Scaling Laws from Sequential Feature Recovery: A Solvable Hierarchical Model

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

We propose a simple mechanism by which scaling laws emerge from feature learning in multi-layer networks. We study a high-dimensional hierarchical target that is a globally high-degree function, but that can be represented by a combination of latent compositional features whose weights decrease as a power law. We show that a layer-wise spectral algorithm adapted to this compositional structure achieves improved scaling relative to shallow, non-adaptive methods, and recovers the latent directions sequentially: strong features become detectable at small sample sizes, while weaker features require more data. We prove sharp feature-wise recovery thresholds and show that aggregating these transitions yields an explicit power-law decay of the prediction error. Technically, the analysis relies on random matrix methods and a resolvent-based perturbation argument, which gives matching upper and lower bounds for individual eigenvector recovery beyond what standard gap-based perturbation bounds provide. Numerical experiments confirm the predicted sequential recovery, finite-size smoothing of the thresholds, and separation from non-hierarchical kernel baselines. Together, these results show how smooth scaling laws can emerge from a cascade of sharp feature-learning transitions.

1. Introduction

Despite the empirical success of neural networks, we still lack a predictive theory answering a deceptively simple question: given a structured learning problem, which features are learned first, and how does their sequential discovery translate into statistical efficiency? This question lies at the intersection of three active lines of research. First, neural scaling laws suggest that the performance of large models follows power laws in data, compute, or model size [4, 14, 34, 37]. Yet most mathematical theories rely on linearized, kernel, or random-feature models, where the relevant representation is fixed in advance and learning is controlled by the spectrum of this representation [10, 16, 19, 20, 26, 57]. Second, many works have emphasized that feature learning is not necessarily smooth: training can exhibit plateaus, abrupt drops in risk, and the sequential emergence of features or concepts [28, 54–56, 64]. Third, recent theory has begun to isolate the computational advantage of depth in compositional tasks, where deeper architectures can discover intermediate representations inaccessible to shallow methods [15, 22, 31, 32, 47, 58, 63].

This paper asks whether scaling laws can arise not from a fixed spectral bias, but from the progressive uncovering of the relevant features in the data, as happen in deep neural nets. We investigate a mathematically tractable high-dimensional task which requires uncovering hidden features in multiple layers, which are combined together, but carry different weights. Statistically detecting an individual feature requires a minimum sample size which is proportional to its weight.

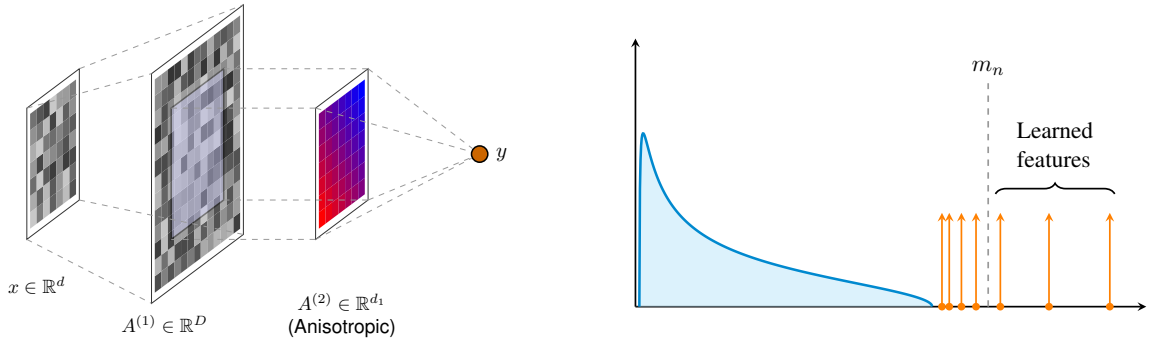

Figure 1: (**Left**) Illustration of the compositional function defined in Equation (3.3) and studied throughout this paper, where the target is given by an anisotropic combination of high-degree features $\mathrm{He}_q(x)$ of the input data $x \in \mathbb{R}^d$. (**Right**) The key conceptual idea in our proof is to show that the relevant features of the target can be efficiently learned by a spectral method adapted to the compositional structure of the target. Its spectrum is composed of a bulk (blue), representing the noise, and spikes (orange), representing the signal, which is only resolved up to a scale m_n depending on the sample size n. The MSE is then dominated by directions $i > m_n$ which are not learned.

Strong features are therefore learned first, weaker features later, and the prediction error is governed by the tail of the hidden spectrum that has not yet been recovered. Solving the task efficiently requires untangling the compositional structure. This combination of compositional and hierarchical features lead to different scaling in the performance of predictors which are adapted (or not) to the task geometry.

The key technical idea in our analysis is to view feature learning as a sequence of spectral transitions: as the sample size grows, progressively weaker hidden features become resolvable, see Figure 1 (right). We combine recent spectral methods for compositional targets [31, 47, 58, 63] with the scaling-law viewpoint of power-law feature strengths [28, 29, 54]. Technically, the proof controls empirical Hermite moment matrices and the alignment of their outlier eigenvectors. Since adjacent power-law spikes have shrinking gaps, Davis–Kahan-type bounds [24] are too coarse for sharp feature-wise thresholds; we instead use a resolvent-based perturbation expansion in the spirit of [30, 33]. This yields matching recovery and non-recovery at the scale $n \asymp D/a_i^2$, where a_i is the i-th feature weight.

