A Proof of Theorem 5.2 for $p = 2$ in a simple scenario

Setup To make our arguments clear, we illustrate the proof of Theorem 5.2 on the simplest non-trivial case, with only a single matrix $A = A^1$, a single initial vector $g^1 \sim \mathcal{N}(0, I)$, and with c^1, c^2, and g^3 ignored. The objects of the interpolated program (Definition 5.1) are as follows:

$$g^2(t) = A(t)\phi(g^1(t)), \quad c(t) = \frac{1}{n}\sum_{\alpha=1}^{n} x_\alpha(t), \quad \text{where } x_\alpha(t) = \psi(g^1_\alpha(t), g^2_\alpha(t))$$

where, for brevity, we have shorthanded $c = c^3$, $x = x^3$, $\phi = \phi^1$, and $\psi = \phi^3$.[23]

For simplicity, we will assume ϕ and ψ have derivatives of all orders bounded by 1 in absolute value. Our goal in this section is to demonstrate Theorem 5.2 with $p = 2$:

$$\sup_t \mathbb{E}\,\dot{c}(t)^2 = O(n^{-1}). \tag{8}$$

In what follows, we will only talk about the interpolated program, so we will suppress the argument (t) to lighten notation.

A.1 Bounding Derivatives of c against A

We will denote $x'_\alpha \overset{\text{def}}{=} \partial_{g^2_\alpha}\psi(g^1_\alpha, g^2_\alpha)$, $x''_\alpha \overset{\text{def}}{=} \partial^2_{g^2_\alpha}\psi(g^1_\alpha, g^2_\alpha)$, and so on, which shall cause no confusion because we will never take other derivatives of x or ψ. Then we can calculate

$$\frac{\partial c}{\partial A_{\alpha\beta}} = \frac{1}{n}x'_\alpha\phi(g^1_\beta), \quad \frac{\partial^2 c}{(\partial A_{\alpha\beta})^2} = \frac{1}{n}x''_\alpha\phi(g^1_\beta)^2, \quad \text{etc.} \tag{9}$$

By the assumption that ψ and ϕ have derivatives of any order bounded by 1, this shows c's derivative against $A_{\alpha\beta}$ of any order ≥ 1 is bounded in absolute value by $1/n$. Generalizing the above calculation to mixed derivatives, one can easily see more generally,

$$|\partial^r c| \leq 1/n \quad \text{for any order } r \geq 1 \text{ mixed derivative } \partial^r \text{ in entries of } A. \tag{10}$$

If we allow ψ and ϕ to be polynomially smooth in general, then this is still true in expectation, up to multiplicative constants; see Lemma K.6.

A.2 Expanding $\dot{c}^2$

At this point, it helps to think of the entries of A as a big vector a of size $N = n^2$. We will index entries of a using letters like κ, which stands for a pair $(\alpha, \beta) \in [n]^2$. We also let $D \in \mathbb{R}^N$ be the vector of derivatives $D_\kappa \overset{\text{def}}{=} \partial_{a_\kappa} c$ (with exact values given by Eq. (9)), so that, by chain rule,

$$\dot{c} = \sum_\kappa D_\kappa \dot{a}_\kappa.$$

Squaring this sum, we get

$$\dot{c}^2 = \sum_{\kappa,\lambda} D_\kappa D_\lambda \dot{a}_\kappa \dot{a}_\lambda = \sum_\kappa D^2_\kappa \dot{a}^2_\kappa + \sum_{\kappa\neq\lambda} D_\kappa D_\lambda \dot{a}_\kappa \dot{a}_\lambda$$

To show Eq. (8), it suffices to show that both sums $\sum_\kappa$ and $\sum_{\kappa\neq\lambda}$ in expectation can be bounded by $O(n^{-1})$ with a constant independent of $t \in [0, 1]$.

A.3 Bounding $\sum_\kappa D^2_\kappa \dot{a}^2_\kappa$.

Because κ ranges over a set of size $N = n^2$, we have

$$\mathbb{E}\sum_\kappa D^2_\kappa \dot{a}^2_\kappa \leq n^2 \max_\kappa \mathbb{E}\, D^2_\kappa \dot{a}^2_\kappa.$$

So it suffices to show $\mathbb{E}\, D^2_\kappa \dot{a}^2_\kappa$ is bounded by Cn^{-3} where C is a constant independent of κ and $t \in [0, 1]$. But by Cauchy-Schwarz,

$$\mathbb{E}\, D^2_\kappa \dot{a}^2_\kappa \leq \sqrt{\mathbb{E}\, D^4_\kappa} \cdot \sqrt{\mathbb{E}\, \dot{a}^4_\kappa} = n^{-2} \cdot O(n^{-1}) = O(n^{-3})$$

where we used Eq. (10) and Lemma 5.3. The hidden constant is independent of κ and t, as desired.

[23] So we would have $M_0 = 1$, $M = 3$, but these values are not important for our purposes.

A.4 Bounding $\sum_{\kappa \neq \lambda} D_\kappa D_\lambda \dot{a}_\kappa \dot{a}_\lambda$

Because (κ, λ) ranges over a set of size $\leq N^2 = n^4$, we have

$$\mathbb{E} \sum_{\kappa \neq \lambda} D_\kappa D_\lambda \dot{a}_\kappa \dot{a}_\lambda \leq n^4 \max_{\kappa \neq \lambda} \mathbb{E}\, D_\kappa D_\lambda \dot{a}_\kappa \dot{a}_\lambda.$$

So it suffices to show $\mathbb{E}\, D_\kappa D_\lambda \dot{a}_\kappa \dot{a}_\lambda$ is bounded by Cn^{-5} where C is a constant independent of κ, λ (as long as $\kappa \neq \lambda$) and $t \in [0, 1]$.

Taylor Expansion Now $D_\kappa D_\lambda$ is a function of a. In particular, we will think of it as a function of a_κ and a_λ, keeping other entries of a fixed. To this end, we will write $D_\kappa D_\lambda = f(a_\kappa, a_\lambda)$. To reduce the amount of subscripts, denote $y = a_\kappa, z = a_\lambda$. Then Taylor expanding f to some order K to be specified below, we have

$$f(y, z) = \sum_{i+j \leq K} y^i z^j \Delta_{ij} + \sum_{i+j=K+1} y^i z^j R_{ij}(y, z)$$

where $\Delta_{ij} = \frac{1}{(i+j)!} \binom{i+j}{i} \partial_y^i \partial_z^j f(0, 0)$ and $R_{ij}(y, z) = \frac{1}{K!} \binom{K+1}{i} \int_0^1 (1-\xi)^K \partial_y^i \partial_z^j f(\xi y, \xi z) d\xi$.

This may look like a big scary expression, but as we will see soon enough, the exact form of Δ_{ij} does not matter beyond the fact that it is independent from y and z, and we can easily bound R_{ij} using Eq. (10).

Cancellation Using Independence and Zero-Mean Since Δ_{ij} is independent from y and z (as random variables), we have

$$\mathbb{E}\, \dot{y}\dot{z} y^i z^j \Delta_{ij} = (\mathbb{E}\, \dot{y} y^i)(\mathbb{E}\, \dot{z} z^j) \mathbb{E}\, \Delta_{ij}. \tag{11}$$

We see immediately that if $i < 2$ or $j < 2$ then this expectation will vanish due to Lemma 5.3. This motivates us to take K (the order of Taylor expansion) to be 3, so that

$$\mathbb{E}\, \dot{y}\dot{z} f(y, z) = \mathbb{E}\, \dot{y} y^2 \dot{z} z^2 R_{ij}(y, z)$$

Now $R_{ij}(y, z)$ is not independent from y and z unlike Δ_{ij}, but by Cauchy-Schwarz, we have

$$\mathbb{E}\, \dot{y} y^2 \dot{z} z^2 R_{ij}(y, z) \leq \sqrt{\mathbb{E}\, (\dot{y} y^2 \dot{z} z^2)^2} \sqrt{\mathbb{E}\, R_{ij}(y, z)^2}$$

By Eq. (10) and the product rule, we see all order-$(K+1)$ (mixed) derivatives of f are bounded by $(K+2)/n^2 = 5/n^2$, so that $R_{ij}(y, z) \leq C/n^2$ for some absolute constant C. This implies $\sqrt{\mathbb{E}\, R_{ij}(y, z)^2} \leq C/n^2$ as well.

Finally, by Lemma 5.3, $\mathbb{E}\, (\dot{y} y^2 \dot{z} z^2)^2 = \mathbb{E}\, (\dot{y} y^2)^2 \mathbb{E}\, (\dot{z} z^2)^2 = O(n^{-6})$ independent of κ and λ. So altogether,

$$\mathbb{E}\, \dot{y} y^2 \dot{z} z^2 R_{ij}(y, z) \leq O(n^{-3}) O(n^{-2}) = O(n^{-5})$$

with a constant independent of t, κ, and λ. This finishes the proof of Theorem 5.2 with $p = 2$ in our simple scenario.

From this example, one can see the importance of $\mathbb{E}\, a\dot{a} = 0$ in Lemma 5.3: had it not been the case, then Eq. (11) could be nonvanishing for $i = 2, j = 0$, so we could only Taylor expand f to order 1 instead of 3, losing a factor of n in the final bound as a result.

A.5 Going Beyond This Simple Example

To get the full result of Theorem 5.2, we need to extend this argument to 1) all powers p instead of just 2 (i.e., bounding $\mathbb{E}\, |\dot{c}|^p$), and to 2) all polynomially smooth nonlinearities.

For 1), the extension follows more or less the line of reasoning demonstrated here, which can be extracted as a general *moment argument* which we explain in Appendix I.

For 2), significant effort is needed to establish the property corresponding to Eq. (10). In fact, this is the bulk of the technical content of our work, done in Appendix J and Appendix K.1, culminating in Lemma K.6 that generalizes Eq. (10).

B Gaussian smoothing makes polynomially bounded nonlinearities polynomially smooth

Lemma B.1. *Let $\phi : \mathbb{R}^k \to \mathbb{R}$. Then for any $\delta > 0$ there exists $\phi^\delta \in C^\infty(\mathbb{R}^k)$ such that*

1. *For any $x \in \mathbb{R}^k$,*
$$|\phi(x) - \phi^\delta(x)| \le \sup_{x_1, x_2 \in \bar{B}_\delta(x)} |\phi(x_1) - \phi(x_2)|,$$
where $\bar{B}_\delta(x)$ is a closed δ-vicinity of x;
2. *If ϕ is polynomially bounded then ϕ^δ is polynomially smooth;*
3. *If ϕ is Lipschitz with a Lipschitz constant λ, then ϕ^δ is also Lipschitz with a Lipschitz constant λ and*
$$\sup_{x \in \mathbb{R}^k} |\phi(x) - \phi^\delta(x)| \le 2\lambda\delta.$$

Proof. Let ζ be a standard bump function with k arguments:
$$\zeta(z) = Ce^{-\frac{1}{1-\|z\|_2^2}} \mathbb{I}(\|z\|_2 \le 1). \tag{12}$$
It is a non-negative $C^\infty(\mathbb{R}^k)$ function that is supported supported on a unit ball. We choose the constant C so that ζ sums to one.

Claim 1 Let $\zeta^\delta(x) \stackrel{\text{def}}{=} (1/\delta)^k \zeta(x/\delta)$. Set $\phi^\delta(x) \stackrel{\text{def}}{=} [\phi * \zeta^\delta](x)$. Then
$$\begin{aligned}
\phi^\delta(x) &= \int_{\mathbb{R}^k} \zeta^\delta(y)\phi(x-y)\,dy \\
&= (1/\delta)^k \int_{\mathbb{R}^k} \zeta(y/\delta)\phi(x-y)\,dy \\
&= \int_{\mathbb{R}^k} \zeta(z)\phi(x - z\delta)\,dz.
\end{aligned}$$
We have for any $x \in \mathbb{R}^k$,
$$\begin{aligned}
|\phi^\delta(x) - \phi(x)| &= \left| \int_{\mathbb{R}^k} \zeta(z)\left(\phi(x - z\delta) - \phi(x)\right)\,dz \right| \\
&\le \left| \int_{\|z\|_2 \le 1} \zeta(z)\,dz \right| \sup_{\|z\|_2 \le 1} |\phi(x - z\delta) - \phi(x)| \\
&\le \sup_{x_1, x_2 \in \bar{B}_\delta(x)} |\phi(x_1) - \phi(x_2)|.
\end{aligned}$$

Claim 2 Suppose ϕ is polynomially bounded, i.e. $|\phi(x)| \le C(1 + \|x\|_p^p)$ for some $p \ge 1$. For any $r \ge 0$ and any $j_1, \ldots, j_r \in [k]$,
$$\begin{aligned}
\partial_{j_1,\ldots,j_r}\phi^\delta(x) &= (1/\delta)^{k+\sum_{i=1}^r j_i} \int_{\mathbb{R}^k} \partial_{j_1,\ldots,j_r}\zeta((x-y)/\delta)\phi(y)\,dy \\
&= (1/\delta)^{k+\sum_{i=1}^r j_i} \int_{\|y\|_2 \le \delta} \partial_{j_1,\ldots,j_r}\zeta(y/\delta)\phi(x-y)\,dy \\
&= (1/\delta)^k \int_{\|z\|_2 \le 1} \partial_{j_1,\ldots,j_r}\zeta(z)\phi(x - z\delta)\,dz.
\end{aligned}$$
Therefore,
$$\begin{aligned}
\left|\partial_{j_1,\ldots,j_r}\phi^\delta(x)\right| &\le (1/\delta)^k \sup_{\|z\|_2 \le 1} |\phi(x - z\delta)| \left| \int_{\|z\|_2 \le 1} \partial_{j_1,\ldots,j_r}\zeta(z)\,dz \right| \\
&\le C(1/\delta)^k \sup_{\|z\|_2 \le 1} (1 + \|x - z\delta\|_p^p) \le C(1/\delta)^k \sup_{\|z\|_2 \le 1} (1 + 2^p\|x\|_p^p + 2^p\delta^p\|z\|_p^p) \\
&\le C(1/\delta)^k \sup_{\|z\|_2 \le 1} (1 + 2^p\|x\|_p^p + 2^p\delta^p p^p\|z\|_\infty^p) \le C(1/\delta)^k(1 + 2^p\|x\|_p^p + (2p\delta)^p),
\end{aligned}$$

which implies that $\partial_{j_1,\ldots,j_r}\phi^\delta$ is polynomially bounded.

Claim 3 Suppose ϕ is Lipschitz with a Lipschitz constant λ. Let us check that ϕ^δ is also Lipschitz with a Lipschitz constant λ:

$$|\phi^\delta(x_1)-\phi^\delta(x_2)| = \left|\int_{\mathbb{R}^k}\zeta(z)\left(\phi(x_1-z\delta)-\phi(x_2-z\delta)\right)\,dz\right|$$
$$\le \lambda|x_1-x_2|\int_{\mathbb{R}^k}\zeta(z)\,dz = \lambda|x_1-x_2|.$$

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

$$|\phi(x)-\phi^\delta(x)| \le \sup_{x_1,x_2\in\bar{B}_\delta(x)}|\phi(x_1)-\phi(x_2)| \le \lambda\sup_{x_1,x_2\in\bar{B}_\delta(x)}\|x_1-x_2\|_2 \le 2\lambda\delta. \tag{13}$$

□

C Additional applications of Theorem 3.7

TP3 demonstrated the following applications of Theorem 3.4: 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 generalize all the above results to non-Gaussian weight initializations.

C.1 Semicircle law for non-Gaussian Wigner ensembles

The semicircle law for GOE proven in [31] is the following result:

Theorem C.1. *For each $n\ge 1$, 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

$$\frac{1}{n}\sum_{\alpha=1}^n\phi(\lambda_\alpha)\xrightarrow{\text{a.s.}}\mathbb{E}_{\lambda\sim\mu_{sc}}\phi(\lambda). \tag{14}$$

In order to be able to apply Tensor Programs machinery to prove this result, [31] used the moment method which states that it suffices to prove almost sure convergene of moments of μ_n:

$$\mathbb{E}_{\lambda\sim\mu_n}\lambda^r=\frac{1}{n}\sum_{\alpha=1}^n\lambda_\alpha^r=n^{-1}\operatorname{tr}(A^r)\xrightarrow{\text{a.s.}}\mathbb{E}_{\lambda\sim\mu_{sc}}\lambda^r\quad\forall r\in\mathbb{N}. \tag{15}$$

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^rz can be expressed as a vector in the following Tensor Program:

$$g^1=z,\quad g^2=Wz,\quad g^3=W^\top z, \tag{16}$$
$$g^{2k}=W(g^{2k-1}+g^{2k-2}),\quad g^{2k+1}=W^\top(g^{2k-1}+g^{2k-2})\quad\forall k\in[2:r]. \tag{17}$$

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

$$n^{-1}\operatorname{tr}(A^r)=\frac{1}{n}\mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^\top A^r z]=$$
$$=\frac{1}{n}\sum_{\alpha=1}^n\mathbb{E}_{g_\alpha^1\sim\mathcal{N}(0,1)}\left[g_\alpha^1\left(g_\alpha^{2r}+g_\alpha^{2r+1}\right)\right]=\frac{1}{n}\sum_{\alpha=1}^n\mathbb{E}_{g_\alpha^1\sim\mathcal{N}(0,1)}\psi(g_\alpha^1,\ldots,g_\alpha^{2r+1}), \tag{18}$$

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 3.4 to the expression above is conditional expectation. [31] proves the following conditional theorem:

Theorem C.2 (Gaussian Conditional Master Theorem, [31])**.** *Consider Setup 3.3 and assume ψ to be quadratically bounded and all the nonlinearities to be linearly bounded. Let S be a subset of initial vectors. Then, as $n \to \infty$,*

$$\frac{1}{n}\sum_{\alpha=1}^{n} \mathbb{E}_{\,S}\psi\left(g_\alpha^1, \dots, g_\alpha^M, c^1, \dots, c^M\right) \xrightarrow{\text{a.s.}} \mathring{\Psi}, \tag{19}$$

where $\mathring{\Psi}$ is the same as in Theorem 3.2.

