Activation Functions Control Finite-Width Concentration in Wide Neural Networks

Soumya Ganguly SOUMYA.GANGULY@RUTGERS.EDU
Department of Mathematics, Rutgers University, Piscataway, NJ, USA

Nilava Metya NILAVA.METYA@RUTGERS.EDU
Department of Mathematics, Rutgers University, Piscataway, NJ, USA

Alexandre V. Morozov MOROZOV@PHYSICS.RUTGERS.EDU
Department of Physics and Astronomy, Rutgers University, Piscataway, NJ, USA

Anirvan M. Sengupta ANIRVANS.PHYSICS@GMAIL.COM
Department of Physics and Astronomy, Rutgers University, Piscataway, NJ, USA
Center for Computational Quantum Physics and Center for Computational Mathematics
Flatiron Institute, Simons Foundation, New York, NY, USA

Abstract

Wide randomly initialized neural networks are known to converge to Gaussian processes in the infinite-width limit. While this asymptotic limit is well understood, much less is known about the finite-width fluctuation behavior of empirical neural kernels and how this behavior depends on the activation function. In this work, we study finite-width concentration of random feature kernels through the lens of Orlicz and sub-Weibull tail theory. We show that activation growth directly controls the universality class of kernel fluctuations. Bounded activations such as $\tanh$ and erf satisfy Hoeffding-type concentration, ReLU activations exhibit sub-exponential Bernstein behavior, whereas polynomial activations generate sub-Weibull concentration regimes whose order depends explicitly on the polynomial degree. In particular, for $\varphi(x) = x^p$, the activation value $\varphi(G)$ has stretched-exponential tail order $2/p$ and Gaussians G, G_i, G_j, while the kernel summand $\varphi(G_i)\varphi(G_j)$ has sub-Weibull order $1/p$. This yields concentration inequalities with Weibull-type large-deviation behavior governed by the product tail. We derive entrywise concentration bounds and corresponding finite-dimensional operator-norm bounds for empirical neural kernels and illustrate the predicted scaling numerically in two-layer random networks. Our results suggest that activation growth provides a natural organizing principle for finite-width fluctuation regimes in wide neural networks.

Keywords: wide neural networks; random feature kernels; Gaussian process limits; Orlicz norms; sub-Weibull distribution; ReLU; polynomial activations.

1. Introduction

Wide neural networks [21] are closely connected to Gaussian processes and kernel methods [1, 13, 16, 17, 19, 27], leading to extensive work on lazy and feature-learning regimes [2, 4, 7, 26]. The demonstration of how the covariance of the intermediate-layer neuron

activity converges with the width of the network is a key part of these studies. Here, we look at this convergence through the lens of concentration inequalities, discovering the crucial role of the nature of the nonlinear activation function. We do this in the simplest yet non-trivial setting of a two-layer neural network.

2. Related work

Wide-network Gaussian-process limits. A major line of work studies the infinite-width limit of randomly initialized neural networks, where network outputs converge to Gaussian processes [17, 19]. These works establish convergence in distribution of finite collections of preactivations and outputs, together with recursive descriptions of the limiting covariance kernels. Subsequent developments clarified different scaling and learning regimes, including lazy and feature-learning limits. While this literature gives a detailed understanding of the infinite-width limit itself, the finite-width fluctuation behavior of empirical neural kernels has received comparatively little attention.

Finite-width fluctuations and moment methods. Recent work by Hanin studies finite-width corrections to Gaussian-process behavior through characteristic functions, collective variables, and perturbative expansions [10, 11]. In particular, moment estimates for collective-variable fluctuations show the expected decay with width and provide a perturbative description of deviations from the infinite-width limit. However, these analyses do not explicitly characterize activation-dependent tail regimes or concentration classes in terms of Orlicz norms (or quasi-norms) [22] or sub-Weibull tails.

Sub-Weibull random variables and neural networks. Vladimirova et al. [23] studied heavy-tailed behavior induced by Bayesian neural-network priors at the level of hidden units and network activations. In particular, they observed that compositions of nonlinear transformations can naturally generate sub-Weibull distributions. This perspective was further developed in Ref. [24], which introduced a systematic probabilistic framework for sub-Weibull random variables and their concentration properties. Our work differs in focus as follows: rather than studying the marginal distribution of hidden units induced by Bayesian priors, we study finite-width concentration of empirical neural kernels and how their fluctuation behavior depends explicitly on the activation growth.

Why activation growth controls concentration. The empirical neural kernel is an average over hidden-unit contributions,

$$(Q_N^{(1)})_{ij} = \frac{1}{N}\sum_{k=1}^{N} \varphi(h_k(x_i))\varphi(h_k(x_j)).$$

In the infinite-width limit, the law of large numbers gives convergence to the deterministic kernel $K_{ij} = \mathbb{E}\left[\varphi(G_i)\varphi(G_j)\right]$, where (G_i, G_j) is Gaussian. Finite-width concentration is therefore governed by the tails of the summands $\varphi(h_k(x_i))\varphi(h_k(x_j))$. The activation function determines these tails. For bounded activations such as $\tanh$ or erf, each summand is bounded, yielding Hoeffding-type concentration. For ReLU activations, Gaussian preactivations produce sub-exponential tails and Bernstein-type concentration. Polynomial activa-

tions generate qualitatively heavier tails. Indeed, if $Z \sim \mathcal{N}(0,1)$ and $\varphi(z) = z^p$, then

$$\mathbb{P}\left[|\varphi(Z)| > t\right] = \mathbb{P}\left[|Z| > t^{1/p}\right] \approx \exp(-ct^{2/p}).$$

For kernel entries, however, the relevant summand is a product $\varphi(G_i)\varphi(G_j) = G_i^p G_j^p$, which is sub-Weibull of order $1/p$. Thus increasing activation growth weakens finite-width concentration and changes the universality class of kernel fluctuations.

3. Model

Fix deterministic inputs $x_1, \ldots, x_P \in \mathbb{R}^D$, width N, and activation $\varphi : \mathbb{R} \to \mathbb{R}$. Let $W^{(1)} \in \mathbb{R}^{N \times D}$ have i.i.d. entries $N(0, \sigma_1^2)$, and let $w^{(2)} \in \mathbb{R}^N$ have i.i.d. entries $N(0, \tau^2/N)$, independent of $W^{(1)}$. Define

$$h_k(x_i) = \sum_{l=1}^{D} W_{kl}^{(1)}(x_i)_l, \qquad \Phi_{ik} = \varphi(h_k(x_i)), \qquad f = \Phi w^{(2)}.$$

The empirical feature Gram matrix or feature kernel is

$$Q_N^{(1)} = \frac{1}{N}\Phi\Phi^\top, \qquad (Q_N^{(1)})_{ij} = \frac{1}{N}\sum_{k=1}^{N} \varphi(h_k(x_i))\varphi(h_k(x_j)).$$

Let $Q_{ij}^{(0)} = x_i^\top x_j$ and $G \sim N(0, \sigma_1^2 Q^{(0)})$. The deterministic limiting kernel is

$$K_{ij} = \mathbb{E}\left[\varphi(G_i)\varphi(G_j)\right].$$

Under the second moment condition $\mathbb{E}\left[\varphi(G_i)^2\right] < \infty$, the strong law gives $(Q_N^{(1)})_{ij} \to K_{ij}$ almost surely for every fixed pair (i, j). Moreover, conditional on Φ, the output is Gaussian,

$$f \mid \Phi \sim N(0, \tau^2 Q_N^{(1)}),$$

so convergence of $Q_N^{(1)}$ is exactly the covariance convergence underlying the finite-dimensional Gaussian-process limit. We study rates of this convergence for various activations φ including $\mathrm{erf}, \tanh, \mathrm{ReLU}$ and polynomial x^p.

Finally, we use the Orlicz ψ_α norm, defined in Appendix F. For a real random variable X and $\alpha > 0$,

$$\|X\|_{\psi_\alpha} := \inf\left\{ s > 0 \;\middle|\; \mathbb{E}\left[e^{\left(\frac{|X|}{s}\right)^\alpha}\right] \le 2 \right\}.$$

If $\|X\|_{\psi_\alpha} < \infty$, then X is said to be sub-Weibull of order α.

4. Main Result

The following theorem summarizes the entrywise concentration mechanism used throughout the paper.

Theorem 1 *Let $\{(X_k, Y_k)\}_{k=1}^N$ be i.i.d. copies of a centered bivariate Gaussian pair (X, Y), and assume that, for some $\alpha > 0$ and $K > 0$,*

$$\|\varphi(X_k)\varphi(Y_k)\|_{\psi_\alpha} \le K \qquad \text{for all } k.$$

Then sub-Weibull Bernstein inequality says that $\exists\ c_\alpha > 0$, depending only on α, such that for all $t \ge 0$,

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{k=1}^{N}\varphi(X_k)\varphi(Y_k) - \mathbb{E}\left[\varphi(X)\varphi(Y)\right]\right| \ge t\right] \le 2\exp\left[-c_\alpha \min\left\{\frac{Nt^2}{K^2}, \left(\frac{Nt}{K}\right)^\alpha\right\}\right].$$

- *For bounded activations like $\varphi = \mathrm{erf}$ and $\varphi = \tanh$, the summands are bounded, so the Nt^2/K^2 term dominates with $\alpha = 2$ and Hoeffding-type concentration follows.*
- *For $\varphi = \mathrm{ReLU}$, Gaussianity implies that $\mathrm{ReLU}(X)\mathrm{ReLU}(Y)$ is sub-exponential, so $\alpha = 1$ and sub-exponential-type concentration follows.*
- *For $\varphi(x) = x^p$, the summand X^pY^p is sub-Weibull of order $1/p$, so $\alpha = 1/p$ and sub-Weibull Bernstein inequality above applies.*

The proof is given in the appendix, immediately after Theorem 30.

This theorem states convergence for one entry of the random feature kernel $Q^{(1)}$. To get operator norm convergence, one simply uses union bound over all P^2 entries of $Q^{(1)}$, resulting in a high probability estimate of the operator norm of $Q^{(1)} - \mathbb{E}\left[Q^{(1)}\right]$. The treatments for the $\mathrm{erf}, \tanh, \mathrm{ReLU}$ activations have been given in Appendices B, C, D respectively. A similar theory for polynomial activations [8, 14, 18] requires a discussion of Orlicz norms developed in Appendix F. Here are the limiting kernels for each of these activations.

