,m^p_j=M_0+1}^{m}\sum_{x^1_1,\dots,x^1_j\in\{g,c\}}\dots\sum_{x^p_1,\dots,x^p_j\in\{g,c\}}$$
$$\mathbb{E}\left[\prod_{r=1}^{p}\partial^{x^r_1}_{m^r_1}\dots\partial^{x^r_j}_{m^r_j}\phi(\bar g^1_\alpha,\dots,\bar g^m_\alpha,\bar c^1,\dots,\bar c^m)\frac{1}{j_r!}\sum_{s\colon P\twoheadrightarrow[j_r]}\prod_{i=1}^{j_r}\begin{cases}\partial_{s^{-1}(i)}\bar g^{m^r_i}_\alpha, & x^r_i=g\\ \partial_{s^{-1}(i)}\bar c^{m^r_i}, & x^r_i=c\end{cases}\right]. \quad (134)$$

Since the sums have $O(1)$ terms, it suffices to bound the following expectations:

$$\mathbb{E}\left|\prod_{r=1}^{p}\partial^{x^r_1}_{m^r_1}\dots\partial^{x^r_j}_{m^r_j}\phi(\bar g^1_\alpha,\dots,\bar g^m_\alpha,\bar c^1,\dots,\bar c^m)\prod_{i=1}^{j_r}\partial_{s^{-1}(i)}\bar g^{m^r_i}_\alpha\right| \le$$
$$\le\sqrt{\mathbb{E}\left[\prod_{r=1}^{p}\partial^{x^r_1}_{m^r_1}\dots\partial^{x^r_j}_{m^r_j}\phi(\bar g^1_\alpha,\dots,\bar g^m_\alpha,\bar c^1,\dots,\bar c^m)\right]^2\mathbb{E}\left[\prod_{i=1}^{j_r}\partial_{s^{-1}(i)}\bar g^{m^r_i}_\alpha\right]^2}, \quad (135)$$

and

$$\mathbb{E}\left|\prod_{r=1}^{p}\partial^{x^r_1}_{m^r_1}\dots\partial^{x^r_j}_{m^r_j}\phi(\bar g^1_\alpha,\dots,\bar g^m_\alpha,\bar c^1,\dots,\bar c^m)\prod_{i=1}^{j_r}\partial_{s^{-1}(i)}\bar c^{m^r_i}\right| \le$$
$$\le\sqrt{\mathbb{E}\left[\prod_{r=1}^{p}\partial^{x^r_1}_{m^r_1}\dots\partial^{x^r_j}_{m^r_j}\phi(\bar g^1_\alpha,\dots,\bar g^m_\alpha,\bar c^1,\dots,\bar c^m)\right]^2\mathbb{E}\left[\prod_{i=1}^{j_r}\partial_{s^{-1}(i)}\bar c^{m^r_i}\right]^2}. \quad (136)$$

Since $j_1,\dots,j_p\ge 1$ and all the derivatives of ϕ are bounded, it suffices to bound the second expectations. For this, it suffices to bound the following for any even p', any collecition of weights P', and any $m'\in[M_0+1:m-1]$:

$$\mathbb{E}\left|\partial_{P'}\bar c^{m'}\right|^{p'} = \mathbb{E}\left|\frac{1}{n}\sum_\beta\partial_P\phi\left(\bar g^1_\beta,\dots,\bar g^{m-1}_\beta,\bar c^1,\dots,\bar c^{m-1}\right)\right|^{p'} \le$$
$$\le\sup_\beta\mathbb{E}\left|\partial_P\phi\left(\bar g^1_\beta,\dots,\bar g^{m-1}_\beta,\bar c^1,\dots,\bar c^{m-1}\right)\right|^{p'}, \quad (137)$$

and

$$\mathbb{E}\left|\partial_{P'}\bar{g}_\alpha^{m'}\right|^{p'} =$$

$$= \mathbb{E}\Bigg[\sum_{(\mu,\beta,\gamma)\in P',T_{m'}} 1_{m'\in\mu}1_{\alpha=\beta}\partial_{P'\setminus\{(\mu,\beta,\gamma)^{T_{m'}}\}}\phi\left(\bar{g}_\gamma^1,\ldots,\bar{g}_\gamma^{m'-1},\bar{c}^1,\ldots,\bar{c}^{m'-1}\right) +$$