We had to come back to the original form of Master theorems (see Theorem 3.2) since there is a distinction between nonlinearities (which have to be linearly bounded) and test functions ψ, which could be quadratically bounded.

The above conditional theorem implies $n^{-1}\operatorname{tr}(A^r) \xrightarrow{\text{a.s.}} \mathring{\Psi}$ for certain $\mathring{\Psi}$. Comparing with Eq. (15), what remains to establish Theorem C.1, is to prove that $\mathring{\Psi}$ equals to the r-th moment of the semicircle law. One can find the proof in [31]; we do not reproduce it here.

Non-Gaussian conditional Master theorem. Consider Setup 3.6 and let S be a subset of initial vectors. Then for any scalar c^i in the program, $\mathbb{E}\,|\mathbb{E}_{\,S}c^i - \mathring{c}^i|^p \le \mathbb{E}\,|c^i - \mathring{c}^i|^p$, which converges to zero by Theorem 3.7. Therefore $\mathbb{E}_{\,S}c^i$ converges to $\mathring{c}^i$ in L^p.

Consider now the same interpolation between Gaussian and non-Gaussian weights $A^l(t)$ as in the proof of Theorem 3.7, see Section 5. Theorem 5.2 then gives $\sup_t \mathbb{E}\,|\mathbb{E}_{\,S}c^i(t)|^p \le \sup_t \mathbb{E}\,|c^i(t)|^p = O(n^{-p/2})$ for any $p \in [1,\infty)$. Following the argument in Section 5, we get $\mathbb{E}_{\,S}c^i(1) - \mathbb{E}_{\,S}c^i(0) \xrightarrow{\text{a.s.}} 0$. Assuming the conditions of the above Gaussian theorem, Theorem C.2, namely, that ϕ^i is quadratically bounded, while ϕ^j for all $j < i$ are linearly bounded, we get $\mathbb{E}_{\,S}c^i(0) \xrightarrow{\text{a.s.}} \mathring{c}^i$, and therefore $\mathbb{E}_{\,S}c^i(1) \xrightarrow{\text{a.s.}} \mathring{c}^i$:

Theorem C.3 (Non-Gaussian Conditional Master Theorem, ours)**.** *Consider Setup 3.6 and let S be a subset of initial vectors. Then every scalar c^i conditioned on S converges to the same $\mathring{c}^i$ as in Theorem 3.4 in L^p for any $p \in [1,\infty)$:*

$$\mathbb{E}_{\,S}c^i \xrightarrow{L^p} \mathring{c}^i \quad \forall p \in [1,\infty). \tag{20}$$

If moreover all the nonlinearities are linearly bounded then for any quadratically bounded polynomially smooth ψ,

$$\frac{1}{n}\sum_{\alpha=1}^{n} \mathbb{E}_{\,S}\psi\left(g_\alpha^1, \dots, g_\alpha^M, c^1, \dots, c^M\right) \xrightarrow{\text{a.s.}} \mathring{\Psi} \tag{21}$$

as $n \to \infty$, where $\mathring{\Psi}$ is the same as in Theorem 3.2.

Since the program used to compute the moments conforms not only Setup 3.3 but also a stronger Setup 3.6, we get a full analogue of Theorem C.1:

Theorem C.4. *For each $n \ge 1$, define the random matrix $A = W + W^\top$ for W being an $n \times n$ matrix with iid entries with zero mean, variance $1/(2n)$, and with all higher moments existing. Let $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ be the eigenvalues of A and $\mu_n = \frac{1}{n}\sum_{\alpha=1}^{n}\delta_{\lambda_\alpha}$ be their empirical distribution. Then $\mu_n \xrightarrow{\text{a.s.}} \mu_{sc}$, where μ_{sc} is the distribution with density $\propto \sqrt{4-x^2}$.*

Relaxing boundedness. Could we relax the assumption on linear and quadratic boundedness in our Theorem C.3? Inspired by Central Limit Theorem and our matrix derivative bound below, Lemma K.6, we conjecture the following moment bound that quantifies the rate of L^p convergence:

Conjecture C.5. *Under Setup C.6 below, $\mathbb{E}\,|c^i - \mathring{c}^i|^p = O(n^{-p/2})$ for any scalar c^i in the program and the corresponding almost sure limit $\mathring{c}^i$, and any $p \in [1,\infty)$.*

Setup C.6. *Consider Setup 3.6 but replace 5) with 5*) for any initial scalar c^i and any $p \in [1,\infty)$, $\mathbb{E}\,|c^i - \mathring{c}^i|^p = O(n^{-p/2})$, where $\mathring{c}^i$ is the almost sure limit of c^i which exists due to 2).*

If the above conjecture holds, we will get a similar bound for scalars conditioned on S, i.e. $\mathbb{E}\,|\mathbb{E}_{\,S}c^i - \mathring{c}^i|^p = O(n^{-p/2})$, which will imply $\mathbb{E}_{\,S}c^i \xrightarrow{\text{a.s.}} \mathring{c}^i$ by Borel-Cantelli lemma:

Conjecture C.7. *Consider Setup C.6 and let S be a subset of initial vectors. Then every scalar c^i conditioned on S converges to the same $\mathring{c}^i$ as in Theorem 3.4 almost surely and in L^p for any $p \in [1, \infty)$:*

$$\mathbb{E}_S c^i \xrightarrow{\text{a.s. \& } L^p} \mathring{c}^i \quad \forall p \in [1, \infty). \tag{22}$$

We leave proving Conjecture C.5 for future work.

C.2 Marchenko-Pastur law for non-Gaussian Wishart ensembles

Using exactly the same machinery as for the semi-circle law, [31] 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 3.7 (with a remark on programs with variable dimensions, see Section 3) gives a similar result with no assumptions on Gaussianity of A:

Theorem C.8 (Ours)**.** *For each $n \geq 1$, let A be an $m \times n$ matrix with iid entries with zero mean, variance $1/n$, and with all higher moments existing. Let $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ be the eigenvalues of $AA^\top$ and $\mu_n = \frac{1}{n}\sum_{\alpha=1}^n \delta_{\lambda_\alpha}$ be their empirical distribution. Let m go to infinity as $n \to \infty$ such that $m/n \to \rho \in (0, \infty)$. Then $\mu_n \xrightarrow{a.s.} \mu_{mp}$, where μ_{mp} has "density" $p_{mp}(x)$ given below:*

$$p_{mp}(x) = \max(0, 1 - \rho^{-1})\delta(x) + \frac{1}{\rho 2\pi x}\sqrt{(b-x)(x-a)}1_{x\in[a,b]}, \tag{23}$$

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

C.3 Free Indepedence Principle for Tensor Programs with non-Gaussian weights

Using the same machinery again, [31] proves Free Independence Principle for Tensor Programs:

Theorem C.9 ([31])**.** *Consider Setup 3.3 and assume all nonlinearities to be linearly bounded. Let $\mathcal{D}$ denote the collection of diagonal matrices formed from bounded coordinatewise images of vectors in the program:*

$$\mathcal{D} = \left\{\operatorname{diag}\left(\psi\left(g^1, \ldots, g^M\right)\right) : \ \psi : \mathbb{R}^M \to \mathbb{R} \text{ is bounded}\right\}. \tag{24}$$

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 C.10. Fix k. Consider collections of random matrices $\mathcal{W}_n^1, \ldots, \mathcal{W}_n^k \subseteq \mathbb{R}^{n\times n}$ for each $n \geq 1$, of constant cardinalities (with n). We say $\mathcal{W}_n^1, \ldots, \mathcal{W}_n^k$ are almost surely asymptotically freely independent (or just almost surely asymptotically free), if

$$n^{-1}\operatorname{tr}\left(\prod_{i=1}^k \left(P_i(\mathcal{W}_n^{j_i}) - \tau_i I\right)\right) \xrightarrow{\text{a.s.}} 0, \tag{25}$$

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

The proof strategy for Theorem C.9 is very similar to the one of Theorem C.1. The trace in each τ_i can be expressed as an expectation $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[z^\top P_i(\mathcal{W}_n^{j_i})z]$ and $P_i(\mathcal{W}_n^{j_i})z$ can be expressed as a vector in a program with linearly bounded nonlinearities (that's the reason why we require ψ in the definition of $\mathcal{D}$ to be bounded). Therefore each τ_i converges by Theorem C.2, and therefore can be thought as a scalar in a new Tensor Program. The "outer" trace in Eq. (25) can be expressed as conditional expectation using the same trick, and the corresponding Eq. (25) again converges by Theorem C.2. We refer the reader to [31] for details.

Simply replacing Theorem C.2 in the above reasoning with its non-Gaussian analogue, Theorem C.3, we get a non-Gaussian analogue of Theorem C.9:

Theorem C.11 (Ours)**.** *Consider Setup 3.6 and assume all nonlinearities to be linearly bounded. Let $\mathcal{D}$ denote the collection of diagonal matrices formed from bounded polynomially smooth coordinatewise images of vectors in the program:*

$$\mathcal{D} = \left\{ \operatorname{diag}\left(\psi\left(g^1, \dots, g^M\right)\right) : \ \psi : \mathbb{R}^M \to \mathbb{R} \text{ is bounded and polynomially smooth} \right\}. \tag{26}$$

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

We had to require all nonlinearities to be linearly bounded and the functions ψ in the definition of diagonal matrices $\mathcal{D}$ to be bounded in order to apply a conditional theorem, Theorem C.2 or Theorem C.3. Note that our original non-Gaussian Master theorem, Theorem 3.7, gives also convergence in L^1, which implies convergence of (full, non-conditional) expectations without requiring linearly bounded nonlinearities[24]. Convergence of full expectations allows us to prove more traditional asymptotic freeness in expectation in our non-Gaussian case:

Definition C.12. 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

$$n^{-1}\mathbb{E}\left[\operatorname{tr}\left(\prod_{i=1}^{k}\left(P_i(\mathcal{W}_n^{j_i}) - \tau_i I\right)\right)\right] \to 0, \tag{27}$$

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 C.13 (Ours)**.** *Consider Setup 3.6. Let $\mathcal{D}$ denote the collection of diagonal matrices formed from polynomially smooth (not necessarily bounded) coordinatewise images of vectors in the program:*

$$\mathcal{D} = \left\{ \operatorname{diag}\left(\psi\left(g^1, \dots, g^M\right)\right) : \ \psi : \mathbb{R}^M \to \mathbb{R} \text{ is polynomially smooth} \right\}. \tag{28}$$

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

We use the same tensor program to compute $c^i = \frac{1}{n}\sum_{\alpha \in [n]} z_\alpha^i (P_i(\mathcal{W}_n^{j_i}) z^i)_\alpha$ required to compute the traces $\tau_i = \mathbb{E}_{z^i \sim \mathcal{N}(0,I)} c^i$. Since we now rely on a theorem that does not require linearly bounded nonlinearities, the nonlinearities of the program at hand and ψ in the definition of $\mathcal{D}$ can be polynomially smooth. We embed the programs for $P_i(\mathcal{W}_n^{j_i})z$ into the "outer" program that computes the scalar $\frac{1}{n}\sum_{\alpha \in [n]} \bar{z}_\alpha \left(\prod_{i=1}^k \left(P_i(\mathcal{W}_n^{j_i}) - c_i I\right) \bar{z}\right)_\alpha$. Taking the expectation over all randomness and applying Theorem C.3 gives Eq. (27).

D Non-Gaussian Master theorem for Lipschitz nonlinearities

One of the limitations of our Theorem 3.7 is smoothness requirement. In the present section, we prove a similar result that assumes nonlinearities to be Lipschitz but not necessarily smooth:

Setup D.1. *Consider Setup 3.6, but replace 3*) and 4*) with the following: 3**) all matrices A^i have independent entries drawn from sub-Gaussian distributions with zero mean and variance n^{-1}; 4**) all the nonlinearities ϕ are Lipschitz (but not necessarily smooth).*

Theorem D.2 (Non-Gaussian Master theorem for Lipschitz nonlinearities)**.** *Consider Setup D.1. Then, as $n \to \infty$, every scalar c^i converges to the same $\mathring{c}^i$ as in Theorem 3.4 almost surely and in mean:*

$$c^i \xrightarrow{\text{a.s. \& } L^1} \mathring{c}^i.$$

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

[24] This proves Conjecture A.4 of [31] for polynomially smooth nonlinearities

Corollary D.3 (Convergence to GP at initialization)**.** *Consider a neural network whose forward pass can be expressed as Eq.* (4) *with each matrix W^i corresponding to some A^j but not its transposed. Suppose 1) all the activation functions are Lipschitz; 2) input and output layer weights are initialized with iid standard Gaussian images; 3) entries of all hidden bias vectors are initialized with iid standard Gaussians; 4) weights of any other layer are initialized according to 3**) of Setup D.1; 5) output layer bias vector entries are initialized with zeros. Then at initialization, as width tends to infinity, the pre-activation output of any hidden neuron of any layer except for the first one converges weakly to a Gaussian Process (GP).*

Proof. We use the fact that when all matrices in Eq. (4) are not transposed, the limit in Theorem 3.4 (and hence in our Theorem D.2) 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 [29].

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, $\mathbb{E}\left[\psi(g^1_\alpha, \dots, g^B_\alpha)\right] \to \mathbb{E}_{z\sim\mathcal{N}(\mu,\Sigma)}\psi(z)$ from Theorem D.2. This means that on any batch of inputs of size B, outputs of a given neuron $g^1_\alpha, \dots, g^B_\alpha$ 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. □

Proof of Theorem D.2. Let λ be the maximal Lipschitz constant among all ϕ^i in the program. Fix some $\delta > 0$. Let $\phi^{i,\delta}$ be the corresponding polynomially smooth Lipschitz functions given by Lemma B.1. W.l.o.g., we substitute δ with $\delta/2L$ so that $\sup_{x\in\mathbb{R}^{2(i-1)}} |\phi^{i,\delta}(x) - \phi^i(x)| \le \delta$.

Define the "smoothed" version of the given Tensor Program:

$$g^{i,\delta}_\alpha \leftarrow \sum_{\beta=1}^{n} W^i_{\alpha\beta} x^{i,\delta}_\beta, \quad c^{i,\delta} \leftarrow \frac{1}{n}\sum_{\beta=1}^{n} x^{i,\delta}_\beta, \quad \text{where } x^{i,\delta}_\alpha = \phi^i(g^{1,\delta}_\alpha, \dots, g^{i-1,\delta}_\alpha; c^{1,\delta}, \dots, c^{i-1,\delta}). \tag{29}$$

for $i \in [M_0 + 1 : M]$, where all W^i are the same as in the original program. All input vectors and scalars coincide with the original program: $g^{i,\delta} = g^i$, $c^{i,\delta} = c^i$ for $i \in [M_0]$.

We consider the same weight interpolation $A^l(t)$ as in the main: $A^l(0)$ corresponds to Gaussian weights $\tilde{A}^l$ with the same mean and variance as A^l, while $A^l(1)$ corresponds to the original weights A^l. Consequently, $g^{i,\delta}(t)$ denotes a vector obtained using $A^1(t), \dots, A^L(t)$, and similarly for $c^{i,\delta}(t)$.

Lemma D.4. *Under premise of Theorem D.2, for any $i \in [M_0 + 1, M]$, there exists a polynomial P^i such that for any $n \ge 1$, any $\delta > 0$, and any $t \in [0, 1]$,*

$$\left\|x^i(t) - x^{i,\delta}(t)\right\|_2 \le \sqrt{n}\delta P^i\left(\lambda\|A^1(t)\|_2, \dots, \lambda\|A^L(t)\|_2\right). \tag{30}$$

Proof. We will drop the t-argument in the proof to lighten the exposition. We prove the statement by induction on i.

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

$$\begin{aligned}&\left\|x^{M_0+1}(t) - x^{M_0+1,\delta}(t)\right\|_2 \le \\ &\le \left\|\phi^{M_0+1}(g^1, \dots, g^{M_0}; c^1, \dots, c^{M_0}) - \phi^{M_0+1,\delta}(g^{1,\delta}, \dots, g^{M_0,\delta}, c^{1,\delta}, \dots, c^{M_0,\delta})\right\|_2 \le \sqrt{n}\delta.\end{aligned} \tag{31}$$

Here $P^{M_0+1} \equiv 1$; in particular, it does not depend on t.