Our contributions are threefold. First, we introduce a high-dimensional task combining hierarchical structure with a power-law spectrum of latent feature strengths. Second, we prove sharp sample-complexity thresholds for recovering individual latent directions by a hierarchy-aware spectral algorithm, and show that aggregating these transitions yields an explicit scaling law for the prediction error. Third, we provide experiments confirming the predicted recovery transitions, finite-size smoothing, and separation from shallow kernel baselines.

2. Related work and positioning

Hierarchy, multi-index structure, and spectral recovery. The role of depth in exploiting compositional structure has been studied through depth-separation and compositional-target models [23, 44, 46, 53, 59], random hierarchy and high-dimensional hierarchical targets [15, 22, 32], hi-

erarchical polynomial targets and three-layer feature learning [31, 47, 63], and the hierarchical spectral method of Tabanelli et al. [58]. Our model also connects to multi-index learning, including statistical-computational gaps, weak recovery, and limitations of kernel methods [1, 3, 5, 6, 8, 21, 60], as well as spectral estimators for low-dimensional structure in Gaussian models [27, 38, 39, 41, 45]. We depart from these settings by adding an anisotropic power-law spectrum over the latent features and by proving matching upper and lower thresholds for individual feature recovery.

Scaling laws, power-law spectra, and polynomial features. Most theoretical scaling-law analyses use fixed representations, such as kernels or random features, where generalization is controlled by the spectrum of the associated feature map [2, 4, 10, 11, 16, 19, 20, 26, 42, 51, 57]; another line studies how trainable-parameter scaling affects optimization, initialization, and expressivity [12, 17, 18, 67]. More recent works show that scaling laws can arise from feature learning itself in quadratic or shallow neural-network models [7, 9, 27, 29, 54], with Defilippis et al. [27, 29] closest to our sequential-recovery picture. We show that analogous rates arise in a genuinely multi-layer hierarchical setting. Technically, our proof is related to Gaussian equivalence and universality for polynomial features, random feature matrices, and high-dimensional kernels [35, 40, 65, 66]; however, rather than replacing Hermite features by Gaussian surrogates, we keep the Hermite structure explicit and use Wiener-chaos tools [48–50].

3. Model and spectral estimator

Let $x \sim \mathcal{N}(0, I_d)$ and let $F(x) = \mathcal{F}[\mathrm{He}_q(x)] \in \mathbb{R}^D$ be the flattened normalized degree-q Hermite feature vector, with $D = B(d,q) = \binom{d+q-1}{q} \asymp d^q$. We take $d_1 = \lfloor d^\varepsilon \rfloor$ hidden directions $A_i^{(1)} \in \mathbb{R}^D$ with independent entries of variance d^{-q}, and define first-layer features

$$h_i^{(1)}(x) = \langle A_i^{(1)}, F(x) \rangle, \qquad i \in [d_1]. \tag{3.1}$$

The second layer is diagonal and anisotropic. For

$$\lambda_i = Z_\gamma z_i i^{-\gamma}, \qquad z_i \sim \mathrm{Rad}(1/2), \qquad Z_\gamma = \Big(\sum_{i=1}^{d_1} i^{-2\gamma} \Big)^{-1/2}, \tag{3.2}$$

we set

$$h^{(2)}(x) = \frac{1}{\sqrt{2}} \sum_{i=1}^{d_1} \lambda_i \Big((h_i^{(1)}(x))^2 - 1 \Big), \qquad y = f_\star(x) = g(h^{(2)}(x)). \tag{3.3}$$

The normalization keeps the signal variance of order one. Although $f_\star$ is a degree-$2q$ function of x when $g(t) = t$, its representation is built from degree-q intermediate features. Orthogonally invariant shallow kernels therefore see a high-degree target and require the d^{2q} scale, while a hierarchy-aware learner can search for the degree-q representation at the $D \asymp d^q$ scale [43, 58].

Given samples $(x_\mu, y_\mu)_{\mu=1}^n$, the first step forms

$$\widehat{C} = \frac{1}{n} \sum_{\mu=1}^{n} y_\mu \, \mathrm{He}_2(F(x_\mu)) \in \mathbb{R}^{D \times D} \tag{3.4}$$

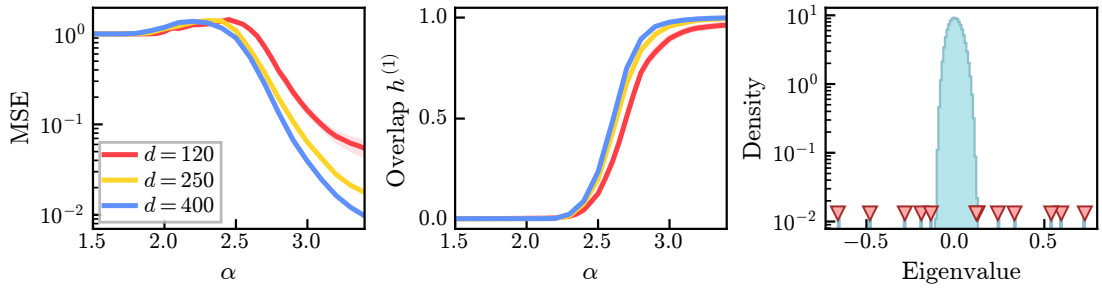

Figure 2: **Recovery transition for the hierarchical spectral estimator.** Parameters: $q = 2$, $\varepsilon = 0.5$, $\gamma = 0.4$, $g^\star = \mathrm{id}$. Left: test MSE versus $\alpha = \log(n)/\log(d)$. Center: first-layer feature overlap $q_h^{(1)}$ versus α. Right: spectrum of $\widehat{C}$ for $d = 140$ and $\alpha = 3.5$; red markers indicate the leading empirical spikes.