$$+\sum_\beta \bar{A}_{\alpha\beta}^{m',T_{m'}}\partial_P\phi\left(\bar{g}_\beta^1,\ldots,\bar{g}_\beta^{m'-1},\bar{c}^1,\ldots,\bar{c}^{m'-1}\right)\Bigg]^{p'} \leq$$

$$\leq 2^{p'}\sum_{(\beta,\gamma)\in P',T_{m'}} 1_{m'\in\mu}1_{\alpha=\beta}\mathbb{E}\left[\partial_{P'\setminus\{(\beta,\gamma)^{T_{m'}}\}}\phi\left(\bar{g}_\gamma^1,\ldots,\bar{g}_\gamma^{m'-1},\bar{c}^1,\ldots,\bar{c}^{m'-1}\right)\right]^{p'} +$$

$$+2^{p'}\times\mathbb{E}\left[\sum_\beta \bar{A}_{\alpha\beta}^{m',T_{m'}}\partial_{P'}\phi\left(\bar{g}_\beta^{1:m'-1},\bar{c}^1,\ldots,\bar{c}^{m'-1}\right)\right]^{p'}. \quad (138)$$

$\mathbb{E}\left|\partial_{P'}\bar{c}^{m'}\right|^{p'}$ is bounded by induction. As for $\mathbb{E}\left|\partial_{P'}\bar{g}_\alpha^{m'}\right|^{p'}$, the first expectation is bounded by $B_{p',|P'|}n^{-p'|P'|/2}$ by $2(m'-1)$ while the second one is $B_{p',|P'|}n^{-p'|P'|/2}$ by $2(m')$. This gives $\mathbb{E}\left[\prod_{i=1}^{j_r}\partial_{s^{-1}(i)}\bar{g}_\alpha^{m_i^r}\right]^2 \leq B_{p,|P|}n^{-j_r|P|}$ and $\mathbb{E}\left|\partial_P\phi(\bar{g}_\alpha^1,\ldots,\bar{g}_\alpha^m,\bar{c}^1,\ldots,\bar{c}^m)\right|^p \leq B_{p,|P|}n^{-p|P|/2}$ for appropriately defined $B_{p,|P|}$, as required. As before, uniformity over α comes from the induction hypothesis. The case of odd p follows from Cauchy-Schwarz. □

Now we generalize the above result to polynomially smooth nonlinearities:

Lemma 15. *Consider a Tensor Program $\bar{\Pi}$ of width n with matrices satisfying $\mathbb{E}\,\bar{A}_{\beta\gamma} = 0$ and $\mathbb{E}\left|\bar{A}_{\beta\gamma}\right|^k \leq \nu_k n^{-k/2}$, $\nu_k \in \mathbb{R}_+\ \forall k \geq 1$, and polynomially smooth nonlinearity ϕ. Denote the corresponding vectors as $\bar{g}^1,\ldots,\bar{g}^M$ and the corresponding scalars as $\bar{c}^1,\ldots,\bar{c}^M$.*

Let P be a collection of weights in $\bar{\Pi}$. Then $\forall m \in [M_0 : M-1]\ \forall p \geq 1$

$$\sup_{\alpha\in[n]}\mathbb{E}\left|\partial_P\phi(\bar{g}_\alpha^1,\ldots,\bar{g}_\alpha^m,\bar{c}^1,\ldots,\bar{c}^m)\right|^p \leq B_{p,|P|}n^{-p|P|/2}, \quad (139)$$

where $B_{p,|P|}$ is a continuous function of a finite number of (scaled) moments ν. Similarly, for a polynomially smooth ψ, we have $\forall p \geq 1$

$$\sup_{\alpha\in[n]}\mathbb{E}\left|\partial_P\psi(\bar{g}_\alpha^1,\ldots,\bar{g}_\alpha^M,\bar{c}^1,\ldots,\bar{c}^M)\right|^p \leq B_{p,|P|}n^{-p|P|/2}. \quad (140)$$