Suppose the induction hypothesis holds for i. Then we get the following for $i + 1$:

$$\begin{aligned}\left\|x^{i+1} - x^{i+1,\delta}\right\|_2 &= \left\|\phi^i(g^1, \dots, g^i; c^1, \dots, c^i) - \phi^{i,\delta}(g^{1,\delta}, \dots, g^{i,\delta}; c^{1,\delta}, \dots, c^{i,\delta})\right\|_2 \le \\ &\le \left\|\phi^i(g^1, \dots, g^i; c^1, \dots, c^i) - \phi^{i,\delta}(g^1, \dots, g^i; c^1, \dots, c^i)\right\|_2 + \\ &+ \left\|\phi^{i,\delta}(g^1, \dots, g^i; c^1, \dots, c^i) - \phi^{i,\delta}(g^{1,\delta}, \dots, g^{i,\delta}; c^{1,\delta}, \dots, c^{i,\delta})\right\|_2 \le \\ &\le \sqrt{n}\delta + \sqrt{\sum_{\alpha=1}^{n}\left|\phi^{i,\delta}(g^1_\alpha, \dots, g^i_\alpha; c^1, \dots, c^i) - \phi^{i,\delta}(g^{1,\delta}_\alpha, \dots, g^{i,\delta}_\alpha; c^{1,\delta}, \dots, c^{i,\delta})\right|^2},\end{aligned} \tag{32}$$

where the last inequality holds due to the approximation property of ϕ^δ.

$$\sum_{\alpha=1}^{n} \left|\phi^{i,\delta}(g_\alpha^1,\dots,g_\alpha^i;c^1,\dots,c^i) - \phi^{i,\delta}(g_\alpha^{1,\delta},\dots,g_\alpha^{i,\delta};c^{1,\delta},\dots,c^{i,\delta})\right|^2 \leq$$

$$\leq \lambda^2 \sum_{\alpha=1}^{n} \sum_{i'=M_0+1}^{i} \left(\left|g_\alpha^{i'} - g_\alpha^{i',\delta}\right|^2 + |c^{i'} - c^{i',\delta}|^2\right) =$$

$$= \lambda^2 \sum_{i'=M_0+1}^{i} \left(\|g^{i'} - g^{i',\delta}\|_2^2 + n|c^{i'} - c^{i',\delta}|^2\right) \leq \lambda^2 \sum_{i'=M_0+1}^{i} \left(\|W^{i'}\|_2^2 + 1\right) \left\|x^{i'} - x^{i',\delta}\right\|_2^2 \leq$$

$$\leq n\delta^2\lambda^2 \sum_{i'=M_0+1}^{i} \left(\|W^{i'}\|_2^2 + 1\right) \left(P^{i'}(\lambda\|A^1\|_2,\dots,\lambda\|A^L\|_2)\right)^2 \quad (33)$$

by induction hypothesis. Plugging this expression back, we get

$$\left\|x^{i+1} - x^{i+1,\delta}\right\|_2 \leq \sqrt{n}\delta \left(1 + \lambda\sqrt{\sum_{i'=M_0+1}^{i} \left(\|W^{i'}\|_2^2 + 1\right) \left(P^{i'}(\lambda\|A^1\|_2,\dots,\lambda\|A^L\|_2)\right)^2}\right) \leq$$

$$\leq \sqrt{n}\delta \left(1 + \lambda \sum_{i'=M_0+1}^{i} \left(\|W^{i'}\|_2 + 1\right) P^{i'}(\lambda\|A^1\|_2,\dots,\lambda\|A^L\|_2)\right). \quad (34)$$

Define $j_{i'}$ such that $W^{i'}$ equals $A^{j_{i'}}$ or its transposition for every $i' \in [M_0+1, M]$. Define the new polynomial P^{i+1} as

$$P^{i+1}(x_1,\dots,x_L) = 1 + \lambda \sum_{i'=M_0+1}^{i} \left(x_{j_{i'}} + 1\right) P^{i'}(x_1,\dots,x_L), \quad (35)$$

which does not depend on t since neither of $P^{i'}$ for $i' \leq i$ do. This proves the induction. □

The above lemma implies for any $t \in [0,1]$,

$$|c^i(t) - c^{i,\delta}(t)| = \frac{1}{n}\|x^i(t) - x^{i,\delta}(t)\|_1 \leq \frac{1}{\sqrt{n}}\|x^i(t) - x^{i,\delta}(t)\|_2 \leq$$

$$\leq \delta P^i(\lambda\|A^1(t)\|_2,\dots,\lambda\|A^L(t)\|_2). \quad (36)$$

For any $p \in [1,\infty)$,

$$\mathbb{E}\,|c^i(1) - c^i(0)|^p \leq 3^p\mathbb{E}\,|c^i(1) - c^{i,\delta}(1)|^p + 3^p\mathbb{E}\,|c^{i,\delta}(1) - c^{i,\delta}(0)|^p + 3^p\mathbb{E}\,|c^i(0) - c^{i,\delta}(0)|^p. \quad (37)$$

The second term goes to zero as $n \to \infty$ for any $\delta > 0$ by our "smooth" Master theorem, Theorem 3.7 (see also the reasoning in Section 5). The first and the last terms are bounded as follows:

$$\mathbb{E}\,|c^i(1) - c^{i,\delta}(1)|^p \leq \delta^p\mathbb{E}\left[\left(P^i(\lambda\|A^1\|_2,\dots,\lambda\|A^L\|_2)\right)^p\right], \quad (38)$$

$$\mathbb{E}\,|c^i(0) - c^{i,\delta}(0)|^p \leq \delta^p\mathbb{E}\left[\left(P^i(\lambda\|\tilde{A}^1\|_2,\dots,\lambda\|\tilde{A}^L\|_2)\right)^p\right], \quad (39)$$

which are both bounded uniformly wrt n by $C\delta^p$ for some constant C independent on n since all $A^1,\dots,A^L$ and $\tilde{A}^1,\dots,\tilde{A}^L$ are sub-Gaussian and therefore have moments, which are uniformly bounded wrt n. Therefore

$$\lim_{n\to\infty} \mathbb{E}\,|c^i(1) - c^i(0)|^p \leq 2C\delta^p. \quad (40)$$

Taking infium over $\delta > 0$ gives simply $\lim_{n\to\infty} \mathbb{E}\,|c^i(1) - c^i(0)|^p = 0$, which means that $c^i(1) - c^i(0)$ converges to zero in L^p. Since this is true for any $p \geq 1$, similarly to the proof of Theorem 3.7, Borel-Cantelli lemma will imply that $c^i(1) - c^i(0)$ converges to zero almost surely. Since Gaussian Master theorem, Theorem 3.4, works for non-smooth nonlinearities, $c^i(1)$ converges almost surely to the same limit as $c^i(0)$.

Since all of our nonlinearities are Lipschitz, they are linearly bounded. For such nonlinearities, Gaussian Master theorem guarantees also convergence in mean, see Theorem A.1 of [31]. Therefore

$$\lim_{n\to\infty} \mathbb{E}\,|c^i(1) - \mathring{c}^i| \le \lim_{n\to\infty} \mathbb{E}\,|c^i(1) - c^i(0)| + \lim_{n\to\infty} \mathbb{E}\,|c^i(0) - \mathring{c}^i| = 0, \tag{41}$$

which means that $c^i(1)$ converges to $\mathring{c}^i$ in mean.

□

E Tensor Program formulation equivalence

The Tensor Program series [29, 30, 31, 34, 32, 33] 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. (4), respectively. In its body, the program can also generate X-vars (that correspond to a vector x in Eq. (4)), 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, $\dots$, g^k : G, c^1 : C, $\dots$, c^l : C. Output: x : X, where $x_\alpha = \phi(g^1_\alpha, \dots, g^k_\alpha, c^1, \dots, c^l)\ \forall \alpha \in [n]$.
- Moment. Input: g^1 : G, $\dots$, g^k : G, c^1 : C, $\dots$, c^l : C. Output: c : C, where $c = \frac{1}{n}\sum_{\alpha=1}^n \phi(g^1_\alpha, \dots, g^k_\alpha, c^1, \dots, c^l)$.

It is easy to see that the iteration of Eq. (4) 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. (4) 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:

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

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. (4).

F Comparison with Chen and Lam [2]

The work of [2] considers the following iteration:

$$\tilde{x}^2_\alpha = \phi^2\left(\sum_\delta \tilde{A}_{\alpha\beta} x^1_\beta\right), \quad \tilde{x}^m_\alpha = \phi^m\left(\sum_\delta \tilde{A}_{\alpha\beta}\tilde{x}^{m-1}_\beta, \tilde{x}^{m-2}_\alpha, \dots, \tilde{x}^2_\alpha, x^1_\alpha\right) \quad \forall m > 2\ \forall \alpha \in [n], \tag{42}$$

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:

$$x_\alpha^2 = \phi^2\left(\sum_\delta A_{\alpha\beta}x_\beta^1\right), \quad x_\alpha^m = \phi^m\left(\sum_\delta A_{\alpha\beta}x_\beta^{m-1}, x_\alpha^{m-2}, \ldots, x_\alpha^2, x_\alpha^1\right) \quad \forall m > 2\ \forall \alpha \in [n], \tag{43}$$

The main result of [2] follows:

Theorem F.1 ([2]). *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*

$$\lim_{n\to\infty}\left|\frac{1}{n}\sum_{\alpha=1}^n\left(\psi(x_\alpha^1,\ldots,x_\alpha^M) - \psi(\tilde{x}_\alpha^1,\ldots,\tilde{x}_\alpha^M)\right)\right| = 0 \tag{44}$$

in probability.

In this formulation, $\frac{1}{n}\sum_{\alpha=1}^n \psi(x_\alpha^1,\ldots,x_\alpha^M)$ does not converge to $\mathbb{E}_{z\sim\mathcal{N}(\mu,\Sigma)}\psi(z)$ for some μ and Σ since $x_\alpha^1,\ldots,x_\alpha^M$ are images of nonlinear functions. Therefore the above formulation does not allow for a clear Gaussian process interpretation as in our Corollary 4.3 or Corollary D.3.

One can show equivalence of our iteration (4) 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 (4), 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_α^j, $\bar{x}_\alpha^j$, 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 (42) for expressing deep learning computations is very limited. First, iteration (42) assumes weight matrices to be symmetric, while such weight initializations are rarely used in practice. Second, iteration (42) 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 (42) 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 4.3) 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 F.1, certain crucial details are different. While Theorem F.1 requires everything to be Lipschitz, our Theorem 3.7 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 4.3 and Corollary 4.4) requires a quadratic ψ-function. Finally, they prove only convergence in probability, while our Theorem 3.7 states almost sure convergence and convergence in L^p for any p, which is much stronger.

G Proof of Lemma 5.3

We prove Lemma 5.3, recalled here for convenience.

Lemma 5.3 (Interpolation Properties). *For any matrix entry $a(t) = A_{\alpha\beta}^i(t)$ of a program in Setup 3.6: (1) $\mathbb{E}\,\dot{a}(t) = \mathbb{E}\,a(t)\dot{a}(t) = 0$ for all t; (2) For any integers $j, k \geq 0$ with sum $\ell = j + k$, $\sup_t \mathbb{E}\,|a(t)^j\dot{a}(t)^k| \leq \pi^\ell\nu_\ell n^{-\ell/2}$, where ν_ℓ is the scaled moment bound in Setup 3.6.*

Beyond our TP setting, our proof shows that these two properties hold for any interpolation $a(t) = a_1\cos\frac{\pi}{2}t + a_2\sin\frac{\pi}{2}t$ between centered random variables a_1 and a_2 with identical variance $1/n$ and satisfying the scaled moment bound $\mathbb{E}\,|a_1|^k, \mathbb{E}\,|a_2|^k \leq \nu_k/n^{-k/2}$ for all $k \geq 3$.

Proof. Write

$$a(t) = \tilde{a}\cos\frac{\pi}{2}t + a\sin\frac{\pi}{2}t$$

where $\tilde{a}$ is the Gaussian random variable and a is the non-Gaussian random variable. Then

$$\dot{a}(t) = \frac{\pi}{2}(-\tilde{a}\sin\frac{\pi}{2}t + a\cos\frac{\pi}{2}t).$$

Then $\mathbb{E}\,\dot{a}(t) = 0$ follows from the zero-mean property of $\tilde{a}$ and a. Likewise, straightforward calculation using the independence between $\tilde{a}$ and a shows $\mathbb{E}\,a(t)\dot{a}(t) = 0$.

To show (2), first note that, for any t,

$$|a(t)|, \frac{2}{\pi}|\dot{a}(t)| \le |\tilde{a}| + |a|.$$

Then

$$\mathbb{E}\,|a(t)^j\dot{a}(t)^k| \le \left(\frac{\pi}{2}\right)^k \sum_{i=0}^{k}\binom{\ell}{i}\mathbb{E}\,|\tilde{a}|^i|a|^{\ell-i}.$$

By Hölder's inequality,

$$\mathbb{E}\,|\tilde{a}|^i|a|^{\ell-i} \le (\mathbb{E}\,|\tilde{a}|^\ell)^{\frac{i}{\ell}}(\mathbb{E}\,|a|^\ell)^{\frac{\ell-i}{\ell}} \le \nu_\ell n^{-\ell/2}.$$

Therefore,

$$\mathbb{E}\,|a(t)^j\dot{a}(t)^k| \le \left(\frac{\pi}{2}\right)^k \sum_{i=0}^{k}\binom{\ell}{i}\nu_\ell n^{-\ell/2} = \left(\frac{\pi}{2}\right)^k 2^\ell\nu_\ell n^{-\ell/2} = \pi^\ell\nu_\ell n^{-\ell/2}.$$

□

H Technical Preliminaries

H.1 L^p Norm

For any vector $v \in \mathbb{R}^k$ and $p \ge 1$, we write $\|v\|_p$ to denote its L^p norm $\|v\|_p \stackrel{\text{def}}{=} \sqrt[p]{|v_1|^p + \cdots + |v_k|^p}$. When $p = 2$, we will just write $\|v\| = \|v\|_2$ when there's no cause for confusion. The following L^p norm bound is standard.

Lemma H.1. *For any vector $b \in \mathbb{R}^k$, if $p \le q$, then*

$$\|b\|_q \le \|b\|_p \le k^{\frac{1}{p}-\frac{1}{q}}\|b\|_q$$

We will use the following trivial but useful fact repeatedly. It follows trivially from Lemma H.1.

Lemma H.2. *For any integer $m \ge 0$ and reals $a_i \in \mathbb{R}$, $i \in [k]$,*

$$\left|\sum_{i=1}^{k} a_i\right|^m \le k^{m-1}\sum_{i=1}^{k}|a_i|^m.$$

H.2 Multisets

Definition H.3 (Multiset)**.** A *multiset* is a set allowing multiple occurences of the same element, e.g., $\{1,1,1,2,2,3\}$ (which is not equal to $\{1,2,3\}$ as a multiset). We will use capital italic font to denote multisets, such as $\mathcal{P}$. Thus, e.g., when we write $\sum_{p\in\mathcal{P}} f(p)$, p could take the same value multiple times. The number of elements in $\mathcal{P}$, counting multiplicity, is denoted $|\mathcal{P}|$. The set of *unique* elements is denoted $\mathrm{uniq}(\mathcal{P})$.

Definition H.4 (Partition of multiset)**.** A *partition* τ of a multiset $\mathcal{P}$ expresses $\mathcal{P}$ as a disjoint union of multisets. Concretely, τ is a multiset $\{\mathcal{P}_1,\ldots,\mathcal{P}_k\}$ such that $\mathcal{P} = \bigsqcup_i \mathcal{P}_i$, and $|\tau| = k$ is the number of sets in the partition. For example, $\mathcal{P} = \{1,1,1,2,2,3\} = \{1,1\}\sqcup\{2,2\}\sqcup\{1,3\}$, so that $\tau = \{\{1,1\},\{2,2\},\{1,3\}\}$ is a partition of $\mathcal{P}$. As another example, $\mathcal{P} = \{1\}\sqcup\{1\}\sqcup\{1\}\sqcup\{2\}\sqcup\{2\}\sqcup\{3\}$ is also a partition, with repeated sets, which emphasizes the fact that τ is allowed to be a multiset.

Obviously, the typical notion of a partition of a *set* is just a special case of the above concept applied to sets.

Definition H.5 (Partition induced by multiset)**.** Given a multiset $\mathcal{P}$, we can count the multiplicity of each element in $\mathcal{P}$ and list them in nonincreasing order, say $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_k \geq 1$ where $\sum_{i=1}^k \pi_i = |\mathcal{P}|$. Then the multiset $\pi = \{\pi_i\}_{i=1}^k$ form a *partition* $\pi(\mathcal{P})$ of the integer $|\mathcal{P}|$, which we call the *partition induced by* $\mathcal{P}$, and we denote $|\pi| \stackrel{\text{def}}{=} k$, the *size of the partition*. Again, note that $\pi(\mathcal{P})$ does not contain any information about the identity of elements in $\mathcal{P}$, only their counts.

For example, $\mathcal{P} = \{\bullet, \bullet, \bullet, \star, \star, \times\}$ would induce the partition $|\mathcal{P}| = 6 = 3 + 2 + 1$ because 3 is the multiplicity of $\bullet$, 2 is the multiplicity of $\star$, and 1 is the multiplicity of $\times$.

Note that the partitions of integer n are equivalent to the partitions of the multiset $\underbrace{\{1, \ldots, 1\}}_{n}$.

Definition H.6 (Multiset from vector)**.** A multiset can be obtained from a vector by forgetting the order of the vector's entries, e.g., $\{1, 1, 1, 2, 2, 3\}$ can obtained this way from the vector $(1, 2, 1, 3, 2, 1)$ or $(2, 3, 1, 1, 2, 1)$. We will use capital bold font to denote such vectors in the context of multisets, such as $\mathbf{P} = (P_1, P_2, \ldots)$, and such obtained multiset would be written as $\tilde{\mathbf{P}} = \{P_1, P_2, \ldots\}$.

Definition H.7 ($N_{\mathcal{P}}$)**.** For any fixed multiset $\mathcal{P}$, we can ask how many distinct vectors $\mathbf{P}$ are there such that $\mathcal{P} = \tilde{\mathbf{P}}$. The answer is the multinomial coefficient $N_{\mathcal{P}} \stackrel{\text{def}}{=} \binom{|\mathcal{P}|}{\pi} = \binom{|\mathcal{P}|}{\pi_1; \cdots; \pi_k}$, where $\pi = \pi(\mathcal{P})$ is the partition induced by $\mathcal{P}$. Note that $N_{\mathcal{P}}$ depends on $\mathcal{P}$ only through π.

For instance, for $\mathcal{P} = \{\bullet, \bullet, \bullet, \star, \star, \times\}$, we have $N_{\mathcal{P}} = \binom{6}{3;2;1}$.

H.3 Monomials

Given a vector $\boldsymbol{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k$ and a vector $\mathbf{P} \in [k]^r$ for some integer $r \geq 0$, we can form the monomial $\boldsymbol{x}^{\mathbf{P}} \stackrel{\text{def}}{=} \prod_{i=1}^r x_{P_i}$. Thus, if 1 appears in $\mathbf{P}$ t times, then $\boldsymbol{x}^{\mathbf{P}}$ contains a factor of x_1^t.

If we also have a multiset $\mathcal{P}$ taking values in $[k]$, then we can also form the monomial $\boldsymbol{x}^{\mathcal{P}} \stackrel{\text{def}}{=} \prod_{\alpha \in \mathcal{P}} x_\alpha$. In this notation, we would have $\boldsymbol{x}^{\mathbf{P}} = \boldsymbol{x}^{\tilde{\mathbf{P}}}$.

H.4 Tensors

Given vectors $x \in \mathbb{R}^k, y \in \mathbb{R}^l$, their tensor $x \otimes y$ is the vector in $\mathbb{R}^{k \times l}$ with entries $(x \otimes y)_{(i,j)} = x_i y_j$, where the index ranges over $(i, j) \in [k] \times [l]$. We will use angle brackets $\langle -, - \rangle$ to denote standard inner product on all Euclidean spaces, so that

$$\langle x \otimes y, x' \otimes y' \rangle = \langle x, x' \rangle \langle y, y' \rangle \tag{45}$$

if $x, x' \in \mathbb{R}^k, y, y' \in \mathbb{R}^l$.