Limiting kernels for standard activations. For (X, Y) centered Gaussian with variances a, b, covariance c, correlation $\rho = c/\sqrt{ab}$, and $\theta = \arccos(\rho)$, the limiting kernels are:

φ	$\mathbb{E}\left[\varphi(X)\varphi(Y)\right]$
ReLU	$\frac{\sqrt{ab}}{2\pi}\left(\sin\theta + (\pi - \theta)\cos\theta\right)$
erf	$\frac{2}{\pi}\arcsin\left(\frac{2c}{\sqrt{(1+2a)(1+2b)}}\right)$
x^p	$\sum_{m=0}^{\lfloor p/2\rfloor}\frac{(p!)^2}{(p-2m)!2^{2m}(m!)^2}c^{p-2m}(ab)^m$

These formulas are classical: the ReLU expression is the arc-cosine kernel of Cho and Saul [6], related formulas appear in the Gaussian-process literature, and the polynomial expression follows directly from Wick's theorem for Gaussian moments. For $\tanh$, the kernel admits an absolutely convergent Hermite expansion in the Gaussian correlation parameter shown in Theorem 7. These formulas, together with the concentration theorem, give a compact finite-width description of the prior random-feature kernel.

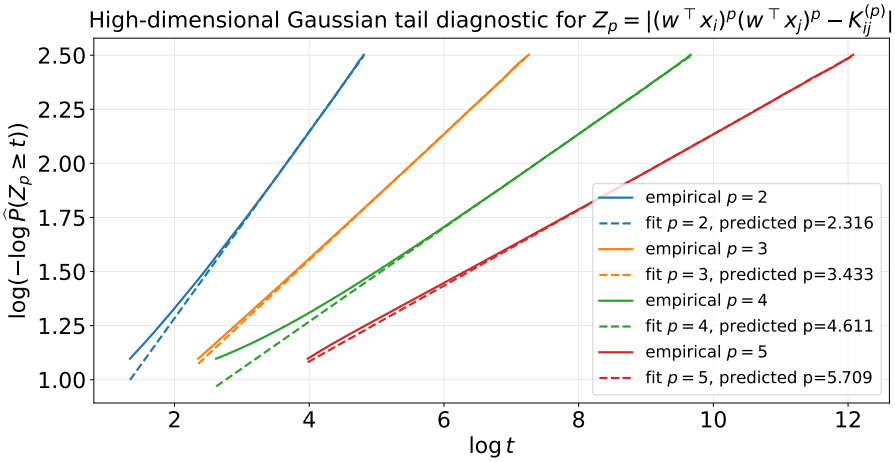

Figure 1: Sub-Weibull scaling for polynomial activations $\varphi(x) = x^p$ with $p \in \{2, 3, 4, 5\}$.

5. Experiments

We empirically diagnose the Orlicz tail exponent in the high-dimensional random-feature model at the level of a single kernel summand. We fix two normalized inputs $x_i, x_j \in \mathbb{R}^D$ and repeatedly draw Gaussian weights $w \sim \mathcal{N}(0, I_D)$. For each polynomial activation $\varphi_p(x) = x^p$, we form the centered one-neuron kernel contribution

$$Z_p = \left| (w^\top x_i)^p (w^\top x_j)^p - K_{ij}^{(p)} \right|, \qquad K_{ij}^{(p)} = \mathbb{E}\left[(w^\top x_i)^p (w^\top x_j)^p \right].$$

The value $K_{ij}^{(p)}$ is computed exactly by Wick's formula as in Theorem 14 using the covariance matrix of the bivariate Gaussian $(w^\top x_i, w^\top x_j)$, whose entries are $\|x_i\|^2, \|x_j\|^2$, and $x_i^\top x_j$. We generate M independent samples of Z_p, but do so in batches of size B to avoid storing the full $M \times D$ Gaussian weight matrix in memory at once. In each batch, we sample B independent weights, compute the corresponding inner products $w^\top x_i$ and $w^\top x_j$, and store the resulting Z_p values. In our experiments we use, for example, $D = 100, M = 5 \times 10^7$ total Gaussian samples, and batch size $B = 10^4$. From the collected samples, we estimate the empirical survival probability $\widehat{P}(Z_p \geq t)$ over upper-tail thresholds t, typically between the 95^{th} and 99.9995^{th} empirical percentiles, using a geometrically spaced grid. Finally, we plot

$$\log\left(-\log \widehat{P}(Z_p \geq t)\right) \quad \text{against} \quad \log t.$$

If Z_p has a sub-Weibull tail with index $1/p$, then $\mathbb{P}(Z_p \geq t) \approx \exp(-Ct^{1/p})$, and therefore the plots in Figure 1 should be approximatel y linear with slope $1/p$ in the upper tail. This experiment directly tests the predicted Orlicz hierarchy for polynomial activations in the high-dimensional Gaussian-weight model.

6. Discussion

Our results suggest that activation growth provides a natural organizing principle for finite-width fluctuation regimes in wide neural networks. While infinite-width Gaussian-process

limits are universal at the level of covariance kernels, finite-width concentration behavior depends strongly on the activation-induced tail structure.

The present work focuses on random two-layer networks as a minimal controlled setting. An important future direction is extending these ideas to trained networks, deeper architectures, and feature-learning regimes beyond the lazy limit. Another natural direction is understanding how activation-dependent fluctuation classes interact with optimization, representation learning, and heavy-tailed phenomena observed in modern deep-learning systems.

Acknowledgements. A.V.M. and A.M.S. acknowledge financial and logistical support from the Center for Quantitative Biology, Rutgers University.

References

[1] Luisa Andreis, Federico Bassetti, and Christian Hirsch. Ldp for the covariance process in fully connected neural networks, 2025. URL https://arxiv.org/abs/2505.08062.

[2] Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation, 2022. URL https://arxiv.org/abs/2205.01445.

[3] Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

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

[5] Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.

[6] Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper_files/paper/2009/file/5751ec3e9a4feab575962e78e006250d-Paper.pdf.

[7] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time, 2025. URL https://arxiv.org/abs/2305.18270.

[8] Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. In *International conference on machine learning*, pages 1329–1338. PMLR, 2018.

[9] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

[10] Boris Hanin. Random neural networks in the infinite width limit as Gaussian processes. *The Annals of Applied Probability*, 33(6A):4798 – 4819, 2023. doi: 10.1214/23-AAP1933. URL https://doi.org/10.1214/23-AAP1933.

[11] Boris Hanin. Random fully connected neural networks as perturbatively solvable hierarchies. *Journal of Machine Learning Research*, 25(267):1–58, 2024. URL http://jmlr.org/papers/v25/23-0643.html.

[12] Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

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

[14] Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks, 2019. URL https://arxiv.org/abs/1905.12207.

[15] Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, 2022. ISSN 2049-8772. doi: 10.1093/imaiai/iaac012. URL http://dx.doi.org/10.1093/imaiai/iaac012.

[16] Clarissa Lauditi, Blake Bordelon, and Cengiz Pehlevan. Adaptive kernel predictors from feature-learning infinite limits of neural networks, 2025. URL https://arxiv.org/abs/2502.07998.

[17] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

[18] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. *Advances in neural information processing systems*, 27, 2014.

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

[20] Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

[21] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[22] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[23] Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo, and Julyan Arbel. Understanding priors in Bayesian neural networks at the unit level. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6458–6467. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/vladimirova19a.html.

[24] Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1), January 2020. ISSN 2049-1573. doi: 10.1002/sta4.318. URL http://dx.doi.org/10.1002/sta4.318.

[25] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

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

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

Appendix A. Setup and basic observations

Fix integers $D \geq 1$ (input dimension), $P \geq 1$ (number of datapoints), and $N \geq 1$ (width). Let $x_1, \ldots, x_P \in \mathbb{R}^D$ be deterministic inputs, and write the *input Gram matrix*

$$Q^{(0)} \in \mathbb{R}^{P \times P}, \qquad Q^{(0)}_{ij} \coloneqq x_i^\top x_j.$$

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a measurable function (activation), applied coordinatewise.

Random weights. Let $W^{(1)} \in \mathbb{R}^{N \times D}$ have i.i.d. entries $W^{(1)}_{\alpha k} \sim \mathcal{N}(0, \sigma_1^2)$. Let $w^{(2)} \in \mathbb{R}^N$ have i.i.d. entries $w^{(2)}_\alpha \sim \mathcal{N}(0, \tau^2/N)$, independent of $W^{(1)}$. (This is the standard scaling that yields a non-degenerate limit.) Let $z \in \mathbb{R}^P$ be standard Gaussian $z \sim \mathcal{N}(0, I_P)$, independent of everything else, and fix $\sigma \geq 0$.

Hidden features and outputs. For each datapoint x_i, define hidden pre-activations

$$h(x_i) \coloneqq W^{(1)} x_i \in \mathbb{R}^N, \qquad h_\alpha(x_i) = \sum_{k=1}^{D} W^{(1)}_{\alpha k} (x_i)_k,$$

and define the (random feature) matrix $\Phi \in \mathbb{R}^{P \times N}$ by

$$\Phi_{i\alpha} \coloneqq \varphi\big(h_\alpha(x_i)\big).$$

Define the network output vector $f \in \mathbb{R}^P$ and noisy observations $y \in \mathbb{R}^P$ by

$$f \coloneqq \Phi w^{(2)}, \qquad y \coloneqq f + \sigma z.$$

Integrability assumption. We assume

$$\mathbb{E}\big[\varphi(G)^2\big] < \infty \quad \text{for } G \sim \mathcal{N}(0, \sigma_1^2 \, \|x\|^2) \text{ for every } x \in \{x_1, \ldots, x_P\}. \tag{1}$$

Equivalently, for the P-variate Gaussian appearing below, all second moments of φ exist.