Proof. Observe that there were only three places where we have used boundedness of derivatives of ϕ: they are Eqs. (130), (135) and (136). Let us elaborate the first equation when ϕ is bounded with a product of two polynomials

$$\mathcal{P}(x_1,\ldots,x_m) = \sum_{\substack{i_1,\ldots,i_m\geq 0:\\ 0\leq i_1+\ldots+i_m\leq d}} b_{i_{1:m}}x_1^{i_1}\ldots x_m^{i_m}; \quad (141)$$

and

$$\tilde{\mathcal{P}}(x_1,\ldots,x_m) = \sum_{\substack{j_1,\ldots,j_m\geq 0:\\ 0\leq j_1+\ldots+j_m\leq d}} \tilde{b}_{j_{1:m}}x_1^{j_1}\ldots x_m^{j_m}: \quad (142)$$

$$\mathbb{E}\,|\phi(\bar{g}^1_\alpha,\dots,\bar{g}^m_\alpha,\bar{c}^1,\dots,\bar{c}^m)|^p \le \mathbb{E}\,|\mathcal{P}(\bar{g}^1_\alpha,\dots,\bar{g}^m_\alpha)\tilde{\mathcal{P}}(\bar{c}^1,\dots,\bar{c}^m)|^p \le$$

$$\le d^{(2m-M_0)p}\left[\sum_{\substack{i_1,\dots,i_m\ge 0:\\ 0\le i_1+\dots+i_m\le d}} |b_{i_{1:m}}|^p\,\mathbb{E}\left|\left(\bar{g}^1_\alpha\right)^{i_1}\dots\left(\bar{g}^m_\alpha\right)^{i_m}\right|^p\right]\times$$

$$\times\left[\sum_{\substack{j_1,\dots,j_m\ge 0:\\ 0\le j_1+\dots+j_m\le d}} \left|\tilde{b}_{j_{1:m}}\right|^p\,\mathbb{E}\left|\left(\bar{c}^1\right)^{j_1}\dots\left(\bar{c}^m\right)^{j_m}\right|^p\right]. \quad (143)$$

Due to the Cauchy-Schwarz inequality, it suffices to bound $\mathbb{E}\left|g^{m'}_\alpha\right|^{p'}$ for $m'\in[m]$ and $\mathbb{E}\left|c^{m'}\right|^{p'}$ for $m'\in[M_0+1:m]$, for arbitrary $p'\ge 1$. The first one is given by Claim 2 for $|P|=0$ and $\phi(x)=x$ (it is smooth with all the derivatives bounded): $\mathbb{E}\left|g^{m'}_\alpha\right|^{p'}=O(1)$. The second one is given again by our Claim 2 for $|P|=0$:

$$\mathbb{E}\left|c^{m'}\right|^{p'} \le \frac{1}{n^{p'}}\sum_{\beta_1,\dots,\beta_{p'}}\mathbb{E}\left|\prod_{k=1}^{p'}\phi(g^{m'}_{\beta_k})\right| = O(1) \quad (144)$$

by Cauchy-Schwarz. Hence $\mathbb{E}\,|\phi(\bar{g}^1_\alpha,\dots,\bar{g}^m_\alpha)|^p=O(1)$ as before.

The two latter cases, Eqs. (135) and (136), relied on the fact that $\mathbb{E}\left[\prod_{r=1}^p \partial_{m^r_1}\dots\partial_{m^r_j}\phi(\bar{g}^1_\alpha,\dots,\bar{g}^m_\alpha,\bar{c}^1,\dots,\bar{c}^m)\right]^2$ is bounded when ϕ has bounded derivatives. Again, using Cauchy-Schwarz, it suffices to bound $\mathbb{E}\left|\partial_{m^r_1}\dots\partial_{m^r_j}\phi(\bar{g}^1_\alpha,\dots,\bar{g}^m_\alpha,\bar{c}^1,\dots,\bar{c}^m)\right|^{p'}$ for arbitrary p'. But it follows from the previous speculation since $\partial_{m^r_1}\dots\partial_{m^r_j}\phi$ is smooth and has all the derivatives polynomially bounded. □

H.4 Proof of Lemma 9

$$\mathbb{E}\left|\partial_P\left[\prod_{i=1}^u f_i(\bar{g}^1_{\alpha_i},\dots,\bar{g}^{m_i}_{\alpha_i},\bar{c}^1,\dots,\bar{c}^{m_i})\right]\right| =$$

$$= \mathbb{E}\left|\sum_{\sqcup_{i=1}^u P_i=P}\prod_{i=1}^u \partial_{P_i} f_i(\bar{g}^1_{\alpha_i},\dots,\bar{g}^{m_i}_{\alpha_i},\bar{c}^1,\dots,\bar{c}^{m_i})\right| \le$$