We can tensor a vector $x \in \mathbb{R}^k$ with itself $p - 1$ times to form a *tensorial power* $x^{\otimes p} \in \mathbb{R}^{k^p}$ of x. Its entries would be indexed by elements $\mathbf{P}$ of $[k]^p$, with $(x^{\otimes p})_{\mathbf{P}} = x^{\mathbf{P}}$. Like in Eq. (45), we have

$$\langle x^{\otimes p}, y^{\otimes p} \rangle = \langle x, y \rangle^p.$$

for any x, y in the same Euclidean space.

This is particularly relevant as we will often encounter the expression $(\sum_\alpha x_\alpha)^p$ which can be written as

$$\left(\sum_\alpha x_\alpha \right)^p = \langle 1, x \rangle^p = \langle 1^{\otimes p}, x^{\otimes p} \rangle = \sum_{\mathbf{P} \in [k]^p} x^{\mathbf{P}}. \tag{46}$$

By Definition H.7, we can further rewrite this as

$$\left(\sum_\alpha x_\alpha \right)^p = \sum_{\mathcal{P}} N_{\mathcal{P}} x^{\mathcal{P}} \tag{47}$$

where $\mathcal{P}$ ranges over all multisets of size $|\mathcal{P}| = p$ taking value in $[k]$.

Spectral Norms for Tensors Let A be a r-th order tensor, $A \in \mathbb{R}^{k_1 \times \ldots \times k_r}$. We define its spectral norm, $\|A\|$, as follows:

$$\|A\| = \sup_{x_1 \in \mathbb{R}^{k_1},\, \|x_1\|=1} \cdots \sup_{x_r \in \mathbb{R}^{k_r},\, \|x_r\|=1} = \left| \left\langle A, \bigotimes_{i=1}^{r} x_i \right\rangle \right|. \tag{48}$$

Note that this definition is consistent with the definition of spectral norm of matrices. For any r-th order tensor A and vectors $x_1 \in \mathbb{R}^{k_1}, \ldots x_r \in \mathbb{R}^{k_r}$, we have the following inequality:

$$\left| \left\langle A, \bigotimes_{i=1}^{r} x_i \right\rangle \right| \leq \|A\| \prod_{i=1}^{r} \|x_i\|. \tag{49}$$

H.5 Higher Order Differentiation and Taylor Expansion

Scalar Function For a smooth multivariate scalar function $f(x) = f(x_1, \ldots, x_k)$ and a vector $\mathbf{P} \in [k]^r$ with multiset $\mathcal{P} \stackrel{\text{def}}{=} \tilde{\mathbf{P}}$, we write

$$\partial_x^{\mathbf{P}} f(x) = \partial_x^{\mathcal{P}} f(x) \stackrel{\text{def}}{=} \frac{\partial^r}{\partial x_{P_1} \cdots \partial x_{P_r}} f(x).$$

We also define $\nabla_x^p f(x) \in \mathbb{R}^{k^p}$ as the pth order tensor containing all order-p derivatives of f evaluated at x:

$$(\nabla_x^p f(x))_{\mathbf{P}} = \partial_x^{\mathbf{P}} f(x).$$

This is a highly symmetric tensor as $(\nabla_x^p f(x))_{\mathbf{P}} = (\nabla_x^p f(x))_{\mathbf{Q}}$ if $\tilde{\mathbf{P}} = \tilde{\mathbf{Q}}$ as multisets. In all notations, we will drop the subscript, e.g., $\partial^{\mathbf{P}} = \partial^{\mathbf{P}}$, when the variables of the differentiation are clear from context.

Note that we have used a superscript notation for $\partial^{\mathbf{P}}$ and $\partial^{\mathcal{P}}$ to emphasize that they denote a "product of (1st order) partial derivatives operators." In particular, $\partial^{\mathbf{P}} f(x)$ and $\partial^{\mathcal{P}} f(x)$ should be distinguished from the vector of 1st order partial derivatives $(\frac{\partial}{\partial x_{P_i}} f(x))_{i=1}^r$. They will always be scalar quantities when f is scalar, whereas we always use ∇^p for when we want to think of the collection of (higher order) partial derivatives as a vector or tensor.

In this notation, f's Taylor expansion around $x = 0$ can be written as

$$f(0) + \langle \nabla f(0), x \rangle + \frac{1}{2!} \langle \nabla^2 f(0), x^{\otimes 2} \rangle + \cdots + \frac{1}{p!} \langle \nabla^p f(0), x^{\otimes p} \rangle + \cdots .$$

Of course, in general this Taylor expansion is not exact unless f is analytic in the appropriate neighborhood. For our purpose, the Taylor expansion with remainder is more useful, as it is exact:

Lemma H.8 (Taylor expansion with remainder)**.** *If $f : \mathbb{R}^k \to \mathbb{R}$, $f(x) = f(x_1, \ldots, x_k)$ is C^{p+1} (i.e., $\nabla^p f(x)$ has continuous partial derivatives), then*

$$f(x) = f(0) + \langle \nabla f(0), x \rangle + \cdots + \frac{1}{p!} \langle \nabla^p f(0), x^{\otimes p} \rangle + R_{p+1},$$

where the remainder R_{p+1} is

$$R_{p+1} \stackrel{def}{=} \frac{1}{p!} \int_0^1 (1-t)^p \langle \nabla^{p+1} f(tx), x^{\otimes (p+1)} \rangle dt$$

Vector Function For the vector case $f : \mathbb{R}^k \to \mathbb{R}^l$, $f(x) = (f_1(x), \ldots, f_l(x))$, all of the above considerations apply to the components of f in parallel. For example,

$$\partial^{\mathbf{P}} f(x) = \partial^{\mathcal{P}} f(x) \stackrel{\text{def}}{=} (\partial^{\mathcal{P}} f_1(x), \ldots, \partial^{\mathcal{P}} f_l(x)) \in \mathbb{R}^l$$
$$\nabla^p f(x) \stackrel{\text{def}}{=} (\nabla^p f_1(x), \ldots, \nabla^p f_l(x)) \in \mathbb{R}^{l \times k^p}$$

H.6 Higher Order Chain Rule

Lemma H.9 ([10]). *Suppose $f : \mathbb{R}^l \to \mathbb{R}, y : \mathbb{R}^k \to \mathbb{R}^l$ are C^k and consider their composition $f(y) = f(y(x_1, \ldots, x_k))$. Then for any vector $\mathbf{P} \in [k]^r$,*

$$\partial_x^{\mathbf{P}} f(y) = \sum_{\tau} \langle \nabla_y^{|\tau|} f(y), \bigotimes_{\mathcal{P} \in \mathbf{P}[\tau]} \partial_x^{\mathcal{P}} y \rangle \tag{50}$$

where τ ranges over partitions of $\{1, \ldots, r\}$ and $\mathbf{P}[\tau]$ is the partition $\{\{P_i : i \in \mathcal{S}\} : \mathcal{S} \in \tau\}$ of $\tilde{\mathbf{P}}$.

Eq. (50) may look somewhat scary at first, but when we apply it, the range of τ (in the outer sum) and the range of $\mathcal{P}$ (in the inner tensor product) are both $O(1)$-sized, as $n \to \infty$. So when we want to bound $\partial_x^{\mathbf{P}} f(y)$ (which is the only time we will use Eq. (50)), what we end up caring about is only the individual norms of $\nabla_y^{|\tau|} f(y)$ and $\partial_x^{\mathcal{P}} y$. May this fact assuage the reader intimidated by the density of Eq. (50).

Nevertheless, let us take some time to digest Eq. (50).

To unpack the tensor notation, notice that 1) $\nabla_y^{|\tau|} f(y) \in \mathbb{R}^{l^{|\tau|}}$, 2) $\partial_x^{\mathcal{P}} y = (\partial_x^{\mathcal{P}} y_1, \ldots, \partial_x^{\mathcal{P}} y_l) \in \mathbb{R}^l$, so 3) $\bigotimes_{\mathcal{P} \in \mathbf{P}[\tau]} \partial_x^{\mathcal{P}} y$, being a tensor over $|\mathbf{P}[\tau]| = |\tau|$ vectors, is in $\mathbb{R}^{l^{|\tau|}}$ as well. So, for example, when $|\tau| = 1$, there is only one partition, consisting of the whole of $[k]$, and $\mathbf{P}[\tau] = \{\tilde{\mathbf{P}}\}$, so that the corresponding term in Eq. (50) is

$$\langle \nabla_y f(y), \partial_x^{\mathbf{P}} y \rangle = \sum_{j=1}^{l} \frac{\partial f}{\partial y_j} \frac{\partial^r y_j}{\partial x^{\mathbf{P}}}.$$

As a helping example, consider the case when $r = 2$ and $\mathbf{P} = (1, 2)$. Then we have

$$\partial_x^{\mathbf{P}} f(y) = \partial_{x_1} \partial_{x_2} f(y) = \partial_{x_1} \left(\langle \nabla_y f(y), \partial_{x_2} y \rangle \right) = \langle \nabla_y f(y), \partial_{x_1} \partial_{x_2} y \rangle + \langle \nabla_y^2 f(y), \partial_{x_1} y \otimes \partial_{x_2} y \rangle.$$

Here the first term corresponds to $\tau = \{[k]\}$; in this case, $\mathbf{P}[\tau] = \{\tilde{\mathbf{P}}\} = \{\{x_1, x_2\}\}$. The second term corresponds to $\tau = \{\{1\}, \{2\}\}$; in that case, $\mathbf{P}[\tau] = \{\{x_1\}, \{x_2\}\}$.

Note that Eq. (50) is not equivalent to the look-alike equation where we allow τ to range over partitions of the multiset $\tilde{\mathbf{P}}$ and letting $\mathcal{P}$ in the inner tensor product range over $\mathcal{P} \in \tau$. This alternative equation would lead to too few terms in the sum over τ, as is apparent when one considers $\mathbf{P} = (1, \ldots, 1)$.

Applying Eq. (49) to Eq. (50), we get

Lemma H.10. *In the same setting as in Lemma H.9, we have*

$$|\partial_x^{\mathbf{P}} f(y)| \le \sum_{\tau} \|\nabla_y^{|\tau|} f(y)\| \cdot \prod_{\mathcal{P} \in \mathbf{P}[\tau]} \|\partial_x^{\mathcal{P}} y\|. \tag{51}$$

We will in particular need to apply this repeatedly to the case when $f(y) = \prod_j y_j$:

Lemma H.11. *Suppose $y : \mathbb{R}^k \to \mathbb{R}^l$ is C^k, with $y(x)$ denoting the value of y on input x. Then for any vector $\mathbf{P} \in [k]^r$,*

$$\left\| \partial_x^{\mathbf{P}} \prod_j y_j \right\| \le \sum_{t=0}^{l-1} \sqrt{l^{l-t-1} \|y\|_{2t}^{2t}} \cdot \sum_{\tau} \prod_{\mathcal{P} \in \mathbf{P}[\tau]} \|\partial_x^{\mathcal{P}} y\|. \tag{52}$$

where τ ranges over partitions of $\{1, \ldots, r\}$ of size $l - t$ and $\mathbf{P}[\tau]$ is as in Lemma H.9. Here $\|y\|_0^0$ is by convention defined to be l.

The proof below actually shows the l^{l-t-1} in Eq. (52) can be refined to the falling factorial $(l-1) \cdots (t+1)$ but we will not need this. Also note that there are no partitions of size $> r$, so that the summand vanishes when $t < l - r$.

Proof. For $f(y) = \prod_j y_j$, by Eq. (51), we just need to show

$$\|\nabla_y^{|\tau|} f(y)\|^2 \le l^{|\tau|-1} \|y\|_{2(l-|\tau|)}^{2(l-|\tau|)}.$$

Each nonzero entry of $\nabla_y^{|\tau|} f(y)$ has the form y^S for some size-$(l-|\tau|)$ subset (i.e., having no duplicates) S of $[l]$. By power-mean inequality, $|y^S|^2 \le \frac{1}{|S|}\sum_{j\in S} |y_j|^{2|S|}$. Taking sum over all possible S of size $(l-|\tau|)$, we have

$$\sum_S |y^S|^2 \le \binom{l}{|\tau|} \frac{1}{l} \sum_{j=1}^{l} |y_j|^{2(l-|\tau|)}$$

Finally, each y^S appears $|\tau|!$ times in $\nabla_y^{|\tau|} f(y)$, so

$$\|\nabla_y^{|\tau|} f(y)\|^2 = |\tau|! \sum_S |y^S|^2 \le l^{|\tau|-1} \sum_{j=1}^{l} |y_j|^{2(l-|\tau|)}$$

as desired. □

H.7 Smoothness Profile of A Function

Definition H.12 (Smoothness Profile)**.** Consider a smooth (C^∞) function $f : \mathbb{R}^k \to \mathbb{R}$. If there is a sequence $\mathcal{C}$ of positive reals $(C_1, p_1), (C_2, p_2), \dots$ such that for any $\mathbf{P} \in [k]^r$, we have $\left|\partial^{\mathbf{P}} f(x)\right| \le C_r(1 + |x_1|^{p_r} + \cdots + |x_k|^{p_r})$, then we say $\mathcal{C}$ is a *smoothness profile*, or just *profile* for short, of f.

Obviously, f is polynomially smooth (Definition 3.5) iff it has a profile.

As one could guess, the "expected smoothness" of a program's vectors (under perturbation of its matrices) will depend on the nonlinearities ϕ^i of the program *only* through the profiles of ϕ^i. Thus, there are smoothness guarantees that are uniform over all programs whose nonlinearities have the same profiles.

I The Basic Moment Argument

Throughout this work, we need to study many different sums of random variables $X = \sum_{\alpha=1}^{n} x_\alpha$. In particular, we will repeatedly use a basic argument to bound its moment $\mathbb{E}\, X^p$ for even p, which is formally codified in Lemma I.4 at the end of this section. But to motivate this eventual formulation, it's nice to start with a simple example.

Simple Motivating Example The simplest example would be when x_α are iid random variables that don't depend on n (but eventually we will need to cover the case of general non-iid distributions depending on n, with small correlations among x_α). Then for even p, by Eq. (47),

$$\mathbb{E}\, X^p = \sum_{\mathcal{P}} N_{\mathcal{P}} \mathbb{E}\, x^{\mathcal{P}} \tag{53}$$

where $\mathcal{P}$ ranges over all multisets of size $|\mathcal{P}| = p$ taking value in $[n]$ and $N_{\mathcal{P}}$ is the multinomial coefficient defined in Definition H.7. In our case, with each x_α being iid, it's clear that $\mathbb{E}\, x^{\mathcal{P}} = \mathbb{E}\, x^{\mathcal{P}'}$ if the multisets $\mathcal{P}, \mathcal{P}'$ have the samed induced partition: $\pi(\mathcal{P}) = \pi(\mathcal{P}')$. For example, $\mathbb{E}\, x_1 x_1 x_2 x_3 = \mathbb{E}\, x_1 x_1 x_3 x_4 = \mathbb{E}\, x_4 x_4 x_n x_{n-1}$ because the corresponding multisets all have the same induced partition $\{2, 1, 1\}$ of 4.

For any fixed partition τ of p, we can ask how many multisets $\mathcal{P}$ are there with $\pi(\mathcal{P}) = \pi$. The answer is quite straightforward but requires a bit of explanation notationally: we think of τ as another multiset and *take its induced partition* $\pi(\tau)$ of the integer $|\tau|$; then the answer is the multinomial coefficient $\binom{n}{\pi(\tau)}$. For the above example of $\tau = \{2, 1, 1\}$, we have $\pi(\tau) = \{1, 2\}$, because a) 1 in $\pi(\tau)$ is the multiplicity of 2 in τ and b) 2 in $\pi(\tau)$ is the multiplicity of 1 in τ. So there are $\binom{n}{1;2}$ multisets $\mathcal{P}$ with $\pi(\mathcal{P}) = \pi$.

Therefore, we can further rewrite Eq. (53) as a sum over partitions τ of integer p:

$$\mathbb{E}\, X^p = \sum_\tau \binom{n}{\pi(\tau)} N_{\mathcal{P}} \mathbb{E}\, x^{\mathcal{P}} \tag{54}$$

where for every τ, $\mathcal{P}$ is any choice of multiset with $\pi(\mathcal{P}) = \tau$. The advantage of this representation is that $\mathbb{E}\, X^p$ is now clearly a polynomial in n. This is because 1) τ ranges over a set that does not depend on n; 2) $N_{\mathcal{P}}$ does not depend on n; and 3) the distribution of x_α does not depend on n by our assumption. Thus, *as $n \to \infty$, we can obtain a bound on X^p by deriving the leading term of this polynomial.* Because $\binom{n}{\pi(\tau)} = \Theta(n^{|\tau|})$ (e.g., $\binom{n}{\pi(\{1,1,2\})} = \binom{n}{2;1} = \Theta(n^3)$), this means focusing on the the partitions τ with the largest $|\tau|$. In general, of course, this is just $\tau = \{1, \dots, 1\}$ where $|\tau| = p$, corresponding to monomials like $x_1 x_2 \cdots x_p$. But in the cases we are concerned with, $\mathbb{E}\, x^{\mathcal{P}}$ for this τ will vanish, so we need to look at lower order terms.

For example, assume further that x_α has mean 0 and variance 1. Then $\mathbb{E}\, x^{\mathcal{P}} = 0$ for any $\mathcal{P}$ that contains an element appearing with multiplicity 1 (e.g., $\mathbb{E}\, x_1 x_2 x_2 = 0$ because it's equal to $(\mathbb{E}\, x_1)(\mathbb{E}\, x_2^2) = 0 \cdot (\mathbb{E}\, x_2^2) = 0$). Therefore, the leading term in the polynomial Eq. (53) corresponds exactly to the partition $\tau = \{2, \dots, 2\}$ (where $|\tau| = p/2$) of p (corresponding to monomials like $\mathbb{E}\, x_1^2 \cdots x_{p/2}^2$ which is equal to 1 by our variance-1 assumption), and any other τ must either have smaller $|\tau|$ or it has $\mathbb{E}\, x^{\mathcal{P}} = 0$. Therefore, in this case, we deduce

$$\mathbb{E}\, X^p = \Theta(n^{p/2}), \quad \text{as } n \to \infty. \tag{55}$$

We can of course say more precise things about $\mathbb{E}\, X^p$ under these assumptions, which we record below.