Structure of hidden layer. For each neuron index $\alpha \in [N]$, define the P-vector

$$g_\alpha \coloneqq \big(h_\alpha(x_1), \ldots, h_\alpha(x_P)\big) \in \mathbb{R}^P.$$

It is clear that

$$g_\alpha \sim \mathcal{N}\big(0, \ \sigma_1^2 Q^{(0)}\big).$$

Define the random *feature Gram matrix* by

$$Q^{(1)}_N \coloneqq \frac{1}{N} \Phi \Phi^\top \in \mathbb{R}^{P \times P} \tag{2}$$

Recall that this means

$$(Q^{(1)}_N)_{ij} = \frac{1}{N} \sum_{\alpha=1}^{N} \varphi(h_\alpha(x_i)) \varphi(h_\alpha(x_j)).$$

Given Φ, the map $w^{(2)} \mapsto f = \Phi w^{(2)}$ is linear. Since $w^{(2)} \sim \mathcal{N}\left(0, \frac{\tau^2}{N} I_N\right)$, it follows that $f|\Phi$ is Gaussian with covariance

$$\mathrm{Cov}(f \mid \Phi) = \Phi \left(\frac{\tau^2}{N} I_N\right) \Phi^\top = \tau^2 \frac{1}{N} \Phi\Phi^\top = \tau^2 Q_N^{(1)}.$$

Therefore

$$y\big|\Phi = (f + \sigma z)\big|\Phi \sim \mathcal{N}\left(0, \tau^2 Q_N^{(1)} + \sigma^2 I_P\right).$$

Limiting kernel. Let $G \sim \mathcal{N}(0, \sigma_1^2 Q^{(0)})$ be a P-variate Gaussian. Define

$$K \in \mathbb{R}^{P\times P}, \qquad K_{ij} \coloneqq \mathbb{E}\big[\varphi(G_i)\varphi(G_j)\big].$$

Under (1), K_{ij} is finite for all i, j. Here is the standard LLN argument to show that K is the limiting kernel under the integrability assumption. For each α, define the random vector $u_\alpha \coloneqq (\varphi(g_{\alpha,i}))_{i\in[P]} \in \mathbb{R}^P$ so that $(Q_N^{(1)})_{ij} = \frac{1}{N}\sum_{\alpha=1}^{N} u_{\alpha,i}u_{\alpha,j}$. The g_α are i.i.d. $\mathcal{N}(0, \sigma_1^2 Q^{(0)})$, hence the u_α are i.i.d. in $\mathbb{R}^P$ with $\mathbb{E}\left[u_{\alpha,i}u_{\alpha,j}\right] = K_{ij}$ and $\mathbb{E}\left[u_{\alpha,i}^2\right] < \infty$ by (1). For each fixed pair (i, j), the random variables $u_{\alpha,i}u_{\alpha,j}$ are i.i.d. and integrable. Indeed, by Cauchy–Schwarz, $\mathbb{E}\left[|u_{\alpha,i}u_{\alpha,j}|\right] \le \left(\mathbb{E}\left[u_{\alpha,i}^2\right]\right)^{1/2}\left(\mathbb{E}\left[u_{\alpha,j}^2\right]\right)^{1/2} < \infty$. By the Strong Law of Large Numbers, $\frac{1}{N}\sum_{\alpha=1}^{N} u_{\alpha,i}u_{\alpha,j} \xrightarrow{\text{a.s.}} \mathbb{E}\left[u_{\alpha,i}u_{\alpha,j}\right] = K_{ij}$.

Large-N Gaussian process limit of the output distribution. We show that the unconditional distribution of f converges to a *deterministic* Gaussian law with covariance $\tau^2 K$, and the same for y with added noise.

Lemma 2 (Finite-dimensional GP limit) *Assume* (1). *Let $f_N \in \mathbb{R}^P$ denote the width-N output vector $f_N = \Phi w^{(2)}$, and $y_N = f_N + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Then, as $N \to \infty$,*

$$f_N \overset{d}{\Rightarrow} \mathcal{N}(0, \tau^2 K), \qquad y_N \overset{d}{\Rightarrow} \mathcal{N}(0, \tau^2 K + \sigma^2 I_P).$$

Proof We first prove the statement for f_N. Fix $t \in \mathbb{R}^P$ and consider the characteristic function $\varphi_N(t) \coloneqq \mathbb{E}\left[\exp(it^\top f_N)\right]$. Conditioned on Φ we have $f_N|\Phi \sim \mathcal{N}(0, \tau^2 Q_N^{(1)})$, hence $\mathbb{E}\big[e^{it^\top f_N} \mid \Phi\big] = \exp\left(-\frac{1}{2}\tau^2 t^\top Q_N^{(1)} t\right)$. By law of total expectation, $\varphi_N(t) = \mathbb{E}\Big[\exp\left(-\frac{1}{2}\tau^2 t^\top Q_N^{(1)} t\right)\Big]$. By above, $Q_N^{(1)} \to K$ almost surely, so $t^\top Q_N^{(1)} t \to t^\top K t$ almost surely. Moreover, for every N, $0 \le \exp\left(-\frac{1}{2}\tau^2 t^\top Q_N^{(1)} t\right) \le 1$. The dominated convergence theorem gives $\varphi_N(t) \to \exp\left(-\frac{1}{2}\tau^2 t^\top K t\right)$, for every t, which is the characteristic function of $\mathcal{N}(0, \tau^2 K)$. By Lévy's continuity theorem, $f_N \Rightarrow \mathcal{N}(0, \tau^2 K)$.

For $y_N = f_N + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_P)$ is independent of f_N, the characteristic function factorizes:

$$\mathbb{E}\left[e\right]^{it^\top y_N} = \mathbb{E}\left[e\right]^{it^\top f_N}\ \mathbb{E}\left[e\right]^{it^\top \varepsilon}.$$

The first factor converges to $\exp(-\frac{1}{2}\tau^2 t^\top K t)$, and the second equals $\exp(-\frac{1}{2}\sigma^2\|t\|^2)$. Hence

$$y_N \Rightarrow \mathcal{N}(0, \tau^2 K + \sigma^2 I_P).$$

■

Appendix B. Computing $K_{ij} = \mathbb{E}\big[\varphi(G_i)\varphi(G_j)\big]$ for $\varphi(x) = \mathrm{erf}(x)$

Recall the earlier setup: $G \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \sigma_1^2 Q^{(0)}$. Fix i, j and set

$$a \coloneqq \Sigma_{ii}, \qquad b \coloneqq \Sigma_{jj}, \qquad c \coloneqq \Sigma_{ij}, \qquad \rho \coloneqq \frac{c}{\sqrt{ab}} \in [-1, 1],$$

with $\rho = 0$ if $ab = 0$. Also recall the definition of the function

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt.$$

Let $\varphi(x) = \mathrm{erf}(x)$ and $K_{ij} \coloneqq \mathbb{E}\big[\mathrm{erf}(G_i)\mathrm{erf}(G_j)\big]$.

Theorem 3 (Closed form for the erf kernel) *Let (X, Y) be a centered bivariate Gaussian with variances a, b and covariance c. Then*

$$\mathbb{E}\big[\mathrm{erf}(X)\mathrm{erf}(Y)\big] = \frac{2}{\pi} \arcsin\left(\frac{2c}{\sqrt{(1+2a)(1+2b)}}\right).$$

In particular, with $(X, Y) = (G_i, G_j)$ this equals

$$K_{ij} = \frac{2}{\pi} \arcsin\left(\frac{2\Sigma_{ij}}{\sqrt{(1+2\Sigma_{ii})(1+2\Sigma_{jj})}}\right).$$

Proof Let

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} a & c \\ c & b \end{bmatrix}\right).$$

Recall the identity $\mathrm{erf}(x) = 2\Phi(\sqrt{2}x) - 1$, where Φ is the standard normal CDF. Equivalently, if $U \sim \mathcal{N}(0, 1)$ is independent of x, then $\mathrm{erf}(x) = \mathbb{E}\left[\mathrm{sign}\left(\sqrt{2}x - U\right)\right]$, since

$$\mathbb{E}\left[\mathrm{sign}\left(\sqrt{2}x - U\right)\right] = \mathbb{P}\left[U \le \sqrt{2}x\right] - \mathbb{P}\left[U > \sqrt{2}x\right] = 2\Phi\left(\sqrt{2}x\right) - 1.$$

Let $U, V \sim \mathcal{N}(0, 1)$ be independent of each other and of (X, Y). Then

$$\mathbb{E}\left[\mathrm{erf}(X)\mathrm{erf}(Y)\right] = \mathbb{E}\left[\mathrm{sign}(\sqrt{2}X - U)\,\mathrm{sign}(\sqrt{2}Y - V)\right].$$

Define

$$A \coloneqq \sqrt{2}X - U, \qquad B \coloneqq \sqrt{2}Y - V.$$

The pair (A, B) is jointly Gaussian with mean zero and covariance matrix

$$\begin{bmatrix} 2a+1 & 2c \\ 2c & 2b+1 \end{bmatrix}.$$

Hence their correlation is

$$r = \mathrm{Corr}(A, B) = \frac{2c}{\sqrt{(2a+1)(2b+1)}}.$$

Applying the arcsine law for any zero-mean jointly Gaussian pair (A, B) with correlation r: $\mathbb{E}\left[\mathrm{sign}(A)\,\mathrm{sign}(B)\right] = \frac{2}{\pi}\arcsin(r)$. Rewritten, this is simply

$$\mathbb{E}\left[\mathrm{erf}(X)\mathrm{erf}(Y)\right] = \frac{2}{\pi}\arcsin\left(\frac{2c}{\sqrt{(2a+1)(2b+1)}}\right).$$

■

Concentration using Hoeffding Since $\mathrm{erf}(x) \in [-1, 1]$, the product $\mathrm{erf}(h_\alpha(x_i))\mathrm{erf}(h_\alpha(x_j)) \in [-1, 1]$. Each centered summand is bounded in $[-2, 2]$, thus applying Hoeffding's inequality immediately yields

Theorem 4 (Entrywise exponential concentration for erf) *For all $t \geq 0$,*

$$\mathbb{P}\left[\left|(Q_N^{(1)})_{ij} - K_{ij}\right| \geq t\right] \leq 2\exp\left(-\frac{Nt^2}{8}\right).$$

Corollary 5 *For all $t \geq 0$,*

$$\mathbb{P}\left[\max_{i,j}\left|(Q_N^{(1)})_{ij} - K_{ij}\right| \geq t\right] \leq 2P^2\exp\left(-\frac{Nt^2}{8}\right),$$

and on the complementary event $\left\|Q_N^{(1)} - K\right\|_{\mathrm{op}} \leq \left\|Q_N^{(1)} - K\right\|_F \leq Pt$.

Appendix C. Computing $K_{ij} = \mathbb{E}\big[\varphi(G_i)\varphi(G_j)\big]$ for $\varphi(x) = \tanh(x)$

Recall the earlier setup: $G \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \sigma_1^2 Q^{(0)}$. Fix i, j and set

$$a \coloneqq \Sigma_{ii}, \qquad b \coloneqq \Sigma_{jj}, \qquad c \coloneqq \Sigma_{ij}, \qquad \Sigma_{(i,j)} \coloneqq \begin{pmatrix} a & c \\ c & b \end{pmatrix}.$$

C.1. Limiting kernel

Define $\varphi(x) = \tanh(x)$ and

$$K_{ij} \coloneqq \mathbb{E}\big[\tanh(G_i)\tanh(G_j)\big].$$

Theorem 6 (Explicit integral representation) *Let $(X, Y) \sim \mathcal{N}(0, \Sigma_{(i,j)})$. Then*

$$K_{ij} = \iint_{\mathbb{R}^2} \tanh(x) \tanh(y)\, \varphi_{a,b,c}(x, y)\, dx\, dy,$$

where $\varphi_{a,b,c}$ is the bivariate centered Gaussian density with covariance $\Sigma_{(i,j)}$. Moreover, the integral is absolutely convergent and satisfies $|K_{ij}| \leq 1$.