and uses its leading eigenvectors as estimates of the hidden directions $A_i^{(1)}$. The learned features are then $\widehat{h}_{\mu,i}^{(1)} = \langle \widehat{A}_i^{(1)}, F(x_\mu)\rangle$. The second layer is estimated from the low-dimensional moment

$$\widehat{A}^{(2)} = \frac{1}{n}\sum_{\mu=1}^{n} y_\mu \,\mathrm{He}_2(\widehat{h}_\mu^{(1)}) \in \mathbb{R}^{d_1\times d_1}, \tag{3.5}$$

and the final one-dimensional readout g is fit by ridge regression on $\widehat{h}_\mu^{(2)} = \langle \widehat{A}^{(2)}, \mathrm{He}_2(\widehat{h}_\mu^{(1)})\rangle$. The full pipeline is stated in Algorithm 1 in the appendix.

The signal-to-noise prediction is immediate. The population moment is aligned with the hidden subspace,

$$\mathbb{E}[\widehat{C}] \simeq \nu_1 A^{(1)\top} A^{(2)} A^{(1)}, \qquad \nu_1 = \mathbb{E}[g'(Z)], \tag{3.6}$$

whereas the empirical fluctuation has operator scale $\sqrt{D/n}$ in the orthogonal complement. Thus the i-th spike is detectable when $|\lambda_i| \gtrsim \sqrt{D/n}$, i.e.

$$n_i \asymp \frac{D\, i^{2\gamma}}{Z_\gamma^2}. \tag{3.7}$$

The theorem below makes this threshold sharp and converts it into a prediction-error rate.

4. Main result and numerical evidence

We work in either of two readout regimes: (i) $g(t) = t$ with $\gamma > 0$, or (ii) $0 < \gamma < 1/2$ with g centered, Lipschitz, and of information exponent one, $\mathbb{E}[g'(Z)] \neq 0$ for $Z \sim \mathcal{N}(0,1)$. The appendix states the technical growth conditions on ε and the full proof.

Theorem 1 (Sequential recovery and induced scaling) *Let $u_k = A_k^{(1)}/\|A_k^{(1)}\|$ and let $\widehat{u}_k$ be the empirical eigenvector associated with the k-th population spike. Under the model above and either readout regime, the k-th latent direction is recovered at the sharp sample scale $n_k \asymp Dk^{2\gamma}/Z_\gamma^2$. More precisely, for each $k \in [d_1]$,*

$$|\langle \widehat{u}_k, u_k\rangle| = 1 - O_d\!\left(\frac{D\,k^{2\gamma}}{nZ_\gamma^2}\right) \quad \text{if } n = \omega_d(Dk^{2\gamma}Z_\gamma^{-2}), \tag{4.1}$$

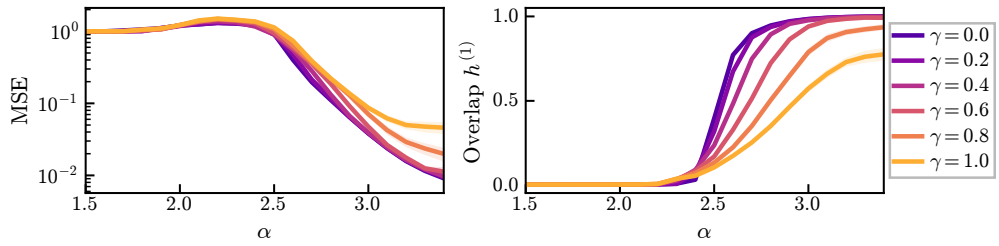

Figure 3: **Dependence on the power-law exponent.** Parameters: $d = 400$, $q = 2$, $\varepsilon = 0.5$, $g^\star = \mathrm{id}$. Larger γ makes the first directions stronger but pushes the weak tail to larger samples, spreading recovery across a wider interval of α.

whereas no weak recovery is possible below this scale, e.g. at $n = \Theta(Dk^{2\gamma} Z_\gamma^{-2} d^{-\delta})$ *for any fixed* $\delta > 0$. *Consequently, after* n *samples the algorithm recovers on the order of* $\min\{d_1, (Z_\gamma^2 n/D)^{1/(2\gamma)}\}$ *directions. Once the final low-dimensional fit is not the bottleneck, the test error is governed by the remaining power-law tail and satisfies*

$$\mathrm{MSE}(n) = \begin{cases} \Theta_d(1) - \Theta_d\left[\left(\dfrac{n}{Dd_1}\right)^{\frac{1}{2\gamma}-1}\right], & 0 < \gamma < \frac{1}{2}, \quad D \ll n \ll Dd_1, \\ \Theta_d\left[\left(\dfrac{n}{D}\right)^{-1+\frac{1}{2\gamma}}\right], & \gamma > \frac{1}{2} \quad \textit{in the identity-readout regime.} \end{cases} \tag{4.2}$$

The proof cannot rely only on a black-box Davis–Kahan argument because the adjacent power-law spikes have shrinking gaps. The key step is a resolvent expansion of the empirical eigenvectors,

$$\widehat{u}_k = u_k + \sum_{j \neq k} \frac{u_j^\top \Delta u_k}{\lambda_k - \lambda_j} u_j + \frac{1}{\lambda_k} P_{\ker(\mathbb{E}\widehat{C})} \Delta u_k + \text{higher order}, \qquad \Delta = \widehat{C} - \mathbb{E}\widehat{C}. \tag{4.3}$$

The projection outside the signal subspace is small exactly above (3.7) and large below it, yielding the matching recovery and non-recovery statements. After this feature-wise characterization, the power-law rate follows by summing the unrecovered tail $\sum_{i > m_n} \lambda_i^2$.