Proposition I.1. *If $x_\alpha, \alpha \in [n]$, are sampled iid from a distribution with kth moment ν_k, then for any even integer p,*

$$\mathbb{E}\left(\sum_{\alpha=1}^n x_\alpha\right)^p = \sum_\tau \binom{n}{\pi(\tau)} \binom{p}{\tau} \nu^\tau \tag{56}$$

summing over partitions τ of integer p and any n.

Here $\nu^\tau = \prod_{k \in \tau} \nu_k$ following the multiset notation in Appendix H.3.

Slightly more generally, we have

Proposition I.2. *If $x_\alpha, \alpha \in [n]$, are sampled independently and the distribution of x_α has kth moment bounded above by ν_k, then for any even integer p,*

$$\mathbb{E}\left(\sum_{\alpha=1}^n x_\alpha\right)^p \le \sum_\tau \binom{n}{\pi(\tau)} \binom{p}{\tau} \nu^\tau \tag{57}$$

summing over partitions τ of integer p and any n.

General Case Now, when we consider x_α that are not iid and may correlate amongst themselves, Eq. (53) will continue to hold but in general Eq. (54) will not. Furthermore, $\mathbb{E}\, x^{\mathcal{P}}$ in the general case can also depend on n. So even if there's enough symmetry to obtain Eq. (54), we no longer have a polynomial expression of $\mathbb{E}\, X^p$ in n.

Nevertheless, most of the time, our objective is to show that $\mathbb{E}\, X^p$ is $O(1)$ as $n \to \infty$, and the above arguments can be easily adapted to work given the following condition:

Definition I.3 (Small Moment Condition)**.** Consider a sequence x of random vectors $(x(n) \in \mathbb{R}^n)_{n=1}^\infty$. We say x satisfies the *Small Moment Condition (SMC)* if for every even integer p, there exists a constant C such that

$$\mathbb{E}\, x(n)^{\mathcal{P}} \le C n^{-|\mathrm{uniq}(\mathcal{P})|} \tag{58}$$

for every multiset $\mathcal{P}$ of size p and for any n, where $\mathrm{uniq}(\mathcal{P})$ is the set of unique elements of $\mathcal{P}$.

Lemma I.4. *Consider a sequence x of random vectors $(x(n) \in \mathbb{R}^n)_{n=1}^\infty$. If x satisfies the Small Moment Condition (Definition I.3), then for any real $p \ge 1$,*

$$\mathbb{E}\left|\sum_{\alpha=1}^n x(n)_\alpha\right|^p = O(1) \quad \textit{as } n \to \infty. \tag{59}$$

Here the constant in $O(1)$ depends only on p and the constant C in Definition I.3.

Proof. Note that, by the power-mean inequality, it suffices to prove this for even integer p. Let $X(n) \stackrel{\text{def}}{=} \sum_{\alpha=1}^{n} x(n)_\alpha$. Below we suppress the argument (n) notationally. Note that $|\mathrm{uniq}(\mathcal{P})| = |\pi(\mathcal{P})|$, for $\pi(\mathcal{P})$ defined in Definition H.5. By Eq. (53), we have, for every n,

$$\mathbb{E}\, X^p \le \sum_{\mathcal{P}} N_{\mathcal{P}} C n^{-|\pi(\mathcal{P})|} \le C' \sum_{\mathcal{P}} n^{-|\pi(\mathcal{P})|}$$

summing over multisets $\mathcal{P}$ of size p taking value in $[n]$. where C is the constant in Definition I.3 and $C' = C \max_{\mathcal{P}} N_{\mathcal{P}}$. Then the same symmetry argument leading to Eq. (54) yields

$$\mathbb{E}\, X^p \le C' \sum_{\tau} \binom{n}{\pi(\tau)} n^{-|\tau|} = \sum_{\tau} O(1) = O(1),$$

where τ ranges over all partitions of integer p. □

As an example, if $x_\alpha = x(n)_\alpha, \alpha \in [n]$, are sampled independently and the distribution of x_α has kth moment bounded above by $\nu_k / n^{k/2}$, where ν_k are independent of n, then $\mathbb{E}\, x^{\mathcal{P}} \le \nu^{\pi(\mathcal{P})} n^{-p/2}$ (where $\nu^{\pi(\mathcal{P})} = \prod_{k \in \pi(\mathcal{P})} \nu_k$). If each x_α further has mean 0, then as discussed above Eq. (55), $\mathbb{E}\, x^{\mathcal{P}} = 0$ if $|\pi(\mathcal{P})| > p/2$. For $\mathcal{P}$ with $|\pi(\mathcal{P})| \le p/2$, we have $\mathbb{E}\, x^{\mathcal{P}} \le \nu^{\pi(\mathcal{P})} n^{-p/2} \le C n^{-|\pi(\mathcal{P})|/2}$ where $C = \max_\tau \nu^\tau$ taken over all partitions τ of integer p. Thus SMC (Definition I.3) is satisfied, and Lemma I.4 applies to yield

Proposition I.5. *Consider a sequence x of random vectors $(x(n) \in \mathbb{R}^n)_{n=1}^{\infty}$. Suppose $x_\alpha = x(n)_\alpha, \alpha \in [n]$, are sampled independently. Assume there are $\nu_k, k = 1, 2, \ldots$, independent of n such that $\mathbb{E}\,|x_\alpha|^k \le \nu_k n^{-k/2}$ for all n and k. Then x satisfies Small Moment Condition (Definition I.3), and for any $p \ge 1$,*

$$\mathbb{E}\left|\sum_{\alpha=1}^{n} x(n)_\alpha\right|^p = O(1) \quad \text{as } n \to \infty. \tag{60}$$

Here the constant in $O(1)$ depends on p and $\nu_1, \ldots, \nu_p$ (i.e., the first p moment bounds only).

J A Priori Moment Controls

Our main result in this section is Lemma J.3, which bounds the moments of derivatives of vector entries in the program. We shall come to it after stating some definitions.

Oblivious Constants In this and following sections, we need to reason carefully about constants hidden in big-O expressions. In particular, we isolate the following notion of *oblivious constant*, which roughly means that the constant does not depend on the fine details of a program beyond some finite number of smoothness and moment bounds.

For the first time reader, exactly understanding this notion is not a priority. So it may help to skip ahead to Lemma J.3, keeping in mind just this intuitive understanding of *oblivious constants*, and come back only after absorbing key ideas of our proofs.

Definition J.1. Consider a program T in Setup J.2, nonlinearities $\psi^1, \ldots, \psi^l$ for some $l \ge 0$, and multisets $\mathcal{P}_1, \ldots, \mathcal{P}_r$ for some $r \ge 0$. In the context of a bound or a big-O expression, we say a constant C is *$(\mathcal{P}_1, \ldots, \mathcal{P}_r)$-oblivious wrt T and $\psi^1, \ldots, \psi^l$* if all of the following hold.

1. C does not depend on n
2. C depends on each $\mathcal{P}_i$ only through its size $|\mathcal{P}_i|$
3. For each combination of $|\mathcal{P}_1|, \ldots, |\mathcal{P}_r|$, there is an integer $K > 1$ such that
   (a) For any sequence $\nu_2, \nu_3, \ldots \ge 0$ such that $\mathbb{E}\,|a|^k \le \nu_k n^{-k/2}$ for all matrix entries a of T, our constant C can be taken to depend on the distributions of matrix entries in T only through $\nu_2, \ldots, \nu_K$
   (b) For any profile $((C_0, p_0), (C_1, p_1), \ldots)$ (Definition H.12) satisfied by all nonlinearities ϕ^i of the program T as well as $\psi^1, \ldots, \psi^l$, our constant C can be taken to depend on $\{\phi^i\}_i$ and $\{\psi^j\}_j$ only through $(C_0, p_0), \ldots, (C_K, p_K)$

(c) For any sequence $R_1, R_2, \ldots$ such that $\mathbb{E}\,|c^i|^q \le R_q$ for all initial scalars c^i (i.e., where $i \in [M_0]$), all integers $q \ge 1$, and all n, our constant C can be taken to depend on $\{c^i\}_{i=1}^{M_0}$ only through $R_1, \ldots, R_K$.
(d) Furthermore, C is a oblivious function of $\nu_2, \ldots, \nu_K$, $(C_0, p_0), \ldots, (C_K, p_K)$, and $R_1, \ldots, R_K$.

We will just say $(\mathcal{P}_1, \ldots, \mathcal{P}_r)$-*oblivious* if the program and $\psi^1, \ldots, \psi^l$ are clear from context. Finally, any of the sets $\mathcal{P}_i$ can be an integer p, which just stands for the multiset $\{\underbrace{1, \ldots, 1}_{p}\}$.[25]

Thus, if an expression $\Lambda = \Lambda(\mathcal{P}, T, \psi)$ is bounded by a $\mathcal{P}$-oblivious constant wrt a program T and additional nonlinearity ψ, then

$$\sup_{\mathcal{P},T,\psi} \Lambda \le F(p, \mathcal{C}, (\nu_j)_j, (R_j)_j)$$

for some function F, where 1) $\mathcal{P}$ ranges over all multisets of size p, 2) ψ and all nonlinearities of T range over functions satisfying the profile $\mathcal{C}$, 3) the distributions of all matrix entries of T range of those that satisfy the moment bounds given by $(\nu_j)_j$, and 4) all initial scalars of T range over those that satisfy the moment bounds given by $(R_j)_j$.

Smoothness and Moment Control We relax Setup 3.6 slightly for our main result in this section.

Setup J.2. *Consider Setup 3.6, but relax 3*) to allow the variances of matrix entries to differ from n^{-1} but still bounded above by $\nu_2 n^{-1}$ for some $\nu_2 > 0$ common to all matrix entries.*

Lemma J.3 (Expected Smoothness of Vectors in a Program)**.** *Consider a Tensor Program under Setup J.2. Then, for any polynomially smooth $\psi : \mathbb{R}^M \to \mathbb{R}$, any $p \ge 1$, and any multiset $\mathcal{P}$ taking values in the program's matrix entries $\{A^i_{\alpha\beta}\}_{\alpha,\beta,i}$,*

$$\sup_{\alpha \in [n]} \mathbb{E}\left|\partial^{\mathcal{P}} \psi(g^1_\alpha, \ldots, g^M_\alpha; c^1, \ldots, c^M)\right|^p = O(1) \quad \text{as } n \to \infty. \tag{61}$$

where constant in the big-O is $(p, \mathcal{P})$-oblivious wrt ψ and the program.

We are slightly deviating from the partial derivative notation in Appendix H.5 as $\mathcal{P}$ now directly specify the variables to take derivatives against, rather than their indices.

One can interpret this result as saying that: any higher order derivative (with respect to weights) of any entry of $\psi(\cdots)$ will typically not explode to ∞ with n. The fact that the hidden constant is $(p, \mathcal{P})$-oblivious will be important when we prove our main result Theorem 5.2.

Remark J.4. As a sanity check, we discuss some features of this result before moving on to the proof. First, notice that the $\sup$ is outside the expectation. Were it the other way around, then we typically would expect some (function of) $\log n$ factors on the RHS of the bound.

Second, notice that if each matrix A^i has iid entries (instead of the more general case covered here where entries can come from different distributions), then by symmetry, the above expectation for all α would be identical, so the supremum is extraneous. But in general, this supremum is not extraneous.

Third, we assume the less stringent Setup J.2 instead of Setup 3.6 not just because we can but also because we need this in our inductive proof: we will need to reason about programs where some matrix entries are shrunken to 0, a condition that Setup J.2 captures but Setup 3.6 does not.

Remark J.5. This will hold for programs with variable dimensions, with the addendum that the constants $B_{p,|\mathcal{P}|}$ also can depend on hidden width ratios.

By applying power mean inequality to Lemma J.3, we also easily get

Lemma J.6. *Consider the same setting as Lemma J.3. Then*

$$\mathbb{E}\left|\partial^{\mathcal{P}} \frac{1}{n} \sum_{\alpha \in [n]} \psi(g^1_\alpha, \ldots, g^M_\alpha; c^1, \ldots, c^M)\right|^p = O(1) \quad \text{as } n \to \infty. \tag{62}$$

for the same hidden constant as in Eq. (61)*.*

However, later we will see that this bound is unnecessarily loose when $\mathcal{P}$ is not empty (Lemma K.6).

[25] In this case of $\mathcal{P}_i$ being an integer, condition 2 is then trivially satisfied, so the important condition is condition 3.

J.1 Proof of Lemma J.3: Induction Setup

In everything below, by *constant* we always mean something independent of n that may or may not depend on other data.

Fix the number of initial vectors M_0, as well as the sequence of scaled moment bounds $\nu_2, \nu_3, \ldots$, the profile $\mathcal{C}$, and the initial scalar moment bounds $R_1, R_2, \ldots$ as discussed in Setup J.2. We will prove the following claims simultaneously *for all programs that have M_0 initial vectors/scalars and satisfy the above constraints (specified by $(\nu_k)_k, \mathcal{C}, (R_k)_k$).* We do so by induction on the vector index j:

Claim 1(j) For any integer $p \geq 0$, there is a sequence of constants $B_{\mathrm{j},p,0}, B_{\mathrm{j},p,1}, \ldots$ such that

$$\sup_{\alpha \in [n]} \mathbb{E} \left| \partial^{\mathcal{P}} g_\alpha^{\mathrm{j}} \right|^p \leq B_{\mathrm{j},p,|\mathcal{P}|}$$

for any multiset $\mathcal{P}$ of the program's matrix entries. Furthermore, $B_{\mathrm{j},p,|\mathcal{P}|}$ is $(\mathrm{j}, p, \mathcal{P})$-oblivious.

Claim 2(j) For any polynomially smooth $\psi : \mathbb{R}^{2\mathrm{j}} \to \mathbb{R}$, any integer $p \geq 0$, there is a sequence of constants $B^{\psi}_{\mathrm{j},p,0}, B^{\psi}_{\mathrm{j},p,1}, \ldots$ such that

$$\sup_{\alpha \in [n]} \mathbb{E} \left| \partial^{\mathcal{P}} \psi(g_\alpha^1, \ldots, g_\alpha^{\mathrm{j}}; c^1, \ldots, c^{\mathrm{j}}) \right|^p \leq B^{\psi}_{\mathrm{j},p,|\mathcal{P}|}$$

for any multiset $\mathcal{P}$ of the program's matrix entries. Furthermore, $B^{\psi}_{\mathrm{j},p,|\mathcal{P}|}$ is $(\mathrm{j}, p, \mathcal{P})$-oblivious wrt to ψ and the program.

Note that $|\cdot|^0$ always equal 1 by convention in both claims. Of course, Claim $2(M)$ would yield Lemma J.3.

Obviously, Claim $1(\mathrm{j})$ is a special case of Claim $2(\mathrm{j})$, but our induction proof will go like this

$$\text{Claim } 2(\mathrm{j}-1) \implies \text{Claim } 1(\mathrm{j})$$
$$\text{Claim } 1(1, \ldots, \mathrm{j}) \text{ and Claim } 2(1, \ldots, \mathrm{j}-1) \implies \text{Claim } 2(\mathrm{j})$$

Before we begin the induction proof, we first record several consequences of the claims above.

Proposition J.7. *Recall ϕ^i denotes the ith nonlinearity of the program. Claim 2$(i-1)$ implies that*

$$\mathbb{E} \left| \partial^{\mathcal{P}} c^i \right|^p \leq B^{\phi^i}_{i-1,p,|\mathcal{P}|}$$

for any multiset $\mathcal{P}$ of the program's matrix entries.

Proof. Unwinding the definition of c^i (Eq. (4)), we have

$$\begin{aligned}
\mathbb{E} \left| \partial^{\mathcal{P}} c^i \right|^p &= \mathbb{E} \left| \frac{1}{n} \sum_{\alpha=1}^n \partial^{\mathcal{P}} \phi^i(g_\alpha^1, \ldots, g_\alpha^{i-1}; c^1, \ldots, c^{i-1}) \right|^p \\
&\leq \frac{1}{n} \sum_{\alpha=1}^n \mathbb{E} \left| \partial^{\mathcal{P}} \phi^i(g_\alpha^1, \ldots, g_\alpha^{i-1}; c^1, \ldots, c^{i-1}) \right|^p && \text{applying Lemma H.2} \\
&\leq \sup_{\alpha \in [n]} \mathbb{E} \left| \partial^{\mathcal{P}} \phi^i(g_\alpha^1, \ldots, g_\alpha^{i-1}; c^1, \ldots, c^{i-1}) \right|^p \leq B^{\phi^i}_{i,p,|\mathcal{P}|} && \text{applying Claim } 2(i-1).
\end{aligned}$$

□

Proposition J.8. *Consider any polynomially smooth $\psi : \mathbb{R}^{2\mathrm{j}} \to \mathbb{R}$ and any integers $p \geq 1$, $k \geq 0$. Recall that $\nabla^k \psi : \mathbb{R}^{2\mathrm{j}} \to \mathbb{R}^{(2\mathrm{j})^k}$ is the function that computes the tensor of ψ's kth-order partial derivatives. Then Claim 1$(1, \ldots, \mathrm{j})$ and Claim 2$(1, \ldots, \mathrm{j}-1)$ together imply the following L^p norm bound: There is a constant C, (j, p, k)-oblivious wrt ψ and the program, such that*

$$\sup_{\alpha \in [n]} \mathbb{E} \left\| \nabla^k \psi(g_\alpha^1, \ldots, g_\alpha^{\mathrm{j}}; c^1, \ldots, c^{\mathrm{j}}) \right\|_p^p \leq C$$

Note the form of this bound is very intuitive, since $\nabla^k \psi$ is just a polynomially bounded function, but taking values in a multi- but constant-dimensional space $\mathbb{R}^{(2\mathrm{j})^k}$ instead of $\mathbb{R}$. The proof is just routine manipulation using Lemma H.2 and applications of Claim 1 and Claim 2.