Proof Since $|\tanh| \leq 1$, $\tanh(X)\tanh(Y)$ is integrable and the expectation equals the stated integral by definition of expectation with respect to the density. Absolute convergence and the bound follow from $|\tanh(x)\tanh(y)| \leq 1$. ■

C.2. Hermite-series representation (orthonormal coordinates for the Gaussian measure space)

Assume $ab > 0$ and let $\rho := c/\sqrt{ab} \in [-1, 1]$. Let (U, V) be standard bivariate normal with correlation ρ and write

$$X = \sqrt{a}U, \qquad Y = \sqrt{b}V.$$

Define the probabilists' Hermite polynomials $(\mathrm{He}_n)_{n\geq 0}$ by

$$\mathrm{He}_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2},$$

which form an orthogonal basis of $L^2(\mathbb{R}, \gamma)$ where γ is standard normal measure.

For $n \geq 0$ define coefficients

$$\alpha_n(a) := \frac{1}{n!}\mathbb{E}\left[\tanh(\sqrt{a}Z)\mathrm{He}_n(Z)\right], \qquad \beta_n(b) := \frac{1}{n!}\mathbb{E}\left[\tanh(\sqrt{b}Z)\mathrm{He}_n(Z)\right], \quad Z \sim \mathcal{N}(0, 1).$$

Theorem 7 (Hermite expansion and kernel series) *For each fixed $a > 0$, the function $z \mapsto \tanh(\sqrt{a}z)$ belongs to $L^2(\gamma)$ and admits the $L^2(\gamma)$ expansion*

$$\tanh(\sqrt{a}z) = \sum_{n=0}^{\infty} \alpha_n(a)\mathrm{He}_n(z), \qquad \textit{with} \quad \sum_{n=0}^{\infty} n!\alpha_n(a)^2 = \mathbb{E}\left[\tanh(\sqrt{a}Z)^2\right] < 1.$$

Similarly for b. Moreover, with (U, V) as above,

$$K_{ij} = \mathbb{E}\big[\tanh(\sqrt{a}U)\tanh(\sqrt{b}V)\big] = \sum_{n=0}^{\infty} n!\alpha_n(a)\beta_n(b)\rho^n,$$

and the series converges absolutely for $|\rho| < 1$ and converges at $\rho = \pm 1$ by Cauchy–Schwarz. In particular, since $\tanh$ is odd, $\alpha_n(a) = \beta_n(b) = 0$ for all even n, so the sum is over odd n only.

Proof Since $|\tanh| \leq 1$, we have $\tanh(\sqrt{a}Z) \in L^2(\gamma)$, hence it has an L^2 Hermite expansion with coefficients $\alpha_n(a)$ as defined. Parseval's identity gives $\sum n!\alpha_n(a)^2 = \mathbb{E}[\tanh(\sqrt{a}Z)^2]$.

For correlated standard Gaussian distributions (U, V) with correlation ρ, the Mehler identity implies $\mathbb{E}\left[\mathrm{He}_n(U)\mathrm{He}_m(V)\right] = \delta_{nm} n! \rho^n$. Expanding both $\tanh(\sqrt{a}U)$ and $\tanh(\sqrt{b}V)$ in Hermite series and using orthogonality yields

$$\mathbb{E}\left[\tanh(\sqrt{a}U)\tanh(\sqrt{b}V)\right] = \sum_{n,m} \alpha_n(a)\beta_m(b)\mathbb{E}\left[\mathrm{He}_n(U)\mathrm{He}_m(V)\right] = \sum_{n=0}^{\infty} n!\alpha_n(a)\beta_n(b)\rho^n.$$

Absolute convergence for $|\rho| < 1$ follows by Cauchy–Schwarz and Parseval:

$$\sum_{n\geq 0} n! \left|\alpha_n(a)\beta_n(b)\right| |\rho|^n \leq \left(\sum_{n\geq 0} n!\alpha_n(a)^2\right)^{1/2} \left(\sum_{n\geq 0} n!\beta_n(b)^2\right)^{1/2} < \infty,$$

and convergence at $\rho = \pm 1$ follows from the same bound. Oddness of $\tanh$ forces vanishing of even Hermite coefficients. ∎

Concentration using Hoeffding. Define

$$(Q_N^{(1)})_{ij} \coloneqq \frac{1}{N}\sum_{\alpha=1}^{N} \tanh(h_\alpha(x_i))\ \tanh(h_\alpha(x_j)).$$

As before, $\mathbb{E}\left[(Q_N^{(1)})_{ij}\right] = K_{ij}$.

Since $\tanh \in (-1, 1)$, each centered summand is bounded in $[-2, 2]$, hence Hoeffding applies.

Theorem 8 (Entrywise exponential concentration for $\tanh$) *For all $t \geq 0$,*

$$\mathbb{P}\left[\left|(Q_N^{(1)})_{ij} - K_{ij}\right| \geq t\right] \leq 2\exp\left(-\frac{Nt^2}{8}\right).$$

Corollary 9 *For all $t \geq 0$,*

$$\mathbb{P}\left[\max_{i,j}\left|(Q_N^{(1)})_{ij} - K_{ij}\right| \geq t\right] \leq 2P^2\exp\left(-\frac{Nt^2}{8}\right),$$

and on the complementary event $\left\|Q_N^{(1)} - K\right\|_{\mathrm{op}} \leq \left\|Q_N^{(1)} - K\right\|_F \leq Pt$.

Appendix D. Computing $K_{ij} = \mathbb{E}[\varphi(G_i)\varphi(G_j)]$ for $\varphi(x) = \mathrm{ReLU}(x)$

$\Sigma \coloneqq \sigma_1^2 Q^{(0)}$ with $Q_{ij}^{(0)} = x_i \cdot x_j$, and $G = (G_1, \ldots, G_P) \sim \mathcal{N}(0, \Sigma)$. Fix i, j and set

$$a \coloneqq \Sigma_{ii} = \sigma_1^2 \left\|x_i\right\|^2, \qquad b \coloneqq \Sigma_{jj} = \sigma_1^2 \left\|x_j\right\|^2, \qquad c \coloneqq \Sigma_{ij} = \sigma_1^2(x_i \cdot x_j), \qquad \rho \coloneqq \frac{c}{\sqrt{ab}} \in [-1, 1],$$

with the convention $\rho = 0$ if $ab = 0$. Let $\theta \coloneqq \arccos(\rho) \in [0, \pi]$.

Define $\varphi(x) = \mathrm{ReLU}(x) \coloneqq \max\{0, x\}$ and

$$K_{ij} \coloneqq \mathbb{E}\big[\varphi(G_i)\varphi(G_j)\big] = \mathbb{E}\big[\mathrm{ReLU}(G_i)\mathrm{ReLU}(G_j)\big].$$

The following calculation already exists due to Cho and Saul [6] as the computation of J_1 in their Appendix A, but we show it here for completeness.

Theorem 10 (Closed form for the ReLU kernel) *Let (X, Y) be a centered bivariate Gaussian with $\mathbb{E}[X^2] = a$, $\mathbb{E}[Y^2] = b$, and correlation $\rho = \mathbb{E}[XY]/\sqrt{ab} \in [-1, 1]$ (with $ab > 0$). Then*

$$\mathbb{E}\left[\mathrm{ReLU}(X)\mathrm{ReLU}(Y)\right] = \frac{\sqrt{ab}}{2\pi}\Big(\sin\theta + (\pi - \theta)\cos\theta\Big), \qquad \theta = \arccos(\rho).$$

If $ab = 0$ the average equals 0. In particular, with $(X, Y) = (G_i, G_j)$ this equals K_{ij}.

Proof Let (X, Y) be jointly Gaussian with mean zero, $\mathrm{Var}[X] = a$, $\mathrm{Var}[Y] = b$, and $Cov(X, Y) = c$. Set $\rho = c/\sqrt{ab} \in [-1, 1]$.

Write $X = \sqrt{a}U$ and $Y = \sqrt{b}V$, where (U, V) is standard bivariate normal with $\mathbb{E}[U] = \mathbb{E}[V] = 0$, $\mathrm{Var}[U] = \mathrm{Var}[V] = 1$, and $Cov(U, V) = \rho$. Then

$$\mathbb{E}\left[\mathrm{ReLU}(X)\mathrm{ReLU}(Y)\right] = \sqrt{ab}\mathbb{E}\left[\mathrm{ReLU}(U)\mathrm{ReLU}(V)\right].$$

So it suffices to compute $\mathbb{E}\left[\mathrm{ReLU}(U)\mathrm{ReLU}(V)\right]$.

Represent (U, V) as a correlated linear transform of independent standard normals:

$$U = Z_1, \qquad V = \rho Z_1 + \sqrt{1 - \rho^2}\, Z_2,$$

where $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Let $\varphi(t) = \max\{t, 0\}$. Using $\varphi(t) = t\mathbf{1}\{t > 0\}$,

$$\mathbb{E}\left[\varphi(U)\varphi(V)\right] = \mathbb{E}\left[Z_1(\rho Z_1 + \sqrt{1 - \rho^2}Z_2)\mathbf{1}\left\{Z_1 > 0, \rho Z_1 + \sqrt{1 - \rho^2}Z_2 > 0\right\}\right].$$

Now switch to polar coordinates for (Z_1, Z_2):

$$(Z_1, Z_2) = (R\cos\alpha,\ R\sin\alpha),$$

where $R \geq 0$ and $\alpha \in [0, 2\pi)$ are independent, α is uniform on $[0, 2\pi)$, and R has Rayleigh density $f_R(r) = re^{-r^2/2}$. Then

$$U = R\cos\alpha, \qquad V = R\big(\rho\cos\alpha + \sqrt{1 - \rho^2}\sin\alpha\big).$$

Let $\theta = \arccos\rho \in [0, \pi]$. Note the trigonometric identity

$$\rho\cos\alpha + \sqrt{1 - \rho^2}\sin\alpha = \cos(\alpha - \theta).$$

Hence

$$\varphi(U)\varphi(V) = R^2\cos\alpha\cos(\alpha - \theta)\mathbf{1}\left\{\cos\alpha > 0, \cos(\alpha - \theta) > 0\right\}.$$

Taking expectation and using independence of R and α gives

$$\mathbb{E}\left[\varphi(U)\varphi(V)\right] = \mathbb{E}\left[R^2\right]\mathbb{E}\left[\cos\alpha\cos(\alpha - \theta)\mathbf{1}\left\{\cos\alpha > 0, \cos(\alpha - \theta) > 0\right\}\right].$$

A short calculation with f_R yields $\mathbb{E}\left[R^2\right] = 2$.

