Finite-Width Neural Tangent Kernels from Feynman Diagrams

Max Guillen*, Philipp Misof*, Jan E. Gerken

* Equal contribution

Department of Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, Sweden

arXiv:2508.11522

ICML International Conference On Machine Learning

The idea in two sentences

The infinite-width NTK is exactly computable but misses NTK evolution and feature learning. We import **Feynman diagrams** from quantum field theory to compute the **finite-width $1/n$ corrections** to NTK statistics — turning pages of algebra into a few pictures.

- **Feynman rules** for all statistics of preactivations, the NTK, and the higher-derivative dNTK / ddNTK tensors.
- Provably **complete to all orders** in $1/n$.
- New results: NTK-mean recursion, **gradient stability** at finite width, and **no corrections** for scale-invariant nonlinearities on the diagonal.
- Implemented & verified: matches sampled networks for $n \gtrsim 20$.

Background: the neural tangent kernel

For a network with parameters θ, the (empirical) *NTK* and *NNGP* kernels of layer ℓ are

$$\widehat{\Theta}^{(\ell)}_{ij}(x,x') = \sum_{\mu} \frac{\partial z^{(\ell)}_i(x)}{\partial \theta_\mu} \frac{\partial z^{(\ell)}_j(x')}{\partial \theta_\mu}, \qquad \widehat{K}^{(\ell)}_{ij}(x,x') = z^{(\ell)}_i(x)\, z^{(\ell)}_j(x'),$$

where $z^{(\ell)}_i$ are the channel-i preactivations. The NTK fixes the training dynamics to first order in the learning rate, but it is a **random** variable (initialization) and **evolves nonlinearly** during training.

The infinite-width limit

As the hidden width n goes to infinity, the NTK *freezes*: it becomes a deterministic Gaussian process, constant in training time [1], computable by layer-wise recursions [2]. Both gradient descent and Bayesian inference are then solvable in closed form.

At infinite width the network is **linear in its parameters** and shows **no feature learning**; finite networks deviate noticeably [3].

Going beyond: the $1/\text{width}$ expansion

Expand all statistics in $1/n$ [4]: the leading term is the Gaussian infinite-width limit, the $1/n$ corrections restore **NTK fluctuations, NTK evolution and feature learning**.

Express statistics in terms of mixed cumulants, decomposed over neural indices, e.g.

$$\mathbb{E}^c_\theta[z^{(\ell+1)}_{i_1}(x_1), z^{(\ell+1)}_{i_2}(x_2), \widehat{\Delta\Theta}^{(\ell+1)}_{i_3 i_4}(x_3,x_4)] = \frac{1}{n_\ell}\left(D^{(\ell+1)}_{1234}\delta_{i_1 i_2}\delta_{i_3 i_4} + F^{(\ell+1)}_{1324}\delta_{i_1 i_3}\delta_{i_2 i_4} + F^{(\ell+1)}_{1423}\delta_{i_1 i_4}\delta_{i_2 i_3}\right),$$

with $\widehat{\Delta\Theta} = \widehat{\Theta} - \mathbb{E}_\theta[\widehat{\Theta}]$.

This results in a fixed set of **tensors:** V and $K^{\{1\}}$ for preactivations, $\Theta^{\{1\}}$ for the NTK, A, B, D, F, for joint preactivation-NTK statistics and P, Q, R, S, T, U for dNTK/ddNTK.

Problem: deriving their recursions by hand is *algebraically brutal*.

Feynman diagrams for NTKs

Feynman rules assign an order in $1/n$ and an algebraic term to each graph. To compute a cumulant: **draw every admissible diagram with the right external legs and order, translate, and sum.**

(i) External lines

Preactivations are *solid*, NTK fluctuations are *dotted*:

(ii) Cubic interactions connect external and internal lines

Two external legs meet one *internal line*. Four cubic vertices arise:

with $\widehat{\Omega}^{(\ell+1)}_{i,\alpha\beta} = \sigma^{(\ell)}_{i,\alpha}\sigma^{(\ell)}_{i,\beta} + C^{(\ell+1)}_W \Theta^{(\ell)}_{\alpha\beta}\sigma'^{(\ell)}_{i,\alpha}\sigma'^{(\ell)}_{i,\beta}$ and $\widehat{\Delta\Omega} = \widehat{\Omega} - \langle\widehat{\Omega}\rangle_{K^{(\ell)}}$.

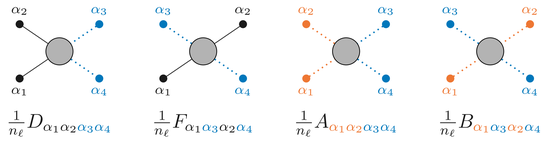
(iii) Propagators connect internal lines

Internal lines end on a *propagator* $\bigcirc$ = a zero-mean Gaussian expectation $\langle\cdot\rangle_{K^{(\ell)}}$ with covariance $K^{(\ell)}_{\alpha\beta}$, e.g.

Selection rules fix how internal lines may attach, forcing equal neural indices, paired sample indices, and the right K/Θ factors.

(iv) Quartic vertices = the rank-4 tensors

The 10 tensors in the cumulant decomposition are quartic vertices; the four with NTK/preactivation legs are



With four *internal* legs instead, the same vertices give the layer-ℓ tensors attached to propagators.