Proof. Let (C, q) be the the kth element of ψ's profile, i.e., such that for all input vectors $v \in \mathbb{R}^{2\mathrm{j}}$ and for all $\mathbf{U} \in [2\mathrm{j}]^k$,

$$|\partial_v^{\mathbf{U}} \psi(v)| \le C(1 + \|v\|_q^q). \tag{63}$$

Let $u \stackrel{\text{def}}{=} (g_\alpha^1, \ldots, g_\alpha^{\mathrm{j}}; c^1, \ldots, c^{\mathrm{j}}) \in \mathbb{R}^{2\mathrm{j}}$. Then

$$\|\nabla_u^k \psi(u)\|_p^p \le (2\mathrm{j})^k \sup_{\mathbf{U}} |\partial_u^{\mathbf{U}} \psi(u)|^p$$

where $\mathbf{U}$ ranges over all vectors in $[2\mathrm{j}]^k$. Thus it suffices to bound $\sup_{\alpha \in [n]} \mathbb{E} \sup_{\mathbf{U}} |\partial_u^{\mathbf{U}} \psi(u)|^p$ by a constant that depends on ψ only through C and q.

Applying Lemma H.1,

$$\mathbb{E} |\partial_u^{\mathbf{U}} \psi(u)|^p \le \mathbb{E} [C(1 + \|u\|_q^q)]^p \le C' \mathbb{E} (1 + \|u\|_{pq}^{pq})$$

where C' is a constant depending only on C, q, p, j continuously. Therefore it remains to show that $\mathbb{E} \|u\|_{pq}^{pq}$ is bounded by a constant independent of α. But, unwinding the definition of u,

$$\mathbb{E} \|u\|_{pq}^{pq} = \mathbb{E} |g_\alpha^1|^{pq} + \cdots + |g_\alpha^{\mathrm{j}}|^{pq} + |c^1|^{pq} + \cdots + |c^{\mathrm{j}}|^{pq}$$

By Claim $1(1, \ldots, \mathrm{j})$,

$$\sup_{\alpha \in [n]} \mathbb{E} |g_\alpha^i|^{pq} \le B_{i,pq,0}$$

For $i \le \mathrm{j}$, by Claim $2(i-1)$ and Proposition J.7, we have

$$\mathbb{E} |c^i|^{pq} \le B_{i-1,pq,0}^{\phi^i}$$

So $\mathbb{E} \|u\|_{pq}^{pq}$ is indeed bounded by a constant independent of α and (j, p, k)-oblivious. □

Proposition J.9. *Consider any integer $p \ge 0$, polynomially smooth $\psi : \mathbb{R}^{2\mathrm{j}} \to \mathbb{R}$ and any multisets $\mathcal{P}, \mathcal{U}$ of matrix entries. Let $z_\alpha \stackrel{\text{def}}{=} \partial^{\mathcal{P}} \psi(g_\alpha^1, \ldots, g_\alpha^{\mathrm{j}}; c^1, \ldots, c^{\mathrm{j}})$. Assume Claim 2$(\mathrm{j})$. Then for any multiset $\mathcal{N}$ taking values in $[n]$,*

$$\mathbb{E} |\partial^{\mathcal{U}} z^{\mathcal{N}}|^p \le C$$

for some constant C that is $(\mathrm{j}, p, \mathcal{N}, \mathcal{P}, \mathcal{U})$-oblivious wrt ψ and the program.

The statement of this bound is a bit more complicated than the previous ones, but again the content is intuitive. It says that some interleaved composition of 1) taking a constant number of partial derivatives and 2) taking a product over a constant number of "neuron index" $\alpha \in [n]$ will still result in an $O(1)$ quantity. The proof is again routine manipulation using Lemma H.2 and standard inequalities after applying the higher order product rule bound in Lemma H.11.

Proof. Let $l \stackrel{\text{def}}{=} |\mathcal{N}|$ and $v \in \mathbb{R}^l$ be the vector $(z_\alpha)_{\alpha \in \mathcal{N}}$ for an arbtirary ordering of $\mathcal{N}$ (including multiplicity). Thus $z^{\mathcal{N}} = \prod_j v_j$. Fix a ordering $\mathbf{U}$ of $\mathcal{U}$. By Lemma H.11 and Lemma H.2, there is a constant G depending only on $|\mathcal{U}|$ and p such that

$$|\partial^{\mathcal{U}} z^{\mathcal{N}}|^p \le G \sum_{t=0}^{l-1} \sqrt{l^{l-t-1} \|v\|_{2t}^{2tp}} \sum_\tau \sqrt{\prod_{\mathcal{Q}} \|\partial^{\mathcal{Q}} v\|^{2p}}$$

where τ ranges over partitions of $\{1, \ldots, |\mathcal{U}|\}$ of size $l - t$ and $\mathcal{Q}$ ranges over $\mathbf{U}[\tau]$ as in Lemma H.9. Here $\|v\|_0^0$ is by convention defined to be l. By Jensen's (concave) inequality, we then have

$$\mathbb{E} |\partial^{\mathcal{U}} z^{\mathcal{N}}|^p \le G \sum_{t=0}^{l-1} \sqrt{l^{l-t-1} \mathbb{E} \|v\|_{2t}^{2tp}} \sum_\tau \sqrt{\mathbb{E} \prod_{\mathcal{Q}} \|\partial^{\mathcal{Q}} v\|^{2p}}$$

Since the sizes of the ranges of t and of τ both depend only on $l = |\mathcal{N}|$ and $|\mathcal{U}|$, it suffices to show that both $\mathbb{E} \|v\|_{2t}^{2tp}$ and $\mathbb{E} \prod_{\mathcal{Q}} \|\partial^{\mathcal{Q}} v\|^{2p}$ have bounds that are $(\mathrm{j}, p, \mathcal{N}, \mathcal{P}, \mathcal{U})$-oblivious wrt ψ and the program.

Bounding $\mathbb{E}\,\|v\|_{2t}^{2t}$ Again, by Lemma H.2, $\mathbb{E}\,\|v\|_{2t}^{2tp} = O(\mathbb{E}\,\|v\|_{2tp}^{2tp})$, so it suffices to bound the latter. Now for all $t \in \{0, \dots, l-1\}$,

$$\mathbb{E}\,\|v\|_{2tp}^{2tp} = \sum_{\alpha \in \mathcal{N}} \mathbb{E}\,|z_\alpha|^{2tp} \le l \sup_{\alpha \in [n]} \mathbb{E}\,|z_\alpha|^{2tp} \le l B^{\psi}_{\mathsf{j},2tp,|\mathcal{P}|}$$

by Claim 2(j). This obviously satisfies the desired property.

Bounding $\mathbb{E}\prod_{\mathcal{Q}} \|\partial^{\mathcal{Q}} v\|^2$ Let $s \stackrel{\text{def}}{=} l - t$, so that the product $\prod_{\mathcal{Q}}$ iterates over s elements. Then by Hölder's Inequality,

$$\mathbb{E}\prod_{\mathcal{Q}} \|\partial^{\mathcal{Q}} v\|^{2p} \le \sqrt[s]{\prod_{\mathcal{Q}} \mathbb{E}\,\|\partial^{\mathcal{Q}} v\|^{2sp}}$$

By Lemma H.1, for a constant R depending on only l and s, we have

$$\begin{aligned}
\mathbb{E}\,\|\partial^{\mathcal{Q}} v\|^{2sp} &\le R\mathbb{E}\,\|\partial^{\mathcal{Q}} v\|_{2sp}^{2sp} \le lR \sup_{\alpha \in [n]} \mathbb{E}\,|\partial^{\mathcal{Q}} z_\alpha|^{2sp} \\
&= lR \sup_{\alpha \in [n]} \mathbb{E}\,|\partial^{\mathcal{Q}} \partial^{\mathcal{P}} \psi(g_\alpha^1, \dots, g_\alpha^{\mathsf{j}}; c^1, \dots, c^{\mathsf{j}})|^{2sp} \\
&\le lRB^{\psi}_{\mathsf{j},2sp,|\mathcal{Q}|+|\mathcal{P}|}
\end{aligned}$$

by Claim 2(j). From this, it's clear that $\mathbb{E}\prod_{\mathcal{Q}} \|\partial^{\mathcal{Q}} v\|^2$ has the desired property as well. □

J.2 Base Case: Claim 1$(1, \dots, M_0)$ and Claim 2$(1, \dots, M_0)$

Here we consider the case of $\mathsf{j} = 1, \dots, M_0$.

When $|\mathcal{P}| > 0$, Claim 1(j) and Claim 2(j) are trivially true since there's no dependence on the matrices A^i yet.

Now assume $|\mathcal{P}| = 0$. Then Claim 1(j) follows from standard Gaussian moment expressions. For Claim 2(j), we note $x \stackrel{\text{def}}{=} \psi(g^1, \dots, g^{\mathsf{j}}; c^1, \dots, c^{\mathsf{j}})$ has

$$\begin{aligned}
|x_\alpha| &\le C(1 + |g_\alpha^1|^q + \cdots + |g_\alpha^{\mathsf{j}}|^q + |c^1|^q + \cdots + |c^{\mathsf{j}}|^q) \\
&\le C(1 + |g_\alpha^1|^q + \cdots + |g_\alpha^{\mathsf{j}}|^q + \mathsf{j}R_q)
\end{aligned}$$

where C, q come from a profile of ψ and R_q is the qth moment bound on all the initial scalars. Then again we can apply standard Gaussian moment expressions to derive Claim 2(j).

J.3 Claim 1$(1, \dots, \mathsf{j})$ and Claim 2$(1, \dots, \mathsf{j}-1)$ Imply Claim 2(j)

Here we assume Claim 1$(1, \dots, \mathsf{j})$ and Claim 2$(1, \dots, \mathsf{j}-1)$ for $\mathsf{j} \ge M_0 + 1$ and derive Claim 2(j).

Let $u \stackrel{\text{def}}{=} (g_\alpha^1, \dots, g_\alpha^{\mathsf{j}}; c^1, \dots, c^{\mathsf{j}}) \in \mathbb{R}^{2\mathsf{j}}$ and let $\mathbf{P}$ be any ordering of $\mathcal{P}$. By Lemma H.10, we have

$$\partial^{\mathbf{P}} \psi(u) = \sum_{\tau} \|D^{|\tau|}\| \cdot \prod_{\mathcal{Q}} \|\partial^{\mathcal{Q}} u\|$$

where $D^k \stackrel{\text{def}}{=} \nabla_u^k \psi(u) \in \mathbb{R}^{(2\mathsf{j})^k}$, τ ranges over partitions of $\{1, \dots, |\mathcal{P}|\}$, and, for each τ, $\mathcal{Q}$ ranges over the elements (which are multisets) of the partition $\mathbf{P}[\tau] = \{\{P_i : i \in \mathcal{S}\} : \mathcal{S} \in \tau\}$ of $\tilde{\mathbf{P}}$.

Thus, applying Lemma H.2,

$$\begin{aligned}
|\partial^{\mathbf{P}} \psi(u)|^p &\le \left(\sum_{\tau} \|D^{|\tau|}\| \cdot \prod_{\mathcal{Q}} \|\partial^{\mathcal{Q}} u\| \right)^p \\
&\le G \sum_{\tau} \|D^{|\tau|}\|^p \cdot \prod_{\mathcal{Q}} \|\partial^{\mathcal{Q}} u\|^p
\end{aligned}$$

where G is the constant from Lemma H.2 that depends only on p and $|\mathcal{P}|$ (through the number of partitions of $\{1,\dots,|\mathcal{P}|\}$). Taking expectation and applying another Cauchy-Schwarz gives

$$\mathbb{E}\,|\partial^{\mathbf{P}}\psi(u)|^p \le G\sum_{\tau}\sqrt{\mathbb{E}\,\|D^{|\tau|}\|^{2p}}\cdot\sqrt{\mathbb{E}\prod_{\mathcal{Q}}\|\partial^{\mathcal{Q}}u\|^{2p}}$$

Since the number of partitions of $\{1,\dots,|\mathcal{P}|\}$ (which is the range of τ) depends only on $|\mathcal{P}|$, it suffices to prove that both $\mathbb{E}\,\|D^{|\tau|}\|^{2p}$ and $\mathbb{E}\prod_{\mathcal{Q}}\|\partial^{\mathcal{Q}}u\|^{2p}$ are bounded by constants that are $(\mathrm{j},p,\mathcal{P})$-oblivious wrt ψ and the program and are independent of α.

Bounding $\mathbb{E}\,\|D^{|\tau|}\|^{2p}$ This follows directly from Proposition J.8 (which is a straightforward bound by replacing each partial derivative of ψ with its polynomial upper bound).

Bounding $\mathbb{E}\prod_{\mathcal{Q}}\|\partial^{\mathcal{Q}}u\|^{2p}$ First, note this quantity clearly does not depend on ψ.

Let $k=|\tau|$. Recall $\mathcal{Q}$ ranges over the k multisets $\{P_i : i\in\mathcal{S}\}$ as $\mathcal{S}$ ranges over elements of the partition τ. By Hölder's Inequality and Lemma H.1,

$$\mathbb{E}\prod_{\mathcal{Q}}\|\partial^{\mathcal{Q}}u\|^{2p} \le \prod_{\mathcal{Q}}\sqrt[k]{\mathbb{E}\,\|\partial^{\mathcal{Q}}u\|^{2pk}} \le R\prod_{\mathcal{Q}}\sqrt[k]{\mathbb{E}\,\|\partial^{\mathcal{Q}}u\|_{2pk}^{2pk}}$$

where R is a constant depending only on $2pk$ and j (coming from Lemma H.1). Since $|\mathcal{Q}|$ and $k=|\tau|$ are both bounded by $|\mathbf{P}|$, it suffices to show $\mathbb{E}\,\|\partial^{\mathcal{Q}}u\|^{2pk}$ is bounded by a constant independent of α and is $(\mathrm{j},p,k,\mathcal{Q})$-oblivious wrt ψ and the program. Now,

$$\begin{aligned}\|\partial^{\mathcal{Q}}u\|_{2pk}^{2pk} &= \sum_{i=1}^{\mathrm{j}}(\partial^{\mathcal{Q}}g_\alpha^i)^{2pk}+(\partial^{\mathcal{Q}}c^i)^{2pk}\\ &\le \sum_{i=1}^{\mathrm{j}} B_{i,2pk,|\mathcal{Q}|}+B^{\phi^i}_{i-1,2pk,|\mathcal{Q}|}\end{aligned}$$

by Claim $1(1,\dots,\mathrm{j})$ and Proposition J.7. This bound indeed is independent of α and has the required oblivious property.

J.4 Claim $2(\mathrm{j}-1)$ Implies Claim $1(\mathrm{j})$

Assume $\mathrm{j}\ge M_0+1$. Let $y_\beta \overset{\text{def}}{=} \phi^{\mathrm{j}}(g_\beta^1,\dots,g_\beta^{\mathrm{j}-1};c^1,\dots,c^{\mathrm{j}-1})$. Recall $g_\alpha^{\mathrm{j}}=\sum_{\beta=1}^n W^{\mathrm{j}}_{\alpha\beta}y_\beta$, where W^{j} is one of the program's matrices A^i or their transposes.

Reduction via Product Rule Then by the product rule of differentiation,

$$\partial^{\mathcal{P}}g_\alpha^{\mathrm{j}} = \left(\sum_{\beta=1}^n W^{\mathrm{j}}_{\alpha\beta}\partial^{\mathcal{P}}y_\beta\right)+remainder$$

where $remainder$ is a sum of at most $|\mathcal{P}|$ elements of the form $\partial^{\mathcal{P}'}y_\beta$ where $\mathcal{P}'$ is $\mathcal{P}$ with some element removed. Therefore, by Lemma H.2,

$$|\partial^{\mathcal{P}}g_\alpha^{\mathrm{j}}|^p \le R\left[\left|\sum_{\beta=1}^n W^{\mathrm{j}}_{\alpha\beta}\partial^{\mathcal{P}}y_\beta\right|^p+|remainder|^p\right]$$

where R depends only on p. We can easily bound $\mathbb{E}\,|remainder|^p$ by a $(\mathrm{j},p,\mathcal{P})$-oblivious constant independent of α using Claim $2(\mathrm{j}-1)$ and Lemma H.2. Thus, it suffices to bound $\left|\sum_{\beta=1}^n W^{\mathrm{j}}_{\alpha\beta}\partial^{\mathcal{P}}y_\beta\right|^p$ by a constant with the same property.

Plan: Show Small Moment Condition (SMC) Holds To do so, we will show the vector with entries $w_\beta z_\beta$ where $w_\beta \stackrel{\text{def}}{=} W^{\mathrm{j}}_{\alpha\beta}$ and $z_\beta \stackrel{\text{def}}{=} \partial^{\mathcal{P}} y_\beta$ satisfies the Small Moment Condition (Definition I.3) and apply Lemma I.4 to it. In particular, we will assume WLOG that p is an even integer and prove that for every multiset $\mathcal{N}$ of $[n]$ of size $|\mathcal{N}| = p$, we have

$$\mathbb{E}\, w^{\mathcal{N}} z^{\mathcal{N}} \le C n^{-|\mathrm{uniq}(\mathcal{N})|} \tag{64}$$

for a constant C that is $(\mathrm{j}, \mathcal{N}, \mathcal{P})$-oblivious, where $\mathrm{uniq}(\mathcal{N})$ is the set of unique elements of $\mathcal{N}$.