It remains to compute the angular integral. Since α is uniform,

$$\mathbb{E}\left[\cdots\right] = \frac{1}{2\pi}\int_0^{2\pi} \cos\alpha\cos(\alpha-\theta)\mathbf{1}\left\{\cos\alpha > 0, \cos(\alpha-\theta) > 0\right\} d\alpha.$$

The constraints $\cos\alpha > 0$ and $\cos(\alpha-\theta) > 0$ define the intersection of two length-π intervals, whose overlap has length $\pi - \theta$. One convenient overlap is $\alpha \in (\theta - \pi/2,\ \pi/2)$. Therefore

$$\mathbb{E}\left[\cdots\right] = \frac{1}{2\pi}\int_{\theta-\pi/2}^{\pi/2} \cos\alpha\cos(\alpha-\theta)d\alpha.$$

Use $\cos\alpha\cos(\alpha-\theta) = \frac{1}{2}\big(\cos\theta + \cos(2\alpha-\theta)\big)$:

$$\begin{aligned}
\int_{\theta-\pi/2}^{\pi/2} \cos\alpha\cos(\alpha-\theta)d\alpha &= \frac{1}{2}\int_{\theta-\pi/2}^{\pi/2}\cos\theta\, d\alpha + \frac{1}{2}\int_{\theta-\pi/2}^{\pi/2}\cos(2\alpha-\theta)d\alpha \\
&= \frac{1}{2}\cos\theta(\pi-\theta) + \frac{1}{4}\Big[\sin(2\alpha-\theta)\Big]_{\theta-\pi/2}^{\pi/2} \\
&= \frac{1}{2}\cos\theta(\pi-\theta) + \frac{1}{4}\big(\sin(\pi-\theta) - \sin(\theta-\pi)\big) \\
&= \frac{1}{2}\cos\theta(\pi-\theta) + \frac{1}{2}\sin\theta.
\end{aligned}$$

Thus

$$\mathbb{E}\left[\cdots\right] = \frac{1}{2\pi}\cdot\frac{1}{2}\big(\sin\theta + (\pi-\theta)\cos\theta\big) = \frac{1}{4\pi}\big(\sin\theta + (\pi-\theta)\cos\theta\big).$$

Multiplying by $\mathbb{E}\left[R^2\right] = 2$ yields

$$\mathbb{E}\left[\varphi(U)\varphi(V)\right] = \frac{1}{2\pi}\big(\sin\theta + (\pi-\theta)\cos\theta\big).$$

Finally, substitute $\cos\theta = \rho$ and $\sin\theta = \sqrt{1-\rho^2}$ to get

$$\mathbb{E}\left[\varphi(U)\varphi(V)\right] = \frac{1}{2\pi}\Big(\sqrt{1-\rho^2} + (\pi - \arccos\rho)\rho\Big).$$

Scaling back gives

$$\mathbb{E}\left[\mathrm{ReLU}(X)\mathrm{ReLU}(Y)\right] = \sqrt{ab}\frac{1}{2\pi}\left(\sqrt{1-\rho^2} + (\pi-\arccos\rho)\rho\right), \qquad \rho = \frac{c}{\sqrt{ab}}.$$

■

Concentration using sub-exponential Bernstein. We now give an exponential deviation bound in N. Since ReLU of a Gaussian is sub-Gaussian by Lemma 11 and the product of two sub-Gaussians is sub-exponential by Lemma 27, a Bernstein inequality applies.

Lemma 11 (ReLU of a Gaussian is sub-Gaussian) *If $X \sim \mathcal{N}(0,a)$ then $\|\mathrm{ReLU}(X)\|_{\psi_2} \le C\sqrt{a}$ for an absolute constant C.*

Proof Since $0 \le \mathrm{ReLU}(X) \le |X|$, monotonicity of the Orlicz norm gives

$$\|\mathrm{ReLU}(X)\|_{\psi_2} \le \|X\|_{\psi_2}.$$

For $X \sim \mathcal{N}(0, a)$, Lemma 24 gives $\|X\|_{\psi_2} \le C\sqrt{a}$. ■

If $ab = 0$, then either X or Y is identically zero, so $\mathrm{ReLU}(X)\mathrm{ReLU}(Y) \equiv 0$ and the deviation probability is zero. Hence the following bound is stated for $ab > 0$.

Theorem 12 (Entrywise exponential concentration for ReLU) *Assume $ab > 0$. There exist absolute constants $c, C > 0$ such that for all $t \ge 0$,*

$$\mathbb{P}\left[\left|(Q_N^{(1)})_{ij} - K_{ij}\right| \ge t\right] \le 2\exp\left[-cN\min\left\{\frac{t^2}{C^2 ab}, \frac{t}{C\sqrt{ab}}\right\}\right].$$

Proof Let $Z_\alpha := \mathrm{ReLU}(h_\alpha(x_i))\mathrm{ReLU}(h_\alpha(x_j))$. Then (Z_α) are i.i.d. with mean K_{ij}. By Lemma 27, $Z_\alpha - K_{ij}$ has uniformly bounded ψ_1 norm of order $\sqrt{ab}$. Apply the standard Bernstein inequality for i.i.d. mean-zero sub-exponential variables. ■

Corollary 13 *Let $S := \max_{1\le k\le P} \sigma_1^2 \|x_k\|^2$ so that $ab \le S^2$ for all i, j. Then for all $t \ge 0$,*

$$\mathbb{P}\left[\max_{i,j}\left|(Q_N^{(1)})_{ij} - K_{ij}\right| \ge t\right] \le 2P^2\exp\left[-cN\min\left\{\frac{t^2}{C^2S^2}, \frac{t}{CS}\right\}\right],$$

and on the complementary event $\left\|Q_N^{(1)} - K\right\|_{\mathrm{op}} \le \left\|Q_N^{(1)} - K\right\|_F \le Pt$.

Appendix E. Computing $K_{ij} = \mathbb{E}\big[\varphi(G_i)\varphi(G_j)\big]$ for $\varphi(x) = x^p$

Fix an integer $p \ge 0$. Let $x_1, \dots, x_P \in \mathbb{R}^D$ be deterministic and set

$$Q_{ij}^{(0)} := x_i \cdot x_j, \qquad \Sigma := \sigma_1^2 Q^{(0)} \in \mathbb{R}^{P\times P}.$$

Let $G = (G_1, \dots, G_P) \sim \mathcal{N}(0, \Sigma)$. For indices i, j define

$$a := \Sigma_{ii} = \sigma_1^2\|x_i\|^2, \qquad b := \Sigma_{jj} = \sigma_1^2\|x_j\|^2, \qquad c := \Sigma_{ij} = \sigma_1^2(x_i^\top x_j).$$

Recall $K_{ij} := \mathbb{E}[\varphi(G_i)\varphi(G_j)] = \mathbb{E}[G_i^p G_j^p]$.

Theorem 14 (Closed form for $\mathbb{E}\left[G_i^p G_j^p\right]$ for a bivariate centered Gaussian) *Let (X, Y) be a centered bivariate Gaussian with $\mathbb{E}[X^2] = a$, $\mathbb{E}[Y^2] = b$, $\mathbb{E}[XY] = c$. Then for every integer $p \ge 0$,*

$$\mathbb{E}[X^pY^p] = \sum_{m=0}^{\lfloor p/2\rfloor} \frac{(p!)^2}{(p-2m)!2^{2m}(m!)^2} c^{p-2m}(ab)^m.$$

In particular, with $(X, Y) = (G_i, G_j)$, this equals K_{ij}.

Proof Let (X, Y) be as stated. Consider the moment generating function

$$M(s,t) \coloneqq \mathbb{E}\big[e^{sX+tY}\big] = \exp\Big(\frac{1}{2}(as^2 + 2cst + bt^2)\Big).$$

Since M is analytic, for integers $p \geq 0$,

$$\mathbb{E}[X^p Y^p] = \partial_s^p \partial_t^p M(s,t)\big|_{s=t=0}.$$

Write

$$M(s,t) = \exp\Big(\frac{1}{2}as^2\Big)\ \exp(cst)\ \exp\Big(\frac{1}{2}bt^2\Big) = \Big(\sum_{u\geq 0}\frac{(a/2)^u}{u!}s^{2u}\Big)\Big(\sum_{k\geq 0}\frac{c^k}{k!}s^k t^k\Big)\Big(\sum_{v\geq 0}\frac{(b/2)^v}{v!}t^{2v}\Big).$$

The coefficient of $s^p t^p$ comes from choices (u,k,v) such that $2u + k = p$ and $k + 2v = p$, hence $u = v =: m$ and $k = p - 2m$, with $m = 0, 1, \ldots, \lfloor p/2 \rfloor$. Thus the coefficient of $s^p t^p$ in M equals

$$\sum_{m=0}^{\lfloor p/2\rfloor}\frac{(a/2)^m}{m!}\cdot\frac{c^{p-2m}}{(p-2m)!}\cdot\frac{(b/2)^m}{m!} = \sum_{m=0}^{\lfloor p/2\rfloor}\frac{(ab)^m\, c^{p-2m}}{2^{2m}(m!)^2(p-2m)!}.$$

Since $\partial_s^p \partial_t^p$ multiplies the coefficient of $s^p t^p$ by $(p!)^2$, the stated identity follows. ■

A Bernstein inequality for concentration of these objects is developed in the next section.

Appendix F. The polynomial activation and the Orlicz norms

In the above, we only showed the convergence assuming well-behaved tails such as that of sub-Gaussian or sub-exponential distributions. However, not all nonlinear transformations of a Gaussian (which is $W^{(1)}X$ in our case) are generally sub-Gaussian or sub-exponential. For example, if G is Gaussian, G^p is (provably) not sub-Gaussian for $p > 1$. Such behavior can be controlled using the Orlicz norm and the corresponding Bernstein inequality, letting us get concentration even for fat-tailed distributions. Our treatment also includes motivation and intuition behind Orlicz norms.