(v) Translating a diagram

Internal lines carry unassigned neural indices; internal solid lines carry unassigned sample indices. Multiply the factors of all vertices and propagators, sum over every unassigned index.

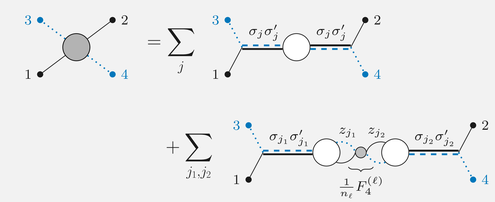
Worked example: the recursion for F

Goal: derive the layer recursion for F (here $C^\ell_W = 1$ and $n_\ell = n_{\ell-1}$)

$$F^{(\ell+1)}_{1324} = \langle\sigma_1\sigma_2\sigma'_3\sigma'_4\rangle_{K^{(\ell)}}\Theta^{(\ell)}_{34} + \sum_{\alpha\beta\gamma\delta}\langle\sigma_1\sigma'_3 z_\alpha\rangle\langle\sigma_2\sigma'_4 z_\beta\rangle K^{\alpha\gamma}K^{\beta\delta}F^{(\ell)}_{\gamma 3\delta 4}$$

from Feynman diagrams.

Draw *all* order-$1/n$ diagrams with two solid + two dotted legs. Only two survive the rules:



These are exactly the two terms of the recursion.

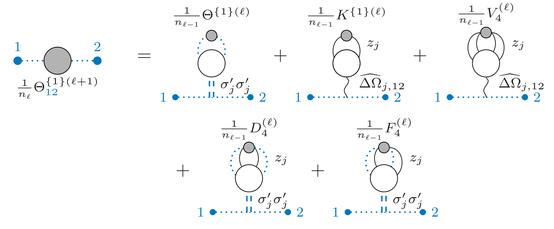
Theoretical guarantees

The rules reproduce the recursions of D, F, A, B *and* the dNTK/ddNTK tensors at order $1/n$, and — extended by higher-valence vertices — give a *complete* characterization at *all* orders.

Three applications

1. New recursion for the NTK mean $\Theta^{\{1\}}$

Previously unknown $1/n$ correction to the mean NTK:



2. Gradient stability at finite width

Thm 4. If the NNGP is critical, every cumulant involving the NTK is critical too.

The forward *and* backward pass share the same finite-width criticality.

3. No diagonal corrections for scale-invariant nonlinearities

For $\sigma(\lambda z) = \lambda\sigma(z)$, $\lambda > 0$, (e.g. ReLU, LeakyReLU):

Thm 5. The diagonal of the NTK mean receives no finite-width corrections — the infinite-width result is exact.

Experiments

PhilippMisofCH/ntk-unlimited

Recursions for $A, B, D, F, V, K^{\{1\}}, \Theta^{\{1\}}$ solved numerically and compared to sampled nets.

Finite-width corrected kernels (GeLU)

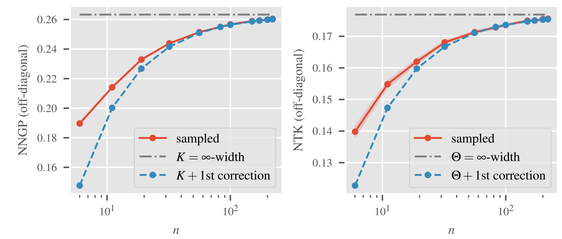


Figure 1. Off-diagonal NNGP $\overline{K}_{01}$ and NTK $\overline{\Theta}_{01}$: the $1/n$-corrected prediction (blue) tracks the sampled mean (red) for $n \gtrsim 20$; infinite width (gray) is off.

Gradient stability (ReLU)

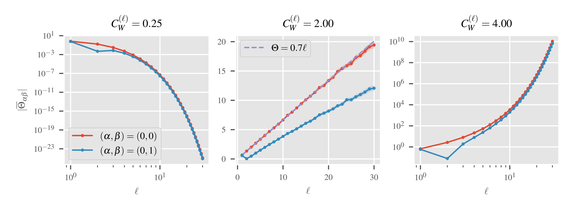


Figure 2. NTK vs. depth ℓ: at criticality $C^c_W = 2$ (middle) the NTK grows linearly, below (left) it shrinks exponentially and above (right) it grows exponentially.

No corrections for ReLU on the diagonal

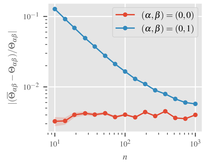


Figure 3. Relative deviation from infinite width: the diagonal gets *no* correction.

References

[1] Jacot et al. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: *Adv. Neural Inf. Process. Syst.* Vol. 31. Curran Associates, Inc., 2018.

[2] Novak et al. "Neural Tangents: Fast and Easy Infinite Neural Networks in Python". In: *Int. Conf. on Learn. Represent.* Apr. 2020.

[3] Lee et al. "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent". In: *Adv. Neural Inf. Process. Syst.* Vol. 32. Curran Associates, Inc., 2019.

[4] Roberts et al. *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge: Cambridge University Press, 2022.