Numerical evidence. The recovery and scaling predictions in Theorem 1 are visible in Figures 2 and 3. In Figure 2, as $\alpha = \log(n)/\log(d)$ crosses the spectral scale predicted by the theorem, the first-layer overlap rises and the MSE falls; the spectrum of $\widehat{C}$ shows the same mechanism at the matrix level, with separated empirical spikes emerging from the noise bulk. In Figure 3, increasing γ separates the strongest directions from the weak tail, so the MSE and representation-overlap curves become more gradual, as predicted by the thresholds $n_i \asymp Di^{2\gamma}/Z_\gamma^2$. Direction-wise recovery rates, the nonlinear-readout experiment, metrics, and implementation details are given in Appendix A.

Discussion. We have isolated a representation-learning mechanism for scaling laws: depth exposes a lower-degree latent representation, while a power-law spectrum spreads coordinate recovery across sample sizes. Smooth learning curves can therefore arise from many sharp feature-learning transitions, rather than from a fixed kernel spectrum. The Gaussian inputs, specified hierarchy, and layer-wise spectral learner are stylized, but they enable sharp recovery/non-recovery guarantees and suggest extensions to richer nonlinearities, less structured data, and end-to-end training dynamics.

References

[1] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

[2] Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

[3] Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, Lenka Zdeborová, et al. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

[4] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

[5] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[6] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.

[7] Gérard Ben Arous, Murat A Erdogdu, Nuri Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: Sgd dynamics and scaling laws. *arXiv preprint arXiv:2508.03688*, 2025.

[8] Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in Neural Information Processing Systems*, 35:9768–9783, 2022.

[9] Fabrizio Boncoraglio, Vittorio Erba, Emanuele Troiani, Yizhou Xu, Florent Krzakala, and Lenka Zdeborová. Single-head attention in high dimensions: A theory of generalization, weights spectra, and scaling laws. In *Workshop on Scientific Methods for Understanding Deep Learning*, 2025.

[10] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

[11] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, pages 4345–4382, 2024.

[12] Blake Bordelon, Lorenzo Noci, Mufan Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

[13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[14] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[15] Francesco Cagnetta, Leonardo Petrini, Umberto M. Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: The random hierarchy model. *Physical Review X*, 14:031001, 2024. doi: 10.1103/PhysRevX.14.031001.

[16] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[17] Louis-Pierre Chaintron, Lénaïc Chizat, and Javier Maas. Resnets of all shapes and sizes: Convergence of training dynamics in the large-scale limit. *arXiv preprint arXiv:2603.18168*, 2026.

[18] Lénaïc Chizat and Praneeth Netrapalli. The feature speed formula: a flexible approach to scale hyper-parameters of deep neural networks. *Advances in Neural Information Processing Systems*, 37:62362–62383, 2024.

[19] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In *Advances in Neural Information Processing Systems*, volume 34, pages 10131–10143, 2021.

[20] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Error scaling laws for kernel classification under source and capacity conditions. *Machine Learning: Science and Technology*, 4(3):035033, 2023.

[21] Alex Damian, Loucas Pillaud-Vivien, Jason D. Lee, and Joan Bruna. Computational-statistical gaps in gaussian single-index models. In *Proceedings of the 37th Annual Conference on Learning Theory (COLT)*, 2024.

[22] Yatin Dandi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The computational advantage of depth: Learning high-dimensional hierarchical functions with gradient descent, 2025. URL https://arxiv.org/abs/2502.13961.

[23] Amit Daniely. Depth separation for neural networks. In *Conference on Learning Theory*, pages 690–696. PMLR, 2017.

[24] Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

[25] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

[26] Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. In *Advances in Neural Information Processing Systems*, volume 37, pages 104630–104693, 2024.

[27] Leonardo Defilippis, Yatin Dandi, Pierre Mergny, Florent Krzakala, and Bruno Loureiro. Optimal spectral transitions in high-dimensional multi-index models. In *Advances in Neural Information Processing Systems*, volume 38, pages 174966–175002. Curran Associates, Inc., 2025.

[28] Leonardo Defilippis, Florent Krzakala, Bruno Loureiro, and Antoine Maillard. Optimal scaling laws in learning hierarchical multi-index models. *arXiv preprint arXiv:2602.05846*, 2026.

[29] Leonardo Defilippis, Yizhou Xu, Julius Girardin, Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Scaling laws and spectra of shallow neural networks in the feature learning regime. In *The Fourteenth International Conference on Learning Representations*, 2026.

[30] Justin Eldridge, Mikhail Belkin, and Yusu Wang. Unperturbed: spectral analysis beyond davis-kahan. In *Algorithmic learning theory*, pages 321–358. PMLR, 2018.

[31] Hengyu Fu, Zihao Wang, Eshaan Nichani, and Jason D. Lee. Learning hierarchical polynomials of multiple nonlinear features with three-layer networks. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UZ893n8FXr.

[32] Jerome Garnier-Brun, Marc Mezard, Emanuele Moscato, and Luca Saglietti. How transformers learn structured data: Insights from hierarchical filtering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 18831–18847. PMLR, 13–19 Jul 2025.