Taylor Expansion of $z^{\mathcal{N}}$ Fix $\mathcal{N}$. We now consider $z^{\mathcal{N}}$ as a function of w_β for unique elements $\beta \in \mathrm{uniq}(\mathcal{N})$ of $\mathcal{N}$ (keeping other weight entries fixed). Let $v = (w_\beta)_{\beta \in \mathrm{uniq}(\mathcal{N})} \in \mathbb{R}^{|\mathrm{uniq}(\mathcal{N})|}$ be the vector of such elements, so that we write $z^{\mathcal{N}} = z^{\mathcal{N}}(v)$ as function of v. By Lemma H.8, we Taylor expand $z^{\mathcal{N}}$ to the rth order, for some r to be determined later:

$$\begin{aligned} z^{\mathcal{N}} &= z^{\mathcal{N}}(0) + \langle \nabla z^{\mathcal{N}}(0), v\rangle + \cdots + \frac{1}{r!}\langle \nabla^r z^{\mathcal{N}}(0), v^{\otimes r}\rangle + R_{r+1} \\ &= R_{r+1} + \sum_{s=0}^{r} \frac{1}{s!}\langle \nabla^s z^{\mathcal{N}}(0), v^{\otimes s}\rangle \\ \text{where}\quad R_{r+1} &\stackrel{\text{def}}{=} \frac{1}{r!}\int_0^1 (1-t)^r \langle \nabla^{r+1} z^{\mathcal{N}}(tv), v^{\otimes (r+1)}\rangle dt \end{aligned}$$

While the tensors appearing in this expansion may seem at first like "large objects", note that in terms of n, the tensors have constant sizes, so their norms are entirely determined by how their entries scale with n. Intuitively, the derivative tensors $\nabla^s z^{\mathcal{N}}$ will have $O(1)$ entry sizes, by induction hypothesis, while $v^{\otimes s}$ has size $O(n^{-s/2})$. But before we keep following this logic of naive bounds, it pays to notice there are a lot of cancellation.

Cancellation Using Independence and Zero-Mean Because now $\nabla^s z^{\mathcal{N}}(0)$ no longer depends on and thus is independent (as a random variable) from v (and thus $w^{\mathcal{N}}$)[26], taking expectation we now have

$$\mathbb{E}\, w^{\mathcal{N}} z^{\mathcal{N}} = \mathbb{E}\, w^{\mathcal{N}} R_{r+1} + \sum_{s=0}^{r} \frac{1}{s!}\langle \mathbb{E}\, \nabla^s z^{\mathcal{N}}(0), \mathbb{E}\, w^{\mathcal{N}} v^{\otimes s}\rangle$$

where $w^{\mathcal{N}} v^{\otimes s}$ is the scalar multiplication of the scalar $w^{\mathcal{N}}$ with the tensor $v^{\otimes s}$.

Now suppose

$$\mathcal{N} \text{ has exactly } k \text{ elements that appear singly (i.e., have multiplicity 1) in } \mathcal{N}.$$

Then $\mathbb{E}\, w^{\mathcal{N}} v^{\otimes s}$ will be 0 for all $s < k$: indeed, every entry of $w^{\mathcal{N}} v^{\otimes s}$ is a monomial $w^{\mathcal{N}'}$ that will have some w_β appearing by itself in the product (i.e., has degree 1), so that

$$\mathbb{E}\, w^{\mathcal{N}'} = \mathbb{E}\, w_\beta \mathbb{E}\, w^{\mathcal{N}'\setminus\{\beta\}} = 0$$

using the fact that w_β is zero-mean and independent from $w^{\mathcal{N}'\setminus\{\beta\}}$.

Therefore, we will take r (the order of the Taylor expansion) to be $k-1$, so that

$$\mathbb{E}\, w^{\mathcal{N}} z^{\mathcal{N}} = \mathbb{E}\, w^{\mathcal{N}} R_{k+1}.$$

Unwinding the definition of R_{k+1} and using Cauchy-Schwarz, we have

$$\mathbb{E}\, w^{\mathcal{N}} z^{\mathcal{N}} \le \frac{1}{(k-1)!}\int_0^1 (1-t)^{k-1} \sqrt{\mathbb{E}\,\|\nabla^k z^{\mathcal{N}}(tv)\|^2 \cdot \mathbb{E}\,\|w^{\mathcal{N}} v^{\otimes k}\|^2}\, dt \tag{65}$$

[26] this was the main purpose of the Taylor expansion

Constructing the SMC Constant Let $\mathcal{N}_1$ be the subset of elements of $\mathcal{N}$ with multiplicity 1 (so that $|\mathcal{N}_1| = k$) and let $\mathcal{N}' = \mathcal{N} \setminus \mathcal{N}_1$. We will show that

$$\mathbb{E}\, w^{\mathcal{N}} z^{\mathcal{N}} \leq B_k n^{-|\mathrm{uniq}(\mathcal{N})|} \tag{66}$$

for some constant B_k that is $(\mathrm{j}, \mathcal{N}_1, \mathcal{N}', \mathcal{P})$-oblivious. In particular, this means B_k depends on $\mathcal{N}$ only through $|\mathcal{N}| = |\mathcal{N}_1| + |\mathcal{N}'| = p$ and $|\mathcal{N}_1| = k$. Then the constant C in Eq. (64) can be taken as $\max_k B_k$ where k ranges from 0 to p.

In light of Eq. (65), to prove Eq. (66), it thus suffices to show that

$$\mathbb{E}\, \|\nabla^k z^{\mathcal{N}}(tv)\|^2 = O(1)$$
$$\mathbb{E}\, \|w^{\mathcal{N}} v^{\otimes k}\|^2 = O(n^{-2|\mathrm{uniq}(\mathcal{N})|})$$

where the big-Os hide $(\mathrm{j}, \mathcal{N}_1, \mathcal{N}', \mathcal{P})$-oblivious constants that furthermore are independent of $t \in [0, 1]$.

Bounding $\mathbb{E}\, \|w^{\mathcal{N}} v^{\otimes k}\|^2$ Each entry of the tensor $w^{\mathcal{N}} v^{\otimes k}$ is just a product of $|\mathcal{N}| + k = p + k$ matrix entries, whose expected square norm can be bounded by $\nu_{2(k+p)} n^{-(k+p)}$, where $\nu_{2(k+p)}$ is the scaled moment bound on the matrix entries we fixed at the beginning of this proof. There are $|\mathrm{uniq}(\mathcal{N})|^k \leq p^k$ entries in this tensor, so

$$\mathbb{E}\, \|w^{\mathcal{N}} v^{\otimes k}\|^2 \leq p^k \nu_{2(k+p)} n^{-(k+p)}$$

Now note that, by the definition of k, we have $k + p \geq 2|\mathrm{uniq}(\mathcal{N})|$. Thus

$$\mathbb{E}\, \|w^{\mathcal{N}} v^{\otimes k}\|^2 = O(n^{-2|\mathrm{uniq}(\mathcal{N})|})$$

where the constant in $O(-)$ is $(p, \mathcal{N}_1, \mathcal{N}', \mathcal{P})$-oblivious.

Bounding $\mathbb{E}\, \|\nabla^k z^{\mathcal{N}}(tv)\|^2$. Recall that all bounds in this proof (Appendix J) are oblivious wrt the program, so they only depend on the program through the bounds $(\nu_q)_q, (R_q)_q, \mathcal{C}$ fixed at the beginning of this proof. Notice that $\nabla^k z^{\mathcal{N}}(tv)$ is just $\nabla^k z^{\mathcal{N}}$ computed in the program where the matrix entries $\{W^{\mathsf{j}}_{\alpha\beta} : \beta \in \mathcal{N}\}$ (which are the entries of v) are scaled *down* and such a program satisfies the exact same data (in particular the moment bounds given by $(\nu_q)_q$).

Thus, by Proposition J.9, *every* entry of $\nabla^k z^{\mathcal{N}}(tv)$ has the same $(\mathrm{j}, \mathcal{N}_1, \mathcal{N}', \mathcal{P})$-oblivious bound C on its expected square norm, *uniformly* over $t \in [0, 1]$. Because $\nabla^k z^{\mathcal{N}}(tv)$ has $|\mathrm{uniq}(\mathcal{N})|^k \leq p^k$ entries,

$$\mathbb{E}\, \|\nabla^k z^{\mathcal{N}}(tv)\|^2 \leq C p^k$$

which is $(\mathrm{j}, \mathcal{N}_1, \mathcal{N}', \mathcal{P})$-oblivious and independent of $t \in [0, 1]$, as desired.

K Program Transformations

K.1 Backpropagation Program

Given a program T as in Eq. (4) and a polynomially smooth function $\psi : \mathbb{R}^{2M} \to \mathbb{R}$, we can create a new program T_ψ that extends T, which we call the *backpropagation program of T with respect to ψ*, or just *backprogram* for short. Intuitively, T_ψ will compute the gradients of

$$c \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\alpha=1}^{n} \psi(g^1_\alpha, \ldots, g^M_\alpha; c^1, \ldots, c^M) \tag{67}$$

with respect to all vectors in the program. In the context of this work, the importance of the backprogram construction is to easily express the partial derivative $\delta = \frac{\partial c}{\partial A^j_{\alpha\beta}}$ (Proposition K.2), which easily shows that $\delta = O(n^{-1})$ *instead of* $O(1)$ *as suggested by Proposition J.7* (see Lemma K.6). This will be crucial in the proof of Theorem 5.2.

Explicitly, T_ψ is constructed as follows. It has the same matrices A^i as T. The first M vectors and scalars g^i and c^i for $i = 1, \ldots, M$ are the same as in T. It additionally has new vectors and scalars constructed after them. We will first describe the *mathematical objects* that will be computed by these new vectors and scalars, before discussing how to represent them in the form of Eq. (4).

Notation In this context, we use Sans Serif font to represent these mathematical objects, to distinguish them from objects in the program itself. Recall that ϕ^i is the nonlinearity used in iteration i in the original program T (as well as in the new program T_ψ). For $M \geq j > i$, we will write $\mathsf{x}^j_{i;} \in \mathbb{R}^n$ and $\mathsf{x}^j_{;i} \in \mathbb{R}^n$ for the vectors with entries

$$(\mathsf{x}^j_{i;})_\alpha \stackrel{\text{def}}{=} \partial_{g^i} \phi^j(g^1, \ldots, g^{j-1}; c^1, \ldots, c^{j-1})$$
$$(\mathsf{x}^j_{;i})_\alpha \stackrel{\text{def}}{=} \partial_{c^i} \phi^j(g^1, \ldots, g^{j-1}; c^1, \ldots, c^{j-1}).$$

In other words, $\mathsf{x}^j_{i;}$ is the partial derivative $\partial_{g^i}\phi^j$ with respect to the argument g^i and likewise $\mathsf{x}^j_{;i}$ is the partial derivative $\partial_{c^i}\phi^j$ with respect to the argument c^i, both evaluated on the original set of inputs for ϕ^j. For convenience, we will also write $\mathsf{x}^{M+1}_{i;}$ and $\mathsf{x}^{M+1}_{;i}$ for the partial derivatives of ψ, i.e.,

$$(\mathsf{x}^{M+1}_{i;})_\alpha \stackrel{\text{def}}{=} \partial_{g^i} \psi(g^1, \ldots, g^M; c^1, \ldots, c^M)$$
$$(\mathsf{x}^{M+1}_{;i})_\alpha \stackrel{\text{def}}{=} \partial_{c^i} \psi(g^1, \ldots, g^M; c^1, \ldots, c^M).$$

In addition, for a given vector $v \in \mathbb{R}^n$, we write $\langle v \rangle \stackrel{\text{def}}{=} \frac{1}{n}\sum_{\alpha=1}^n v_\alpha$ for the average of the entries of v.

Mathematical Idea We iteratively construct the vectors $\mathsf{dx}^i, \mathsf{dg}^i \in \mathbb{R}^n$ for decreasing $i = M, M-1, \ldots, M_0+1$ as follows.

$$\mathsf{dx}^{M+1} \stackrel{\text{def}}{=} 1 \in \mathbb{R}^n$$
$$\mathsf{dg}^i \stackrel{\text{def}}{=} \sum_{k=i+1}^{M+1} \mathsf{x}^k_{i;} \odot \mathsf{dx}^k + \sum_{M+1 \geq j > k > i} \langle \mathsf{dx}^j \odot \mathsf{x}^j_{;k} \rangle x^k_{i;} \tag{68}$$
$$\mathsf{dx}^i \stackrel{\text{def}}{=} W^{i\top} \mathsf{dg}^i \tag{69}$$

where $W^{i\top}$ in the last line is the transpose of the same matrix used in iteration i of the original program T, as in Eq. (4). For example, the first few dg^i and dx^i looks like the following.

$$\begin{aligned}
\mathsf{dx}^{M+1} &= 1 \\
\mathsf{dg}^M &= \mathsf{x}^{M+1}_{M;} \\
\mathsf{dx}^M &= W^{M\top} \mathsf{dg}^M \\
\mathsf{dg}^{M-1} &= \mathsf{x}^{M+1}_{M-1;} + \mathsf{x}^M_{M-1;} \odot \mathsf{dx}^M + \langle \mathsf{x}^{M+1}_{;M} \rangle \mathsf{x}^M_{M-1;} \\
\mathsf{dx}^{M-1} &= W^{M-1\top} \mathsf{dg}^{M-1}
\end{aligned}$$

One can verify the following statement using the chain rule.

Proposition K.1. *For $i = M, M-1, \ldots, M_0+1$, the vector dx^i (Eq. (69)) equals the scaled total gradient $n\nabla_{x^i} c$ of c (scaled up by n) against the vector x^i defined in Eq. (4), and likewise dg^i (Eq. (68)) equals $n\nabla_{g^i} c$.*

The vectors dx^i and dg^i make it easy to express the partial derivative $\frac{\partial c}{\partial A^i_{\alpha\beta}}$ as well:

Proposition K.2. *For any matrix A^j of program T, for ψ, c as in Eq. (67) and dg^i as in Eq. (68), we have*

$$\frac{\partial c}{\partial A^j_{\alpha\beta}} = \sum_{i: W^i = A^j} \frac{1}{n} (\mathsf{dg}^i)_\alpha (x^i)_\beta + \sum_{i: W^i = A^{j\top}} \frac{1}{n} (\mathsf{dg}^i)_\beta (x^i)_\alpha,$$

where in the first sum we iterate over indices i such that $W^i = A^j$ and in the second, i such that $W^i = A^{j\top}$, both in the context of the original program T.

Proof. Each A^j is used in the computation of c only through any W^i that equals A^j or its transpose. Thus

$$\frac{\partial c}{\partial A^j_{\alpha\beta}} = \sum_{i: W^i = A^j} \frac{\partial c}{\partial W^i_{\alpha\beta}} + \sum_{i: W^i = A^{j\top}} \frac{\partial c}{\partial W^i_{\beta\alpha}}$$

But, from Proposition K.1 and chain rule, one easily calculates

$$\frac{\partial c}{\partial W^i_{\alpha\beta}} = \frac{1}{n}(\mathsf{dg}^i)_\alpha (x^i)_\beta.$$

□

Program T_ψ Construction Here we will construct the vectors g^a for $a = M+1, M+2, \ldots$ for the new program T_ψ. To avoid confusion with vectors in the original program, we use superscript index a, b, e for talking about the new vectors, while i, j, k are reserved for indices of the old program. Our goal is to express each dx^i as some g^a with dg^i being the corresponding x^a (c.f. Eq. (4)). To do so, as is apparent from Eq. (68), we need to express $\langle \mathsf{dx}^j \odot \mathsf{x}^j_{:k} \rangle$ for each $M+1 \geq j > k$ as scalars c^b as well.

So the strategy is to, in descending order of i, alternatingly express dx^i, then the scalars $\langle \mathsf{dx}^j \odot \mathsf{x}^j_{:k} \rangle$ for $M+1 \geq j > k > i-1$, then express dg^{i-1} through Eq. (68) and finally dx^{i-1} through Eq. (69), and repeat. At the end, the vectors of T_ψ contain $g^1, \ldots, g^M$ (same as in T) and $\mathsf{dx}^{M+1} \ldots, \mathsf{dx}^{M_0+1}$ (new vectors) and the scalars contain $c^1, \ldots, c^M$ (same as in T) and $\{\langle \mathsf{dx}^j \odot \mathsf{x}^j_{:k} \rangle\}_{M+1 \geq j > k > M_0}$ (new scalars). There are also other "junk" vectors (resp. scalars) that we don't care about, which arise when we only want to express some scalar (resp. vectors). This shows

Proposition K.3. *If T has $M - M_0$ iterations, then T_ψ has at most $(M - M_0)^2$ iterations.*

Combining this with the fact that the nonlinearities of T_ψ are just compositions of polynomials with first order derivatives of those of T, we deduce the following

Proposition K.4. *Any $(\mathcal{P}_1, \ldots, \mathcal{P}_r)$-oblivious constant wrt T_ψ and $\psi^1, \ldots, \psi^l$ is also $(\mathcal{P}_1, \ldots, \mathcal{P}_r)$-oblivious wrt T and $\psi^1, \ldots, \psi^l$.*

It's clear from Eq. (68) that dg^i can be expressed as some nonlinearity applied to vectors of T_ψ. Beyond this fact, the exact construction of T_ψ is not important for our purposes, so we will not detail it further here.

Proposition K.5. *For any program T and c, ψ as in Eq.* (67)*, each $\mathsf{dg}^i = n\nabla_{g^i} c$ can be expressed as some nonlinearity ϕ applied to the vectors and scalars of T_ψ. If all nonlinearities of T are polynomially smooth, then ϕ is polynomially smooth as well.*

Matrix Derivative Bound

Lemma K.6. *Consider a program in Setup J.2. Then for any $p \geq 1$, any nonempty multiset $\mathcal{P}$ taking values in the program's matrix entries $\{A^i_{\alpha\beta}\}_{\alpha,\beta,i}$, and any scalar c of the program,*

$$\mathbb{E}\left|\partial^{\mathcal{P}} c\right|^p = O(n^{-p}) \quad \text{as } n \to \infty. \tag{70}$$

Furthermore, the constant in the big-O is $(p, \mathcal{P})$-oblivious wrt the program.

Note importantly that $\mathcal{P}$ has to be nonempty for this to hold, since without taking derivatives, c can definitely be $\Theta(1)$ (for example, if its limit $\mathring{c}$ is nonzero). This lemma improves on Lemma J.6 drastically when $\mathcal{P}$ is nonempty. The reason that Lemma J.6 is so loose in such cases is that, in the sum over $\alpha \in [n]$ in Eq. (62), only a constant number of α really achieves the sup bound in Lemma J.3. So the naive way Lemma J.6 converts the sup bound to an average bound turned out to leave a lot of room.