F.1. Orlicz norms

Here we develop a self-contained Orlicz-tail framework tailored to random feature kernels and wide neural networks. A centered Gaussian satisfies a tail bound of the form $\mathbb{P}\left[|X| > t\right] \lesssim \exp\left\{-\Theta(t^2/\sigma^2)\right\}$ for $X \sim \mathcal{N}(0,\sigma)$. We use a scale parameter $\|\cdot\|$ on random variables that could replace σ. For general random variables X, with $\|X\| = 1$ (to be defined below), we ask for a function f such that $\mathbb{P}\left[|X| > t\right] \lesssim \exp\left\{-f(t)\right\}$. For such random variables, we would want the tail behavior to stay the same under constant scaling because one should be able to infer about X from $X/\|X\|$. But $\mathbb{P}\left[|cX| > t\right] = \mathbb{P}\left[|X| > t/|c|\right] \lesssim \exp\left\{-f(t/c)\right\} \forall c > 0$. So one imposes a restriction that the 'shape' of the exponent in the scaled and unscaled random variables be the same, thus asking for a constraint $K_a \coloneqq \lim_{t\to\infty}\dfrac{f(at)}{f(t)} < \infty$. Such a shape constraint determines, up to slowly varying factors, $f(t) = t^\alpha$, as stated in the following theorem:

Theorem 15 *Suppose $f : (0,\infty) \to (0,\infty)$ is a measurable function satisfying $K_a \coloneqq \lim_{t\to\infty} \frac{f(at)}{f(t)} < \infty \; \forall \; a > 0$. Then:*

1. *$K_{ab} = K_a K_b \forall a, b > 0$.*
2. *$\exists \alpha$ such that $K_a = a^\alpha \forall a > 0$.*
3. *The function $\ell(t) \coloneqq \frac{f(t)}{t^\alpha}$ is slowly varying, that is, $\lim_{t\to\infty} \frac{\ell(t)}{\ell(at)} = 1 \forall a > 0$.*

Proof

1. *Fix arbitrary $a, b > 0$. Then $K_{ab} = \lim_{t\to\infty} \frac{f(abt)}{f(t)} = \lim_{t\to\infty} \frac{f(abt)}{f(bt)} \cdot \frac{f(bt)}{f(t)}$. Both the limits exist individually, and the last expression can be written as a product of limits, the first of which is K_a and the latter K_b.*
2. *Consider $g(x) = \log K_{e^x}$. This changes our goal to equivalently prove that $g(x) = \alpha x, \; \forall x \in \mathbb{R}$. Let us first note that g is measurable. Indeed, since f is measurable and K is the pointwise limit of a ratio of two measurable functions, K is measurable; and since g is a composition of measurable functions, g must also be measurable. Next note that $g(x+y) = \log K_{e^{x+y}} = \log\left(K_{e^x} K_{e^y}\right) = g(x) + g(y)$. It is well known that the only solution to Cauchy's equation with a measurable function is $g(x) = \alpha x$ for some $\alpha \in \mathbb{R}$.*
3. *$\lim_{t\to\infty} \frac{\ell(t)}{\ell(at)} = \lim_{t\to\infty} \frac{a^\alpha t^\alpha f(t)}{t^\alpha f(at)} = a^\alpha / K_a = 1$.*

■

It is not uncommon to assume that, under scaling, the variables can be separated, that is, $f(at) = K_a f(t)$. This is only a mild additional constraint since ℓ above has slow variation. Under this assumption, we directly get

Corollary 16 *If $f(at) = K_a f(t) \; \forall a > 0, t > 0$ then $K_a = a^\alpha$ for some $\alpha \in \mathbb{R}$.*

Proof In the above, $\ell \equiv 1$, hence $K_a = a^\alpha$. ■

Now, let us move on to further investigating the structure of the Orlicz norms. A norm $\|\cdot\|$ on real-valued random variables should have the following desired properties:

- If $X \overset{D}{=} Y$ then $\|X\| = \|Y\|$, that is, two random variables coming from the same distribution should have the same norm.
- If $|X| \leq |Y|$ a.s. then $\|X\| \leq \|Y\|$, that is, monotonicity should be preserved.
- $\|X\| = \||X|\|$, that is, invariance under lattice symmetry.

Next, let us take a convex body-centric view of a norm. Every norm $\|\cdot\|$ has its corresponding unit ball $\{\|\cdot\| \leq 1\}$ which is symmetric and convex. Alternatively, given a symmetric convex body B with a nonempty interior, its corresponding norm is given by

$\|\cdot\|_B = \inf\{t > 0 \mid X \in tB\}$. Typically, convex bodies are the sublevel sets of convex functions. Let us say that $B = \{X \mid \rho(X) \leq 1\}$, where ρ satisfies:

$(A1)$ ρ is convex: $\rho(\lambda X + (1-\lambda)Y) \leq \lambda\rho(X) + (1-\lambda)\rho(Y)$.

$(A2)$ $\rho(0) = 0, \rho(X) = \rho(Y)$ if $X \overset{D}{=} Y$.

$(A3)$ $0 \leq X \leq Y$ a.s. $\implies \rho(X) \leq \rho(Y)$.

$(A4)$ For each X, $\lim_{t \downarrow 0} \rho(tX) = 0$.

$(A5)$ $\rho(X+Y) = \rho(X) + \rho(Y)$ if $XY = 0$ a.s.

$(A6)$ If $0 \leq X_n \uparrow X$ a.s. then $\rho(X) \leq \liminf_n \rho(X_n)$.

It can be shown that under the above assumptions, $\rho(X) = \mathbb{E}[\Phi(X)]$ for some Φ satisfying

$(B1)$ $\Phi(0) = 0$.

$(B2)$ Φ is non-decreasing.

$(B3)$ Φ is convex.

$(B4)$ $\lim_{t\to\infty} \Phi(t) = \infty$ (unless ρ is trivial).

Thus, our norm becomes $\|X\| = \inf\{t > 0 \mid \mathbb{E}[\Phi(X/t)] \leq 1\}$.

Next, we attempt to extend the proof of Hoeffding inequality to general random variables. Recall that the first step was to use the Markov inequality on e^{aX} to incorporate the moment generating function:

$$\mathbb{P}[X > t] = \mathbb{P}\left[e^{aX} \geq e^{at}\right] \leq \frac{\mathbb{E}[\exp\{aX\}]}{\exp\{at\}}.$$

Instead, the function $X \to aX$ could be replaced by any monotone function Φ such that $\Phi(0) = 0$. Let us use the above Young function Φ. Since we keep track of scaling, let us use $X \mapsto \Phi(aX)$. We also want two-sided tail bounds:

$$\mathbb{P}[|X| > t] = \mathbb{P}[\Phi(a|X|) \geq \Phi(at)] \leq \frac{\mathbb{E}[\Phi(a|X|)]}{\Phi(at)}, \quad \forall\, t \geq 0.$$

Next, we want to incorporate the norm and the tail bounds. Recall we wanted a norm that works well with the tail bound $e^{-f(t)}$. Say $\||X|\| = \|X\| = \inf\{s > 0 \mid \mathbb{E}[\Phi(|X|/s)] \leq 1\}$. By definition, this norm applies to scalars: $\|aX\| = a\|X\| \ \forall a > 0$. Consider a decreasing $\{K_n > 0\}$ such that $\mathbb{E}[\Phi(|X|/K_n)] \leq 1 \ \forall n \geq 1$ and $K_n \downarrow \|X\|$. Clearly, $K_n \geq \|X\|$. Since the above-mentioned tail bound (from the Markov inequality) holds for every $a > 0$, picking $a = \frac{1}{K_n}$ makes the RHS looks like $\dfrac{\mathbb{E}[\Phi(a|X|)]}{\Phi(at)} \leq \dfrac{1}{\Phi(t/K_n)}$. Let us now assume that Φ is also continuous, resulting in

$$\mathbb{P}[|X| > t] \leq \frac{1}{\Phi(t/\|X\|)} \quad \forall\, t \geq 0.$$

This motivates the following norm that imitates the sub-Gaussian parameter for the usual Hoeffding inequality. Thus, if we want the tail to be $\sim \exp\{-f(t/\|X\|)\}$, we obtain $\Phi(x) \simeq \exp f(x)$. By our earlier discussion, we know that for tails to have the same shape under scaling, we should choose $f(x) = x^\alpha$ for some $\alpha \in \mathbb{R}$ so $\Phi(s) \simeq \exp\{x^\alpha\}$. But recall the condition that $\lim_{s\to 0}\Phi(s) = 0$. This can be obtained by taking $\Phi(x) = \exp\{x^\alpha\} - 1$. Using this, we get

$$\mathbb{P}[|X| > t] \lesssim \exp\left\{-\left(\frac{t}{\|X\|}\right)^\alpha\right\}$$

where $\|X\| = \inf\left\{s > 0 \;\middle|\; \mathbb{E}\left[\exp\left(\left(\frac{|X|}{s}\right)^\alpha\right)\right] \le 2\right\}$. We often write this norm with the symbol $\|\cdot\|_{\psi_\alpha}$, where $\psi_\alpha(x) = \exp(x^\alpha) - 1$.

Definition 17 (Orlicz ψ_α norm) *For $\alpha > 0$ and a real random variable U, define*

$$\|U\|_{\psi_\alpha} := \inf\left\{s > 0 \;\middle|\; \mathbb{E}\left[e^{\left(\frac{|U|}{s}\right)^\alpha}\right] \le 2\right\}.$$

Definition 18 (Sub-Weibull) *A random variable U is called* sub-Weibull of order $\alpha > 0$ *if $\|U\|_{\psi_\alpha} < \infty$. Equivalently, there exists a constant $K > 0$ such that*

$$\mathbb{P}(|U| \ge t) \le 2\exp\Big(-(t/K)^\alpha\Big) \quad \forall t \ge 0,$$

with K comparable (up to absolute factors) to $\|U\|_{\psi_\alpha}$.

A random variable X satisfying $\|X\|_{\psi_\alpha} < \infty$ will be called a sub-Weibull(α). It is worth noting that sub-Weibull(2) is equivalent to a sub-Gaussian, and sub-Weibull(1) is equivalent to a sub-exponential random variable.