[33] Anne Greenbaum, Ren-cang Li, and Michael L Overton. First-order perturbation theory for eigenvalues and eigenvectors. *SIAM review*, 62(2):463–482, 2020.

[34] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030, 2022.

[35] Hong Hu, Yue M Lu, and Theodor Misiakiewicz. Asymptotics of random feature regression beyond the linear scaling regime. *arXiv preprint arXiv:2403.08160*, 2024.

[36] Svante Janson. *Gaussian hilbert spaces*. Cambridge university press, 1997.

[37] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

[38] Filip Kovačević, Yihan Zhang, and Marco Mondelli. Spectral estimators for multi-index models: Precise asymptotics and optimal weak recovery. *arXiv preprint arXiv:2502.01583*, 2025.

[39] Yue M Lu and Gen Li. Phase transitions of spectral initialization for high-dimensional non-convex estimation. *Information and Inference: A Journal of the IMA*, 9(3):507–541, 2020.

[40] Yue M Lu and Horng-Tzer Yau. An equivalence principle for the spectrum of random inner-product kernel matrices with polynomial scalings. *The Annals of Applied Probability*, 35(4): 2411–2470, 2025.

[41] Antoine Maillard, Florent Krzakala, Yue M Lu, and Lenka Zdeborová. Construction of optimal spectral methods in phase retrieval. In *Mathematical and Scientific Machine Learning*, pages 693–720. PMLR, 2022.

[42] Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

[43] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

[44] Hrushikesh Mhaskar, Qianli Liao, and Tomaso Poggio. When and why are deep networks better than shallow ones? In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[45] Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. In *Conference On Learning Theory*, pages 1445–1450. PMLR, 2018.

[46] Elchanan Mossel. Deep learning and hierarchal generative models. *arXiv preprint arXiv:1612.09057*, 2016.

[47] Eshaan Nichani, Alex Damian, and Jason D. Lee. Provable guarantees for nonlinear feature learning in three-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[48] Ivan Nourdin and Giovanni Peccati. Stein's method on wiener chaos. *Probability Theory and Related Fields*, 145(1):75–118, 2009.

[49] Ivan Nourdin and Giovanni Peccati. *Normal approximations with Malliavin calculus: from Stein's method to universality*, volume 192. Cambridge University Press, 2012.

[50] DAVID NUALART and GIOVANNI PECCATI. Central limit theorems for sequences of multiple stochastic integrals. *The Annals of Probability*, 33(1):177–193, 2005.

[51] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. *Advances in Neural Information Processing Systems*, 37: 16459–16537, 2024.

[52] Giovanni Peccati and Murad S Taqqu. *Wiener Chaos: Moments, Cumulants and Diagrams: A survey with computer implementation*, volume 1. Springer Science & Business Media, 2011.

[53] Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

[54] Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D. Lee. Emergence and scaling laws in SGD learning of shallow neural networks, 2025. URL https://arxiv.org/abs/2504.19983.

[55] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

[56] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581, 2023.

[57] Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: Empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, 2020.

[58] Hugo Tabanelli, Yatin Dandi, Luca Pesce, and Florent Krzakala. Deep learning of compositional targets with hierarchical spectral methods, 2026. URL https://arxiv.org/abs/2602.10867.

[59] Matus Telgarsky. benefits of depth in neural networks. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1517–1539, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

[60] Emanuele Troiani, Yatin Dandi, Leonardo Defilippis, Lenka Zdeborova, Bruno Loureiro, and Florent Krzakala. Fundamental computational limits of weak learnability in high-dimensional multi-index models. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2467–2475. PMLR, 03–05 May 2025.

[61] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[62] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[63] Zihao Wang, Eshaan Nichani, and Jason D Lee. Learning hierarchical polynomials with three-layer neural networks. *arXiv preprint arXiv:2311.13774*, 2023.

[64] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models, 2022. URL https://arxiv.org/abs/2206.07682.

[65] Garrett G Wen, Hong Hu, Yue M Lu, Zhou Fan, and Theodor Misiakiewicz. When does gaussian equivalence fail and how to fix it: Non-universal behavior of random features with quadratic scaling. *arXiv preprint arXiv:2512.03325*, 2025.

[66] Yizhou Xu, Antoine Maillard, Lenka Zdeborová, and Florent Krzakala. Fundamental limits of matrix sensing: Exact asymptotics, universality, and applications. In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 5757–5823. PMLR, 30 Jun–04 Jul 2025.

[67] Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34: 17084–17097, 2021.

Appendix A. Further numerics

All experiments follow the hierarchical spectral procedure in Algorithm 1, implementation and reproducibility details are given in Section A.4. This appendix also provides additional direction-wise recovery-rate numerics, the nonlinear-readout experiment, and the metrics used in the figures.

Algorithm 1: Hierarchical spectral learning (training procedure)

Data: Data $\{(\boldsymbol{x}_\mu, y_\mu)\}_{\mu=1}^n$, max degree $K_{\max}$
Result: $\widehat{A}^{(1)}, \widehat{A}^{(2)}$
First layer recovery.
Compute flattened degree-q features and moment matrix

$$\phi_\mu = \mathcal{F}[\mathrm{He}_q(x_\mu)] \in \mathbb{R}^D,$$
$$\widehat{C} = \frac{1}{n}\sum_{\mu=1}^n y_\mu \,\mathrm{He}_2(\phi_\mu) \in \mathbb{R}^{D\times D}.$$