$$\le \sum_{\sqcup_{i=1}^u P_i=P}\mathbb{E}\left|\prod_{i=1}^u \partial_{P_i} f_i(\bar{g}^1_{\alpha_i},\dots,\bar{g}^{m_i}_{\alpha_i},\bar{c}^1,\dots,\bar{c}^{m_i})\right| \le$$

$$\le \sum_{\sqcup_{i=1}^u P_i=P}\prod_{i=1}^u \sqrt[u]{\mathbb{E}\left[\left|\partial_{P_i} f_i(\bar{g}^1_{\alpha_i},\dots,\bar{g}^{m_i}_{\alpha_i},\bar{c}^1,\dots,\bar{c}^{m_i})\right|^u\right]} \le$$

$$\le \sum_{\sqcup_{i=1}^u P_i=P}\prod_{i=1}^u\left(\sqrt[u]{B_{u,|P_i|}}\,n^{-|P_i|/2}\right) = n^{-|P|/2}\sum_{\sqcup_{i=1}^u P_i=P}\prod_{i=1}^u \sqrt[u]{B_{u,|P_i|}} \quad (145)$$

by Lemma 10 and Lemma 15. Defining $\bar{B}_{u,|P|}=\sum_{\sqcup_{i=1}^u P_i=P}\prod_{i=1}^u\sqrt[u]{B_{u,|P_i|}}$ gives the result.

I Empirical validation

In the present section, we validate our main result, Theorem 2, by empirically checking some of its corollaries mentioned in Section 3.

I.1 On non-Gaussian biases and input and output layers

Here we first note a very important subtlety concerning distribution universality. When expressing neural network computations (i.e. forward and backward passes) as a tensor program, we use matrices for initial hidden weights and vectors for biases and initial input and output weights. Indeed, since for input and output layers, only one dimension grows to infinity, they cannot be expressed as $n \times n$ matrices in the program for which both dimensions go to infinity. Instead, they are expressed as a set of vectors.

As noted in Appendix A, we model non-Gaussian vector variables as images of Gaussian ones under elementwise nonlinear maps. This requires adding a nonlinearity to a program, which may alter the limit in Theorem 2.

This means that the NNGP and NTK kernels of Corollaries 1 and 2 are not necessarily distribution-invariant. However both corollaries are still true as they merely state that these kernels exist.

I.2 Setup

We perform our experiments with vanilla RNN, vanilla GRU network, and a simple Transformer. We build on the code acompanying [22] and [23].

As for experiments aimed to validate NNGP correspondence, Corollary 1, we compute correlation matrices on two sentences, "The brown fox jumps over the dog" and "The quick brown fox jumps over the lazy dog", embedded using GloVe embeddings [16]. We compute the empirical kernel for different distributions and measure the relative Frobenius norm distance between this empirical kernel and the analytic kernel for Gaussian initialization. We compute mean and standard deviations for this relative Frobenius norm distance using 100 random initializations.

As for experiments aimed to validate convergence to a kernel method, Corollary 2, we compute empirical kernels on two "sentences", the first 5 pixels of the first CIFAR10 image and the first 5 pixels of the second CIFAR10 image. We compute it for different distributions and measure the relative Frobenius norm distance between this empirical kernel and the analytic neural tangent kernel. Same as for NNGP correspondence, we compute mean and standard deviations for this relative Frobenius norm distance using 100 random initializations.

Following Setup 2, we consider distributions with zero mean and variance $1/n$. The distributions we consider are Gaussian, Gaussian truncated at 2σ, uniform, Laplace, and the other one which we call *Cubecauchy*. The Cubecauchy distribution is a distribution of a cubic root of a Cauchy random variable. Such a random variable has finite mean and variance, but does not have any other moments. Since it does not have all moments, it does not follow Setup 2, which makes our Theorem 2 inapplicable. However, as we shall see shortly, the empirical kernels we consider still converge to the same limit suggesting that existence of all moments is not necessary.

I.3 Results

We start with empirically validating NNGP correspondence, Corollary 1. We first swap only hidden weights with their non-Gaussian counterparts. In this case, our Theorem 2 predicts that the limit should not depend on matrix entry distributions. As we see in Fig. 2, this is indeed the case. Moreover, the empirical kernel that corresponds to the Cubecauchy distribution still converges to the same limit as for the other distributions, thus suggesting that coincidence of only the first two moments are necessary for the Master theorem to hold.