Proof. Suppose c is the ith scalar. For $i \leq M_0$, the derivative is 0, so consider the case where $i > M_0$. For brevity, write ψ for ϕ^i (the nonlinearity used to create c^i). Construct the backprogram T_ψ with respect to ψ. By Proposition K.4, it suffices to show the bound for a constant $(p, \mathcal{P})$-oblivious wrt T_ψ.

Since $\mathcal{P}$ is nonempty, there is an element a of $\mathcal{P}$. Write $\mathcal{P}' = \mathcal{P} \setminus \{a\}$. Then by Proposition K.2,

$$\partial^{\mathcal{P}} c = \partial^{\mathcal{P}'}(\partial_a c) = \frac{1}{n} \sum \partial^{\mathcal{P}'}[(\mathsf{dg}^j)_\alpha (x^j)_\beta],$$

where the sum is over some collection of (j, α, β) of size at most $M - M_0$ (an upper bound on how many times a has been used in the original program).

Then a routine combination of Lemma H.2, product rule (Lemma H.11), and Cauchy-Schwarz reduces our problem to bounding $\mathbb{E}\,|\partial^{\mathcal{Q}}(\mathsf{dg}^j)_\alpha|^{2p}$ and $\mathbb{E}\,|\partial^{\mathcal{Q}}(x^j)_\beta|^{2p}$ for all subsets $\mathcal{Q}$ of $\mathcal{P}'$ by a $(p, \mathcal{Q})$-oblivious constant independent of α and β. This is precisely provided by Lemma J.3 applied to the backprogram T_ψ. □

Lemma K.7. *Consider a program in Setup J.2. Let $\mathcal{Q}_1, \dots, \mathcal{Q}_r$ be nonempty and let $\mathcal{P}$ be potentially empty multisets of matrix entries. Then for any scalar c of the program,*

$$\mathbb{E}\left|\partial^{\mathcal{P}} \prod_{i=1}^{r} \partial^{\mathcal{Q}_i} c\right|^p = O(n^{-rp})$$

where the hidden constant is $(p, \mathcal{Q}_1, \dots, \mathcal{Q}_r, \mathcal{P})$-oblivious wrt the program.

Proof. The product rule (Lemma H.11), Lemma H.2, and Hölder's Inequality show that the LHS is, within a constant factor depending only on p and $|\mathcal{P}|$, bounded by a sum of $O(1)$ number (depending only on p) of terms, each of which is a product over r elements of the form

$$\mathbb{E}\,|\partial^{\mathcal{Q}} c|^{rp}$$

for multisets $\mathcal{Q}$ with size at most $|\mathcal{P}| + \max_i |\mathcal{Q}_i|$. Then the desired result follows from Lemma K.6.

□

L Proof of Theorem 5.2

We will prove the following more specific version of Theorem 5.2, which says more about the hidden constant. Recall the dot derivative notation from Section 5 as well as the notion of *oblivious constants* in Definition J.1.

Theorem L.1. *Consider a program in Setup 3.6 and its interpolation (Definition 5.1), and let c be a scalar in it. Then for any finite $p \geq 1$,*

$$\sup_t \mathbb{E}\,\dot{c}(t)^p = O(n^{-p/2})$$

where the hidden constant is p-oblivious wrt the program.

Proof. By the power-mean inequality, it suffices to show this for any even integer p. Let $a = a(t)$ be the vector of all matrix entries in the program, which is an interpolation of the Gaussian and non-Gaussian matrix entries. Let N be its dimension (which is equal to n^2 times the number of matrices in the program). We will use κ to index a's entries. Denote its derivative against t by $\dot{a} \in \mathbb{R}^N$, as in Section 5. For brevity, we will suppress the argument (t) below.

Let $D \stackrel{\text{def}}{=} \nabla_a c \in \mathbb{R}^N$, so that for any multiset $\mathcal{P}$ taking values in $[N]$, $D^{\mathcal{P}} = \prod_{\kappa \in \mathcal{P}} \frac{\partial c}{\partial a_\kappa}$. By chain rule, we have

$$\dot{c} = \langle D, \dot{a} \rangle = \langle 1, D \odot \dot{a} \rangle.$$

From here on, the proof follows an outline similar to that of Appendix J.4, except in how the cancellation happens.

Proof Plan: Small Moment Condition. We will show that the N-dimensional vector $\sqrt{n} D \odot \dot{a}$ (which has entries $\sqrt{n} D_\kappa \dot{a}_\kappa$) satisfies the Small Moment Condition (Definition I.3): For any even $p \geq 0$ and any multiset $\mathcal{P}$ of size p taking values in $[N]$, we need to show there's a $\mathcal{P}$-oblivious constant C wrt the program that is also independent of t such that

$$n^{p/2} D^{\mathcal{P}} \dot{a}^{\mathcal{P}} \leq C N^{-|\mathrm{uniq}(\mathcal{P})|}. \tag{71}$$

Then by Lemma I.4, we have

$$n^{p/2} \mathbb{E}\,|\dot{c}|^p = \mathbb{E}\,|\langle 1, \sqrt{n} D \odot \dot{a} \rangle|^p = O(1)$$

where the hidden constant is independent of $t \in [0, 1]$, which yields Theorem L.1.

Taylor Expansion. Now fix $\mathcal{P}$. Let $v \stackrel{\text{def}}{=} (a_\kappa)_{\kappa \in \text{uniq}(\mathcal{P})}$ be the vector of unique a_κ for $\kappa \in \mathcal{P}$ (for any ordering of $\text{uniq}(\mathcal{P})$). We think of $D^{\mathcal{P}}$ as a function $D^{\mathcal{P}}(v)$ of v. For an integer r to be specified later, we Taylor expand $D^{\mathcal{P}}$ in v around 0 to the rth order (Lemma H.8): with ∇^i denoting differentation wrt v,

$$D^{\mathcal{P}}(v) = D^{\mathcal{P}}(0) + \langle \nabla D^{\mathcal{P}}(0), v\rangle + \cdots + \frac{1}{r!}\langle \nabla^r D^{\mathcal{P}}(0), v^{\otimes r}\rangle + R_{r+1}$$
$$\text{where} \quad R_{r+1} \stackrel{\text{def}}{=} \frac{1}{r!}\int_0^1 (1-\zeta)^r \langle \nabla^{r+1} D^{\mathcal{P}}(\zeta v), v^{\otimes (r+1)}\rangle d\zeta.$$

Cancellation Using Independence and Zero-Mean Now notice $\nabla^i D^{\mathcal{P}}(0)$ is independent from $v^{\otimes i}$ and $\dot{a}^{\mathcal{P}}$. Therefore, for $i = 0, \ldots, r$,

$$\mathbb{E}\,\langle \nabla^i D^{\mathcal{P}}(0), \dot{a}^{\mathcal{P}} v^{\otimes i}\rangle = \langle \mathbb{E}\,\nabla^i D^{\mathcal{P}}(0), \mathbb{E}\,\dot{a}^{\mathcal{P}} v^{\otimes i}\rangle.$$

Furthermore, notice $\mathbb{E}\,\dot{a}^{\mathcal{P}} v^{\otimes i} = 0$ for small values of i. Indeed, if

$$k \text{ is the number of elements in } \mathcal{P} \text{ appearing exactly once}$$

and $i < 2k$, then every entry of the expected tensor $\mathbb{E}\,\dot{a}^{\mathcal{P}} v^{\otimes i}$ has the form

$$(\mathbb{E}\,\dot{a}_\kappa)\cdot(other) \quad \text{or} \quad (\mathbb{E}\,\dot{a}_\kappa a_\kappa)\cdot(other)$$

for some a_κ where $other$ is the expectation of some other monomial in a and $\dot{a}$ independent from a_κ and $\dot{a}_\kappa$. Both cases evaluate to 0 by Lemma 5.3. Thus, we now choose r (the degree of Taylor expansion) to be $2k-1$. Then

$$\mathbb{E}\,\dot{a}^{\mathcal{P}} D^{\mathcal{P}}(v) = \mathbb{E}\,\dot{a}^{\mathcal{P}} R_{2k}.$$

By Lemma H.2 and Cauchy-Schwarz, unwinding the definition of R_{2k}, we now have

$$\mathbb{E}\,\dot{a}^{\mathcal{P}} D^{\mathcal{P}}(v) \le \frac{1}{(2k-1)!}\int_0^1 (1-\zeta)^{2k-1}\sqrt{\mathbb{E}\,\|\nabla^{2k} D^{\mathcal{P}}(\zeta v)\|^2 \cdot \mathbb{E}\,\|\dot{a}^{\mathcal{P}} v^{\otimes 2k}\|^2}\, d\zeta \tag{72}$$

Let $\mathcal{P}_1$ be the subset of $\mathcal{P}$'s elements appearing in $\mathcal{P}$ exactly once, so that $|\mathcal{P}_1| = k$. Let $\mathcal{P}' = \mathcal{P} \setminus \mathcal{P}_1$. So it suffices to show

$$\sqrt{\mathbb{E}\,\|\dot{a}^{\mathcal{P}} v^{\otimes 2k}\|^2} = O(n^{-(2k+p)/2})$$
$$\sqrt{\mathbb{E}\,\|\nabla^{2k} D^{\mathcal{P}}(\zeta v)\|^2} = O(n^{-p})$$

where the hidden constants a) are $(\mathcal{P}_1, \mathcal{P}')$-oblivious and b) are independent of t (the interpolation variable) and ζ (the integration variable in the Taylor remainder R_{2k}). If these are shown, then, plugging into Eq. (72),

$$n^{p/2}\mathbb{E}\,\dot{a}^{\mathcal{P}} D^{\mathcal{P}}(v) = O(n^{-k-p}) = O(n^{-2|\text{uniq}(\mathcal{P})|}) = O(N^{-|\text{uniq}(\mathcal{P})|})$$

where the second equality follows because k is the number of elements of $\mathcal{P}$ with multiplicity 1 and $p = |\mathcal{P}|$, thus $k + p \ge 2|\text{uniq}(\mathcal{P})|$. The hidden constant here will depend on $k = |\mathcal{P}_1|$, but we can take the max of such constants over all $k = 0, 1, \ldots, p$ to arrive at the desired $\mathcal{P}$-oblivious constant in Eq. (71).

Bounding $\dot{a}^{\mathcal{P}} v^{\otimes 2k}$. Each entry of $\dot{a}^{\mathcal{P}} v^{\otimes 2k}$ is a degree $2k + |\mathcal{P}| = 2k + p$ monomial in $\dot{a}$ and a. Thus, by Lemma 5.3, its expected square norm is bounded by $O(n^{-(2k+p)})$ uniformly over all entries, where the hidden constant depends only on the scaled moment bound ν_{2k+p} of the program's matrix entries. Because $\dot{a}^{\mathcal{P}} v^{\otimes k}$ has $|\text{uniq}(\mathcal{P})|^{2k} \le p^{2k} = O(1)$ entries, we have

$$\mathbb{E}\,\|\dot{a}^{\mathcal{P}} v^{\otimes 2k}\|^2 = O(n^{-(2k+p)})$$

where the hidden constant is $(\mathcal{P}_1, \mathcal{P}')$-oblivious, as desired.

Bounding $\nabla^{2k} D^{\mathcal{P}}(\zeta v)$. We can think of $\nabla^{2k} D^{\mathcal{P}}(\zeta v)$ as $\nabla^{2k} D^{\mathcal{P}}$ computed on a program where the matrix entries in v are multiplied by a factor of $0 \le \zeta \le 1$. This program would then satisfy the same scaled moment bounds $(\nu_j)_{j=2}^{\infty}$ as the original program, uniformly over all t. Thus all of our oblivious bounds would apply to $\nabla^{2k} D^{\mathcal{P}}(\zeta v)$. In particular, Lemma K.7 tells us each entry of $\nabla^{2k} D^{\mathcal{P}}(\zeta v)$ has expected square norm bounded uniformly (over all entries) by $O(n^{-2|\mathcal{P}|}) = O(n^{-2p})$ with a $(\mathcal{P}_1, \mathcal{P}')$-oblivious constant independent of ζ and t. Finally, because there are $|\mathrm{uniq}(\mathcal{P})|^{2k} \le p^{2k} = O(1)$ entries, we get

$$\mathbb{E} \|\nabla^{2k} D^{\mathcal{P}}(\zeta v)\|^2 = O(n^{-2p})$$

$(\mathcal{P}_1, \mathcal{P}')$-obliviously and uniformly over ζ and t as desired.

□

M L^p Convergence From Almost Sure Convergence

The following is a standard lemma.

Lemma M.1. *For any nonnegative random variable $X \in \mathbb{R}$ and deterministic $B \ge 0$, we have*

$$\mathbb{E}\, X \mathbb{I}(X > B) = B \Pr[X > B] + \int_B^{\infty} \Pr[X > t] dt$$

Proof. Integration by parts. □

Lemma M.2. *Suppose a sequence of random variables $(X_n)_{n=1}^{\infty}$ converges almost surely to 0. Then for any $p \in [1, \infty)$, X_n converges to 0 in L^p as well if for some $q > p$, there exists a constant C such that $\mathbb{E}\,|X_n|^q < C$ for all n.*

This lemma's proof is a standard truncation trick.

Proof. For any $B > 0$, we have

$$\mathbb{E}\,|X_n|^p = \mathbb{E}\,|X_n|^p \mathbb{I}(|X_n|^p \le B) + \mathbb{E}\,|X_n|^p \mathbb{I}(|X_n|^p > B)$$

The random variable $|X_n|^p \mathbb{I}(|X_n|^p \le B)$ converges almost surely to 0, so by dominated convergence, it also converges in mean:

$$\mathbb{E}\,|X_n|^p \mathbb{I}(X_n \le B) \to 0.$$

This convergence happens for any fixed $B > 0$. So it suffices to show that $\sup_n \mathbb{E}\,|X_n|^p \mathbb{I}(|X_n|^p > B)$ becomes arbitrarily small as $B \to \infty$.

By Markov's Inequality, we have

$$\Pr[|X_n|^p > t] = \Pr[|X_n|^q > t^{q/p}] \le \frac{\mathbb{E}\,|X_n|^q}{t^{q/p}}$$

By Lemma M.1,

$$\begin{aligned}\mathbb{E}\,|X_n|^p \mathbb{I}(|X_n|^p > B) &= B \Pr[|X_n|^p > B] + \int_B^{\infty} \Pr[|X_n|^p > t] dt \\ &\le \frac{\mathbb{E}\,|X_n|^q}{B^{q/p-1}} + \int_B^{\infty} \frac{\mathbb{E}\,|X_n|^q}{t^{q/p}} dt.\end{aligned}$$

Since $\mathbb{E}\,|X_n|^q < C$ and $q/p > 1$ by assumption, this quantity goes to 0 as $B \to \infty$, as desired. □

Theorem M.3. *Consider a program in Setup J.2 and a scalar c in the program. If c converges almost surely to a limit $\mathring{c}$, then c converges in L^p to $\mathring{c}$ as well for every $p \in [1, \infty)$.*

Proof. This follows from Lemma M.2 and Lemma J.6. □

N Empirical validation

In the present section, we validate our main result, Theorem 3.7, by empirically checking some of its corollaries mentioned in Section 4.

N.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 Section 4, 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 3.7.

This means that the NNGP and NTK kernels of Corollaries 4.3 and 4.4 are not necessarily distribution-invariant. However both corollaries are still true as they merely state that these kernels exist.

N.2 Setup

We perform our experiments with vanilla RNN, vanilla GRU network, and a simple Transformer. We build on the code acompanying [29] and [30].

As for experiments aimed to validate NNGP correspondence, Corollary 4.3, 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 [22]. 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 4.4, 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 3.6, 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 3.6, which makes our Theorem 3.7 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.

N.3 Results

We start with empirically validating NNGP correspondence, Corollary 4.3. We first swap only hidden weights with their non-Gaussian counterparts. In this case, our Theorem 3.7 predicts that the limit should not depend on matrix entry distributions. As we see in Fig. 1, 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. 2, sometimes it is indeed the case (left plot), while sometimes the limit could be the same (right plot). We note that even

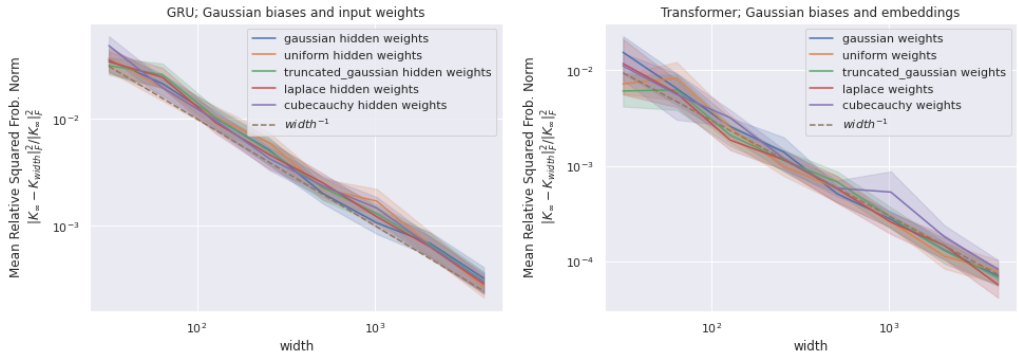

Figure 1: Relative Frobenius norm distance between empirical NNGP kernels for different distributions and the analytic kernel for Gaussian initialization. All trainable parameters expressed with vectors (i.e. biases, embeddings, and input weights) are assumed to be Gaussian. Left: a simple GRU network, right: a simple transformer, see Appendix N.2. As we see, NNGP kernels for all distributions considered converge to the same analytic kernel.

Figure 2: 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 N.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 N.1. Note that the limit kernel still exists and there is no contradiction with Corollary 4.3.

Figure 3: 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 N.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 N.1.

when the limit is different from the one resulted from Gaussian weights, Fig. 2 still demonstrates that the limit exists even for non-Gaussian ones, therefore validating Corollary 4.3.

Next, we validate neural tangent kernel convergence at initialization, checking a part of Corollary 4.4. As we see in Fig. 3, even when all trainable parameters are initialized with non-Gaussians, the limit kernel still remains the same.