Equivalence of a distribution's tail bounds with moments and the MGF, such as Vershynin [22, Proposition 2.5.2], Wainwright [25, Theorem 2.6] for sub-Gaussianity or Vershynin [22, Proposition 2.7.1], Wainwright [25, Theorem 2.13] for sub-exponentiality, can also be derived for a sub-Weibull(α), reproducing the proofs line-by-line:

Proposition 19 (Vladimirova et al. [24, Theorem 2.1]) *Let X be a random variable. Then the following properties are equivalent for a random variable X and parameter $\alpha > 0$ with constants $K_{1,2,3,4}$ multiplicatively differing at most by a universal constant:*

1. *(tail bound) $\exists K_1 > 0$ such that $\mathbb{P}[|X| \ge t] \le 2\exp(-(t/K_1)^\alpha)\ \forall\, t \ge 0$.*
2. *(moment bound) $\exists K_2 > 0$ such that $\|X\|_{L_p} \le K_2 p^{\frac{1}{\alpha}}\ \forall\, p \ge 1$.*
3. *(power mgf bound) $\exists K_3 > 0$ such that $\mathbb{E}\left[\exp\left(\left(\frac{t|X|}{K_3}\right)^\alpha\right)\right] \le \exp(t^\alpha)\ \forall\, t \in (0, 1]$.*
4. *(norm quantification) $\exists K_4 > 0$ such that $\mathbb{E}\left[\exp\left(\left(\frac{|X|}{K_4}\right)^\alpha\right)\right] \le 2$.*

We note the following:

Lemma 20 $\left\|U^\beta\right\|_{\psi_\alpha} = \|U\|^\beta_{\psi_{\alpha\beta}}$ *for any $\alpha, \beta > 0$.*

Proof

$$\begin{aligned}\left\|U^\beta\right\|_{\psi_\alpha} &= \inf\left\{s>0 \;\middle|\; \mathbb{E}\left[\exp\left(\left|U^\beta\right|^\alpha/s^\alpha\right)\right]\le 2\right\}\\ &= \inf\left\{s>0 \;\middle|\; \mathbb{E}\left[\exp\left(\left(|U|/s^{\frac{1}{\beta}}\right)^{\beta\alpha}\right)\right]\le 2\right\}\\ &= \inf\left\{s^\beta>0 \;\middle|\; \mathbb{E}\left[\exp\left((|U|/s)^{\beta\alpha}\right)\right]\le 2\right\}\\ &= \left(\inf\left\{s>0 \;\middle|\; \mathbb{E}\left[\exp\left((|U|/s)^{\beta\alpha}\right)\right]\le 2\right\}\right)^\beta\\ &= \|U\|^\beta_{\psi_{\alpha\beta}}.\end{aligned}$$

■

One important point to note is that $\|\cdot\|_{\psi_\alpha}$ for $\alpha>0$ is a norm iff $\alpha\ge 1$. It can be easily seen that $\|X\|_{\psi_\alpha}$ is always non-negative non-degenerate (that is, positive if X is nonzero on a set of positive measure) and is compatible with scaling (that is, $\|cX\|_{\psi_\alpha}=|c|\,\|X\|_{\psi_\alpha}$) for any $\alpha>0$. What fails is the triangle inequality. This is precisely because the unit norm ball is not convex for $\alpha<1$. However we still have the following 'scaled' triangle inequality:

Lemma 21 *Fix $\alpha>0$ and let $\|\cdot\|$ denote $\|\cdot\|_{\psi_\alpha}$. Then*

$$\|U+V\|\le \max\left\{2^{\frac{1}{\alpha}-1},1\right\}\left(\|U\|+\|V\|\right).$$

Proof Begin with the case $\alpha\ge 1$. Start by noting that the function $x\mapsto e^{x^\alpha}$ is convex. Let $\{u_n\},\{v_n\}$ be decreasing sequences satisfying $\mathbb{E}\left[e^{\left(\frac{|U|}{u_n}\right)^\alpha}\right]\le 2, \mathbb{E}\left[e^{\left(\frac{|V|}{v_n}\right)^\alpha}\right]\le 2\ \forall\ n$ and $u_n\to\|U\|, v_n\to\|V\|$. Then $\exp\left(\left(\frac{|U+V|}{u_n+v_n}\right)^\alpha\right)\le \exp\left(\left(\frac{|U|+|V|}{u_n+v_n}\right)^\alpha\right)=\exp\left(\left(\frac{|U|}{u_n}\frac{u_n}{u_n+v_n}+\frac{|V|}{v_n}\frac{v_n}{u_n+v_n}\right)^\alpha\right)\overset{\text{Jensen}}{\le}\frac{u_n}{u_n+v_n}\exp\left(\left(\frac{|U|}{u_n}\right)^\alpha\right)+\frac{v_n}{u_n+v_n}\exp\left(\left(\frac{|V|}{v_n}\right)^\alpha\right)$.

Taking expectations,

$$\mathbb{E}\left[\exp\left(\left(\frac{|U+V|}{u_n+v_n}\right)^\alpha\right)\right]\le\frac{u_n}{u_n+v_n}\mathbb{E}\left[\exp\left(\left(\frac{|U|}{u_n}\right)^\alpha\right)\right]+\frac{v_n}{u_n+v_n}\mathbb{E}\left[\exp\left(\left(\frac{|V|}{v_n}\right)^\alpha\right)\right]\le 2.$$

But $u_n+v_n\downarrow u+v$. Hence, $\|U+V\|\le u+v$.

Now suppose that $\alpha<1$ and let $p:=1/\alpha$. Then $\frac{1}{p^*}=1-\frac{1}{p}=1-\alpha$. Note that $\|U+V\|^\alpha_{\psi_\alpha}=\||U+V|^\alpha\|_{\psi_1}\le\||U|^\alpha+|V|^\alpha\|_{\psi_1}\le\|U^\alpha\|_{\psi_1}+\|V^\alpha\|_{\psi_1}=\|U\|^\alpha_{\psi_\alpha}+\|V\|^\alpha_{\psi_\alpha}$. Here we used Lemma 20 twice in the first and last equalities. The second step (inequality) follows from the fact that $x\mapsto x^\alpha$ is concave for $\alpha<1$. By Hölder's inequality with the pair (p,p^*), $\|U\|^\alpha_{\psi_\alpha}+\|V\|^\alpha_{\psi_\alpha}\le\left(\|U\|_{\psi_\alpha}+\|V\|_{\psi_\alpha}\right)^\alpha 2^{1-\alpha}$. Combining everything, we obtain

$$\|U+V\|_{\psi_\alpha}\le 2^{\frac{1}{\alpha}-1}\left(\|U\|_{\psi_\alpha}+\|V\|_{\psi_\alpha}\right).$$

■

Proposition 22 $|\mathbb{E}[U]| \le \dfrac{2\|U\|_{\psi_\alpha}\Gamma\left(\frac{1}{\alpha}\right)}{\alpha}$. *Hence* $\|\mathbb{E}[U]\|_{\psi_\alpha} \le \frac{2\|U\|_{\psi_\alpha}\Gamma\left(\frac{1}{\alpha}\right)}{\alpha(\ln 2)^{\frac{1}{\alpha}}}$.

Proof Let $u = \|U\|_{\psi_\alpha}$. By Markov $\mathbb{P}[|U| > t] = \mathbb{P}\left[e^{\left(\frac{|U|}{u}\right)^\alpha} > e^{\left(\frac{t}{u}\right)^\alpha}\right] \le 2e^{-\left(\frac{t}{u}\right)^\alpha}$. So

$$\begin{aligned}
|\mathbb{E}[U]| \le \mathbb{E}[|U|] &= \int_0^\infty \mathbb{P}[|U| > t]\,\mathrm{d}\,t \\
&\le \int_0^\infty 2e^{-\left(\frac{t}{u}\right)^\alpha}\,\mathrm{d}\,t \\
&= \frac{2u}{\alpha}\int_0^\infty e^{-s}s^{\frac{1}{\alpha}-1}\,\mathrm{d}\,s \qquad \left[s = \left(\frac{t}{u}\right)^\alpha \implies \mathrm{d}\,t = \frac{us^{\frac{1}{\alpha}-1}\,\mathrm{d}\,s}{\alpha}\right] \\
&= \frac{2u\Gamma\left(\frac{1}{\alpha}\right)}{\alpha} = \frac{2\|U\|_{\psi_\alpha}\Gamma\left(\frac{1}{\alpha}\right)}{\alpha}
\end{aligned}$$

The bound on $\|\mathbb{E}[U]\|_{\psi_\alpha}$ follows from $\|a\|_{\psi_\alpha} = |a|(\ln 2)^{-\frac{1}{\alpha}}\ \forall\, a \in \mathbb{R}$. ■

Lemma 23 *For* $\alpha > 0$, $\|U - \mathbb{E}[U]\|_{\psi_\alpha} \le \begin{cases} 2\|U\|_{\psi_\alpha} & \text{if } \alpha \ge 1 \\ 2^{\frac{1}{\alpha}-1}\left(1 + \frac{2\Gamma\left(\frac{1}{\alpha}\right)}{\alpha(\log 2)^{\frac{1}{\alpha}}}\right)\|U\|_{\psi_\alpha} & \text{if } 0 < \alpha < 1 \end{cases}$.

Proof Suppress the ψ_α and simply write $\|\cdot\|$ for $\|\cdot\|_{\psi_\alpha}$.

Suppose $\alpha \ge 1$. Let $\{u_n\}$ be a sequence satisfying

$$\mathbb{E}[\exp]\left(\left(\frac{|U|}{u_n}\right)^\alpha\right) \le 2 \qquad \text{for all } n,$$

and $u_n \downarrow \|U\|$. Since $x \mapsto \exp(|x|^\alpha)$ is convex for $\alpha \ge 1$, Jensen's inequality gives

$$\exp\left(\left(\frac{|\mathbb{E}[U]|}{u_n}\right)^\alpha\right) \le \mathbb{E}[\exp]\left(\left(\frac{|U|}{u_n}\right)^\alpha\right) \le 2.$$

Hence $\|\mathbb{E}[U]\| \le \|U\|$.

Suppose $\alpha \in (0,1)$. $\|\mathbb{E}[U]\| \le \frac{2\Gamma(1/\alpha)}{\alpha(\log 2)^{\frac{1}{\alpha}}}\|U\|$ by Proposition 22. Therefore, $\|U - \mathbb{E}[U]\| \overset{\text{Lemma 21}}{\le} 2^{\frac{1}{\alpha}-1}(\|U\| + \|\mathbb{E}[U]\|) \le 2^{\frac{1}{\alpha}-1}\left(1 + \frac{2\Gamma\left(\frac{1}{\alpha}\right)}{\alpha(\log 2)^{\frac{1}{\alpha}}}\right)\|U\|$. ■

F.2. Stronger exponential tails for $\varphi(x) = x^p$

Throughout, $p \in \mathbb{Z}_{\ge 1}$ is fixed. Fix indices $i, j \in \{1, \dots, P\}$ and write

$$a \coloneqq \Sigma_{ii} = \sigma_1^2\|x_i\|^2, \qquad b \coloneqq \Sigma_{jj} = \sigma_1^2\|x_j\|^2, \qquad c \coloneqq \Sigma_{ij} = \sigma_1^2(x_i \cdot x_j), \qquad \Sigma_{(i,j)} \coloneqq \begin{pmatrix} a & c \\ c & b \end{pmatrix}.$$

Let $(X, Y) \sim \mathcal{N}(0, \Sigma_{(i,j)})$ and define

$$K_{ij} \coloneqq \mathbb{E}[X^p Y^p] \qquad \text{and} \qquad (Q_N^{(1)})_{ij} \coloneqq \frac{1}{N} \sum_{\alpha=1}^{N} X_\alpha^p Y_\alpha^p,$$

where (X_α, Y_α) are i.i.d. copies of (X, Y).

The goal is a *subexponential* (in N) deviation bound for $(Q_N^{(1)})_{ij} - K_{ij}$.

F.2.1. Gaussian inputs imply sub-Weibull tails for polynomials

We first state well-known facts for Gaussian variables.