Compute top eigenvectors $\{\widehat{A}_i^{(1)}\}_{i\in[d_1]} \subset \mathbb{R}^D$.
for $\mu = 1, \dots, n$ *and* $i = 1, \dots, d_1$ **do**
  $\widehat{h}_{\mu,i}^{(1)} \leftarrow \langle \widehat{A}_i^{(1)}, \mathrm{He}_q(x_\mu)\rangle$.
end
Second layer recovery.
Compute the second-order moment matrix

$$\widehat{A}^{(2)} = \frac{1}{n}\sum_{\mu=1}^n y_\mu \,\mathrm{He}_2(\widehat{h}_\mu^{(1)}) \in \mathbb{R}^{d_1\times d_1}.$$

for $\mu = 1, \dots, n$ **do**
  $\widehat{h}_\mu^{(2)} \leftarrow \langle \widehat{A}^{(2)}, \mathrm{He}_2(\widehat{h}_\mu^{(1)})\rangle \in \mathbb{R}$.
end
Perform kernel regression on $\{(\widehat{h}_\mu^{(2)}, y_\mu)\}_{\mu=1}^n$.

A.1. Additional experiment: direction-wise recovery rates

Fig. 4 gives a more local test of the recovery theory than the aggregate MSE curves in the main text. The left panel shows that individual directions are recovered sequentially: directions with larger weights $|\lambda_i| = Z_\gamma i^{-\gamma}$ turn on earlier, while weaker directions appear later, in agreement with the threshold prediction $n_i \asymp Di^{2\gamma}/Z_\gamma^2$. The center panel focuses on the post-transition regime. Once a direction has crossed its spectral threshold, the angular error decreases approximately at the predicted $1/n$ rate from Equation (4.1). Finally, the right panel compares the empirical aggregate feature overlap with the theoretical count of recovered directions. The finite-dimensional curve is smoother than the idealized threshold prediction, but its scale and ordering agree with the cascade-of-transitions picture.

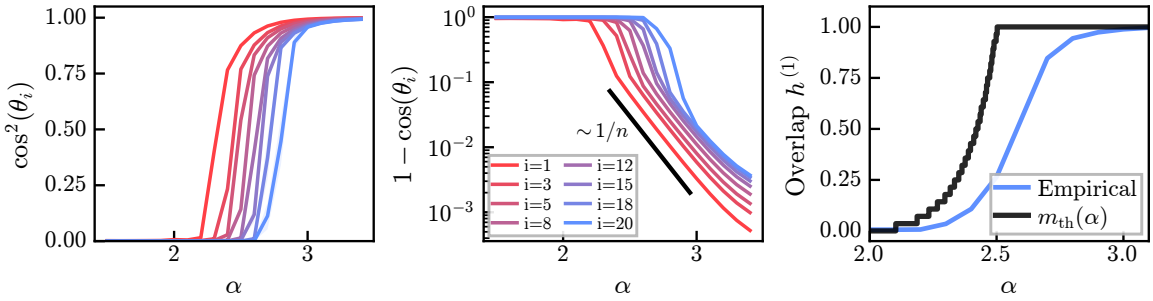

Figure 4: **Direction-wise recovery and post-transition rates.** Parameters: $q = 2$, $\varepsilon = 0.5$, $\gamma = 0.4$, and $g^\star = \mathrm{id}$. **Left:** Direction-wise alignments $\cos^2(\theta_i)$ versus $\alpha = \log(n)/\log(d)$, with $d = 400$. **Center:** Direction-wise angular errors $1 - \cos(\theta_i)$ versus α, with a $1/n$ guide and $d = 400$. **Right:** Aggregate first-layer overlap $q_h^{(1)}$ compared with the theoretical recovered-direction count $m_{\rm th}(\alpha)$, with $d = 800$.

A.2. Additional experiment: nonlinear readout

Fig. 5 repeats the experiment with a nonlinear readout g. The performance is qualitatively comparable to the identity-readout case: the MSE decreases at the same scale at which the first-layer overlap grows. Finite-size effects are more visible at lower dimensions, but the dominant bottleneck remains the recovery of the latent first-layer representation, not the final low-dimensional readout.

A.3. Metrics and evaluation protocol

We use three complementary diagnostics in the numerical experiments. The MSE measures end-to-end prediction performance, the first-layer feature overlap measures recovery of the latent representation as a subspace, and the direction-wise cosines isolate the individual spectral transitions predicted by Theorem 1. These quantities answer different questions: the MSE mixes all recovered and unrecovered directions through their weights, the aggregate overlap summarizes representation recovery, and the direction-wise overlaps reveal which latent directions have crossed their spectral threshold.

Generalization error. The MSE reported in the figures is the empirical test estimate of the generalization error. Given an independent test set $\{(x_\mu^{\rm test}, y_\mu^{\rm test})\}_{\mu=1}^{n_{\rm test}}$, we compute

$$\widehat{\rm MSE} = \frac{1}{n_{\rm test}} \sum_{\mu=1}^{n_{\rm test}} \left(\widehat{f}(x_\mu^{\rm test}) - y_\mu^{\rm test} \right)^2 . \tag{A.1}$$

This is the final performance metric of the algorithm. The overlap quantities below are diagnostic measures tied to the teacher-student setting. They are used to identify whether changes in MSE are caused by first-layer feature recovery or else.