What happens if we swap all other trainable parameters, i.e. biases, input weights, embeddings, with their non-Gaussian counterparts? As discussed above, the Master theorem still holds, but it requires modifying the program itself which may alter the limit. As we observe in Fig. 3, sometimes it is indeed the case (left plot), while sometimes the limit could be the same (right plot). We note that even when the limit is different from the one resulted from Gausian weights, Fig. 3 still demonstrates that the limit exists even for non-Gaussian ones, therefore validating Corollary 1.

Next, we validate neural tangent kernel convergence at initialization, checking a part of Corollary 2. As we see in Fig. 4, even when all trainable parameters are initialized with non-Gaussians, the limit kernel still remains the same.

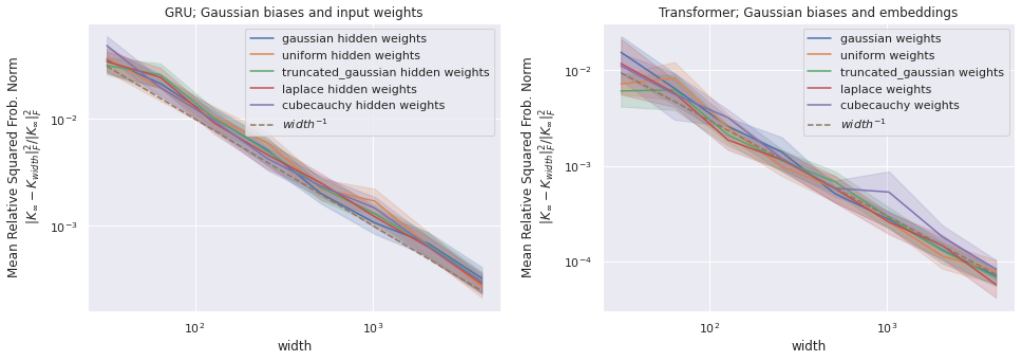

Figure 2: Relative Frobenius norm distance between empirical NNGP kernels for different distributions and the analytic kernel for Gaussian initialization. All trainable parameters expressed with vectors (i.e. biases, embeddings, and input weights) are assumed to be Gaussian. Left: a simple GRU network, right: a simple transformer, see Appendix I.2. As we see, NNGP kernels for all distributions considered converge to the same analytic kernel.

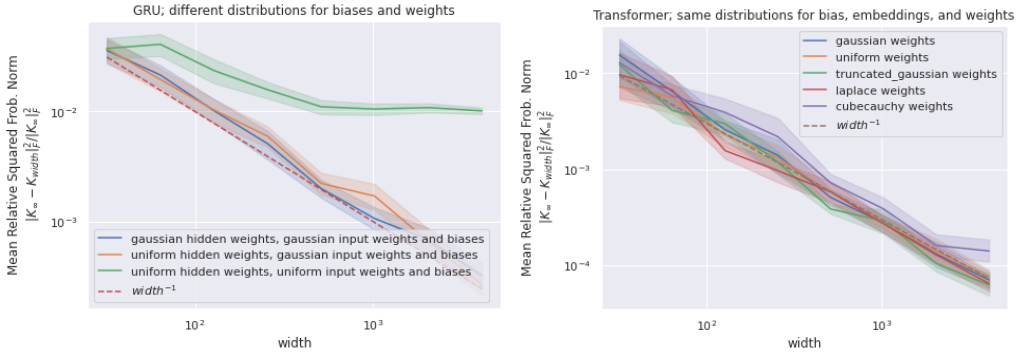

Figure 3: Relative Frobenius norm distance between empirical NNGP kernels for different distributions and the analytic kernel for Gaussian initialization. We assume all trainable parameters to have the same distribution. Left: a simple GRU network, right: a simple transformer, see Appendix I.2. As we see, in some cases (left, green line), the limit NNGP kernel may depend on the parameter distribution since we had to modify the corresponding tensor program in order to model non-Gaussian vector variables, see Appendix I.1. Note that the limit kernel still exists and there is no contradiction with Corollary 1.

Figure 4: Relative Frobenius norm distance between empirical neural tangent kernels for different distributions and the analytical neural tangent kernel for Gaussian initialization. We assume all trainable parameters to have the same distribution. Left: a simple recurrent neural network, right: a simple transformer, see Appendix I.2. As we see, the limit NTK does not depend on the parameter distribution, even when we use non-Gaussian vectors, see discussion in Appendix I.1.