Lemma 24 (Gaussian ψ_2 norm) *If $Z \sim \mathcal{N}(0, v)$, then $\|Z\|_{\psi_2} \le C\sqrt{v}$ for an absolute constant $C > 0$. Consequently, $\mathbb{P}(|Z| \ge t) \le 2\exp(-c\,t^2/v)$ for an absolute $c > 0$.*

Proof For $Z = \sqrt{v}\,G$ with $G \sim \mathcal{N}(0,1)$, it suffices to bound $\|G\|_{\psi_2}$ by a universal constant. Using $\mathbb{E}e^{G^2/C^2} = (1 - 2/C^2)^{-1/2}$ for $C > \sqrt{2}$ (chi-squared mgf), choose C large so that $\mathbb{E}\exp(G^2/C^2) \le 2$, yielding $\|G\|_{\psi_2} \le C$. The tail bound follows from the standard equivalence between ψ_2 and sub-Gaussian tails. ■

Proposition 25 *If $\alpha, \beta > 0$ then $\|U^\alpha\|_{\psi_\beta} = \|U\|_{\psi_{\alpha\beta}}^{\alpha}$.*

Proof $\|U^\alpha\|_{\psi_\beta} = \inf\left\{ s > 0 \;\middle|\; \mathbb{E}e^{\left(\frac{|U^\alpha|}{s}\right)^\beta} \le 2 \right\} = \inf\left\{ s^\alpha > 0 \;\middle|\; \mathbb{E}e^{\left(\frac{|U|}{s}\right)^{\alpha\beta}} \le 2 \right\} = \|U\|_{\psi_{\alpha\beta}}^{\alpha}$. ■

Lemma 26 (Powers of sub-Gaussians are sub-Weibull) *If U is sub-Gaussian, i.e. $\|U\|_{\psi_2} \le L$, then for every real $m \ge 1$, we have $\|U^m\|_{\psi_{2/m}} = \|U\|_{\psi_2}^m \le L^m$.*

Proof Use Proposition 25. ■

Lemma 27 (Product of two sub-Gaussians is sub-exponential) *If $\|U\|_{\psi_2} \le L_1$ and $\|V\|_{\psi_2} \le L_2$, then*

$$\|UV\|_{\psi_1} \le \Gamma L_1 L_2$$

for an absolute constant $\Gamma > 0$.

Proof By homogeneity, it suffices to consider the case $L_1 = L_2 = 1$. Using $|uv| \le (u^2+v^2)/2$, we have

$$\exp\left(\frac{|UV|}{\Gamma}\right) \le \exp\left(\frac{U^2}{2\Gamma}\right) \exp\left(\frac{V^2}{2\Gamma}\right).$$

By Cauchy–Schwarz,

$$\mathbb{E}\left[\exp\right]\left(\frac{|UV|}{\Gamma}\right) \le \left(\mathbb{E}\left[\exp\right]\left(\frac{U^2}{\Gamma}\right)\right)^{1/2} \left(\mathbb{E}\left[\exp\right]\left(\frac{V^2}{\Gamma}\right)\right)^{1/2}.$$

Choosing $\Gamma > 0$ to be sufficiently large, depending only on the absolute constants in the ψ_2 bounds, makes the right-hand side at most 2. Hence $\|UV\|_{\psi_1} \le \Gamma L_1 L_2$. ■

Lemma 28 (Degree-$2p$ Gaussian monomial is sub-Weibull of order $1/p$) *Let (X, Y) be any centered bivariate Gaussian. Then*

$$\|X^p Y^p\|_{\psi_{1/p}} \le \Gamma^p \|X\|_{\psi_2}^p \|Y\|_{\psi_2}^p ,$$

where $\Gamma > 0$ is from Lemma 27.

Proof $\|(XY)^p\|_{\psi_{1/p}} \overset{\text{Proposition 25}}{=} \|XY\|_{\psi_1}^p \overset{\text{Lemma 27}}{\le} \Gamma^p \|X\|_{\psi_2}^p \|Y\|_{\psi_2}^p .$ ■

Corollary 29 (Explicit scale in terms of covariance) *Let (X, Y) be a zero-mean bivariate Gaussian with $Var[X] = a, Var[Y] = b$. Then $\|X\|_{\psi_2} \le C\sqrt{a}, \|Y\|_{\psi_2} \le C\sqrt{b}$, hence*

$$\|X^p Y^p\|_{\psi_{1/p}} \le \widetilde{C}_p (ab)^{p/2}.$$

Moreover, by Lemma 23, there is a constant $C'_p > 0$, depending only on p, such that

$$\|X^p Y^p - \mathbb{E}[X^p Y^p]\|_{\psi_{1/p}} \le C'_p (ab)^{p/2}.$$

Proof Lemma 24 gives

$$\|X\|_{\psi_2} \le C\sqrt{a}, \qquad \|Y\|_{\psi_2} \le C\sqrt{b}.$$

Combining this with Lemma 28, we obtain

$$\|X^p Y^p\|_{\psi_{1/p}} \le C_p (ab)^{p/2},$$

where $C_p > 0$ depends only on p. Applying Lemma 23 with $\alpha = 1/p$ gives another constant $C'_p > 0$, depending only on p, such that

$$\|X^p Y^p - \mathbb{E}[X^p Y^p]\|_{\psi_{1/p}} \le C'_p (ab)^{p/2}.$$

This proves the claim. ■

F.3. Deviation for averages of i.i.d. sub-Weibull variables

There is a Bernstein inequality for the subWeibull type tails, just like the Hoeffdin inequality and the sub-Exponential Bernstein inequality. Here is a simplified version of [15, Theorem 3.1].

Theorem 30 (Sub-Weibull Bernstein inequality) *Let $\alpha > 0$ and let $U_1, \dots, U_N$ be independent mean-zero random variables with*

$$\|U_k\|_{\psi_\alpha} \le K \quad \text{for all } k.$$

Then there exists a constant $c_\alpha > 0$ depending only on α such that for all $t \ge 0$,

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{k=1}^{N} U_k\right| \ge t\right] \le 2\exp\left(-c_\alpha \min\left\{\frac{Nt^2}{K^2}, \left(\frac{Nt}{K}\right)^\alpha\right\}\right).$$

Remark. The two regimes correspond to "small deviations" (Gaussian-type, t^2) and "large deviations" (Weibull-type, t^α). For $\alpha = 1$ this reduces (up to constants) to the standard sub-exponential Bernstein inequality.

F.4. Proof of Theorem 1

Let

$$Z_k := \varphi(X_k)\varphi(Y_k) - \mathbb{E}\left[\varphi(X)\varphi(Y)\right].$$

By assumption, the variables $Z_1, \dots, Z_N$ are independent, mean-zero, and satisfy

$$\|Z_k\|_{\psi_\alpha} \le K \qquad \text{for all } k.$$

Applying Theorem 30 gives, for every $t \ge 0$,

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{k=1}^{N} Z_k\right| \ge t\right] \le 2\exp\left(-c_\alpha \min\left\{\frac{Nt^2}{K^2}, \left(\frac{Nt}{K}\right)^{\alpha}\right\}\right).$$

Since

$$\frac{1}{N}\sum_{k=1}^{N} Z_k = \frac{1}{N}\sum_{k=1}^{N} \varphi(X_k)\varphi(Y_k) - \mathbb{E}\left[\varphi(X)\varphi(Y)\right],$$

this is exactly the claimed bound.

D. Application to X^pY^p and to $(Q_N^{(1)})_{ij}$

Define mean-zero i.i.d. variables

$$U_\alpha := X_\alpha^p Y_\alpha^p - \mathbb{E}[X^pY^p] = X_\alpha^p Y_\alpha^p - K_{ij}, \qquad \alpha = 1, \dots, N.$$

By Corollary 29,

$$\|U_\alpha\|_{\psi_{1/p}} \le K_{p,i,j} := 2^p\left(1 + \frac{2p!}{(\log 2)^p}\right)\Gamma^p C^2 \sqrt{a^p b^p}.$$

Therefore Theorem 30 applies with $\alpha = 1/p$.

Theorem 31 (Exponential tail for each kernel entry) *Fix i, j and integer $p \ge 1$. There exist constants $c_p, C_p > 0$ depending only on p such that for all $t \ge 0$,*

$$\mathbb{P}\left[\left|(Q_N^{(1)})_{ij} - K_{ij}\right| \ge t\right] \le 2\exp\left[-c_p \min\left\{N\left(\frac{t}{C_p(ab)^{p/2}}\right)^2, \left(\frac{Nt}{C_p(ab)^{p/2}}\right)^{1/p}\right\}\right].$$

Proof For $ab > 0$, Corollary 29 gives

$$\|X^pY^p - \mathbb{E}\left[X^pY^p\right]\|_{\psi_{1/p}} \le C_p(ab)^{p/2}.$$

Apply Theorem 30 with $\alpha = 1/p$ and $K = C_p(ab)^{p/2}$. This gives

$$\mathbb{P}\left[\left|(Q_N^{(1)})_{ij} - K_{ij}\right| \ge t\right] \le 2\exp\left(-c_p \min\left\{N\left(\frac{t}{C_p(ab)^{p/2}}\right)^2, \left(\frac{Nt}{C_p(ab)^{p/2}}\right)^{1/p}\right\}\right),$$

after adjusting constants. ∎

Corollary 32 *Let*

$$S_p := \max_{1\le i,j\le P}(ab)^{p/2} = \max_{1\le i,j\le P}\left(\sigma_1^4\,\|x_i\|^2\,\|x_j\|^2\right)^{p/2} = \sigma_1^{2p}\Big(\max_{1\le i\le P}\|x_i\|^2\Big)^p.$$

Then there exist constants $c_p, C_p > 0$ depending only on p such that for all $t \ge 0$,

$$\mathbb{P}\left[\max_{i,j}\left|(Q_N^{(1)})_{ij} - K_{ij}\right| \ge t\right] \le 2P^2 \exp\left[-c_p \min\left\{N\left(\frac{t}{C_pS_p}\right)^2, \left(\frac{Nt}{C_pS_p}\right)^{1/p}\right\}\right].$$

On the complementary event,

$$\left\|Q_N^{(1)} - K\right\|_{\mathrm{op}} \le \left\|Q_N^{(1)} - K\right\|_F \le Pt.$$

Proof Apply Theorem 31 to each pair (i,j) and use the union bound over the P^2 entries. Since

$$(ab)^{p/2} \le S_p$$

for every pair (i,j), the stated probability bound follows after adjusting constants. On the complementary event, every entry of $Q_N^{(1)} - K$ has absolute value at most t. Hence

$$\left\|Q_N^{(1)} - K\right\|_F \le Pt, \qquad \left\|Q_N^{(1)} - K\right\|_{\mathrm{op}} \le \left\|Q_N^{(1)} - K\right\|_F.$$

■