First-layer feature overlap. Let $H^{(1)}, \widehat{H}^{(1)} \in \mathbb{R}^{n_{\rm test} \times d_1}$ denote the true and learned first-layer feature matrices evaluated on the same independent test set. Let Q and $\widehat{Q}$ be orthonormal bases for the column spaces of $H^{(1)}$ and $\widehat{H}^{(1)}$, respectively. We define

$$q_h^{(1)} = \frac{1}{d_1} \|Q^\top \widehat{Q}\|_F^2 . \tag{A.2}$$

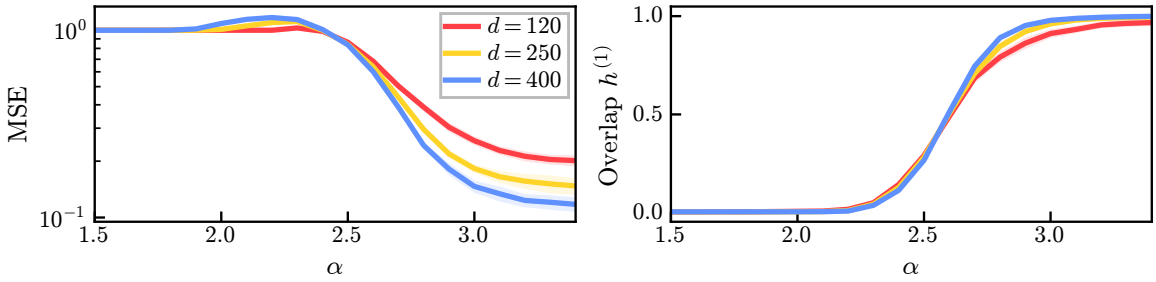

Figure 5: **Non-identity readout.** Parameters: $q = 2$, $\varepsilon = 0.5$, $\gamma = 0.4$, and $g^\star = \tanh$. **Left:** Test MSE versus $\alpha = \log(n)/\log(d)$. **Right:** First-layer feature overlap $q_h^{(1)}$ versus α.

Equivalently, $q_h^{(1)}$ is the average squared canonical correlation between the true and learned latent feature spaces. It belongs to $[0, 1]$, more precisely, it is equal to one when the learned features span the same subspace as the true features, and close to zero when the two subspaces are nearly orthogonal.

This subspace definition is intentional. The learned coordinates are only meaningful up to signs, permutations, and possible finite-size rotations inside the recovered eigenspace. These transformations can be absorbed by the second-layer fit and should not be counted as representation error.

Direction-wise alignment. The aggregate overlap $q_h^{(1)}$ does not show which individual directions have been recovered. To test the feature-wise prediction of Theorem 1, we also measure single direction alignments. Let $u_i = A_i^{(1)}/\|A_i^{(1)}\|$ be the normalized i-th teacher direction, and let $\widehat{U} \in \mathbb{R}^{D\times d_1}$ be an orthonormal basis of the top eigenspace of $\widehat{C}$ (returned by the first spectral step). We define:

$$\cos^2(\theta_i) = \|\widehat{U}^\top u_i\|^2 = u_i^\top \widehat{U}\widehat{U}^\top u_i. \tag{A.3}$$

This is a projector overlap. It is invariant to the sign of u_i and to the choice of basis inside the recovered eigenspace. The value $\cos^2(\theta_i) \simeq 0$ means that direction i is absent from the recovered subspace, while $\cos^2(\theta_i) \simeq 1$ means that it has been recovered. This is the quantity shown in the left panel of Figure 4, where the different curves turn on sequentially according to the spike sizes $|\lambda_i| = Z_\gamma i^{-\gamma}$.

After a direction has been recovered, we are interested not only in whether it is present, but also in how fast its alignment improves with n. For this post-transition regime we write $\cos(\theta_i) = \|\widehat{U}^\top u_i\|$ and plot the angular error $1 - \cos(\theta_i)$. This is the quantity shown in the center panel of Figure 4. We use $1 - \cos(\theta_i)$ rather than $1 - \cos^2(\theta_i)$ because Theorem 1 is stated directly in terms of the absolute eigenvector overlap and predicts

$$1 - \cos(\theta_i) = O_d\left(\frac{d^q i^{2\gamma}}{n Z_\gamma^2}\right) \tag{A.4}$$

after recovery.

Finally, when the test features are well conditioned, the aggregate feature overlap can be viewed as a smoothed average of the direction-wise overlaps. Heuristically, $q_h^{(1)} \approx d_1^{-1}\sum_{i=1}^{d_1}\cos^2(\theta_i) =: m_{\mathrm{th}}(\alpha)$, so the global overlap curves in Figures 2 and 3 summarize the cascade of individual transitions displayed in Figure 4.

A.4. Implementation and reproducibility

All experiments follow the hierarchical spectral pipeline described in Algorithm 1. For each value of the sample exponent α, we use $n = \lfloor d^\alpha \rfloor$ training samples and evaluate the MSE and overlaps on an independent test set. For the identity readout $g^\star = \mathrm{id}$, we use the scalar estimator $\widehat{h}^{(2)}$ directly. For the nonlinear experiment with $g^\star = \tanh$, we fit a polynomial ridge regressor on $\widehat{h}^{(2)}$. This readout is specified by three hyperparameters: the maximal polynomial degree r, the ridge parameter ρ, and the kernel regularization λ_{poly}. For the final curves reported in Figure 5, we use $(r, \rho, \lambda_{\mathrm{poly}}) = (3, 10^{-5}, 10^{-4})$, selected on the grid

$$r \in \{3, 5, 7, 9\}, \qquad \rho \in \{10^{-7}, 10^{-6}, 10^{-5}\}, \qquad \lambda_{\mathrm{poly}} \in \{10^{-5}, 10^{-4}, 10^{-3}\}.$$

The full set of parameters used in the numerical figures is summarized below.

- **Figure 2, left-center.** MSE and $q_h^{(1)}$; $d \in \{120, 250, 400\}$, α on a grid in $[1.5, 3.4]$, $q = 2$, $\varepsilon = 0.5$, $\gamma = 0.4$, $g^\star = \mathrm{id}$. We use a linear readout on $\widehat{h}^{(2)}$ and average over 10 seeds.
- **Figure 2, right.** Spectrum of $\widehat{C}$; $d = 140$, $\alpha = 3.5$, $q = 2$, $\varepsilon = 0.5$, $\gamma = 0.4$, $g^\star = \mathrm{id}$. We plot the full spectrum with no readout fit, for one seed.
- **Figure 3.** MSE and $q_h^{(1)}$; $d = 400$, α on a grid in $[1.5, 3.4]$, $q = 2$, $\varepsilon = 0.5$, $\gamma \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, $g^\star = \mathrm{id}$. We use a linear readout on $\widehat{h}^{(2)}$ and average over 10 seeds.
- **Figure 4, left-center.** Direction-wise quantities $\cos^2(\theta_i)$ and $1 - \cos(\theta_i)$; $d = 400$, α on a grid in $[1.5, 3.4]$, $q = 2$, $\varepsilon = 0.5$, $\gamma = 0.4$, $g^\star = \mathrm{id}$. We track directions $i \in \{1, 3, 5, 8, 12, 15, 18, 20\}$ and average over 10 seeds.
- **Figure 4, right.** Aggregate overlap $q_h^{(1)}$ and theoretical count $m_{\mathrm{th}}(\alpha)$; $d = 800$, α on a grid in $[2.0, 3.2]$, $q = 2$, $\varepsilon = 0.5$, $\gamma = 0.4$, $g^\star = \mathrm{id}$. We use a linear readout on $\widehat{h}^{(2)}$ and average over 10 seeds.
- **Figure 5.** MSE and $q_h^{(1)}$; $d \in \{120, 250, 400\}$, α on a grid in $[1.5, 3.4]$, $q = 2$, $\varepsilon = 0.5$, $\gamma = 0.4$, $g^\star = \tanh$. We use polynomial ridge regression on $\widehat{h}^{(2)}$, with final choice $(r, \rho, \lambda_{\mathrm{poly}}) = (3, 10^{-5}, 10^{-4})$, and average over 10 seeds.

Concerning resources, all experiments were run on single-GPU workers of an internal compute cluster, using up to 32 GB of host memory per job. Depending on the values of d, α, and the number of seeds, runtimes ranged from a few minutes for the smallest jobs to several hours for the largest ones; for the heaviest sweeps, jobs were typically submitted with a wall-clock budget of up to 12–24 hours.

Appendix B. Preliminary Results

B.1. First Preliminary Results

Lemma 2 (Computation of Z_γ) *From the criterion* $\mathrm{Var}[(h^{(2)})^2] = \Theta(1)$*, it comes:*

$$Z_\gamma \propto \begin{cases} 1/\sqrt{d_1} & \text{if } \gamma = 0 \\ d_1^{\gamma - \frac{1}{2}}, & \text{if } \gamma < \frac{1}{2}, \\ 1, & \text{if } \gamma > \frac{1}{2}. \end{cases} \tag{B.1}$$

Proof

$$\mathrm{Var}\left[(h^{(2)})^2\right] = \mathrm{Var}\left(\langle A^{(2)}, He_2(h^{(1)})\rangle\right) \tag{B.2}$$

$$= \mathbb{E}\left[\langle A^{(2)}, \mathrm{He}_2(\boldsymbol{h}^{(1)})\rangle^2\right] \tag{B.3}$$

$$= Z_\gamma^2 \sum_{i=1}^{d_1} i_1^{-2\gamma} \underbrace{\mathbb{E}\left[\left(\left(h_i^{(1)}\right)^2 - 1\right)^2\right]}_{=\Theta(1)} \tag{B.4}$$

$$= \Theta\left(Z_\gamma^2 \sum_{i=1}^{d_1} i_1^{-2\gamma}\right). \tag{B.5}$$

The sum on the RHS has the following asymptotic behavior:

$$\sum_{i=1}^{d_1} i_1^{-2\gamma} = \begin{cases} d_1 & \text{if } \gamma = 0 \\ d_1^{1-2\gamma}, & \text{if } 0 \le \gamma < \frac{1}{2} \\ \log(d_1), & \text{if } \gamma = \frac{1}{2} \\ \Theta(1), & \text{if } \gamma > \frac{1}{2}. \end{cases} \tag{B.6}$$

Then, taking $Z_\gamma = \Theta(\sqrt{\sum_{i=1}^{d_1} i^{-2\gamma}})$ concludes the result. ■

B.2. Auxiliary Concentration Lemmas

Lemma 3 (Lemma F.4 in [65]) *with probability at least* $1 - e^{-cd}$,

$$\|F_\mu\|_2^2 \le Cd^q, \tag{B.7}$$

for a universal constant C.

Lemma 4 (Corollary 5.21 in [13]) *Let* $f(x) = \sum_{i=0}^k a_i x^i$ *be a polynomial of degree* k *of a real variable and let* X *be a standard normal random variable. Then for any* $q > 2$,

$$\left(E\left[|f(X)|^q\right]\right)^{1/q} \le (q-1)^{k/2} \left(E\left[|f(X)|^2\right]\right)^{1/2}.$$

Lemma 5 (Gaussian Poincaré Inequality, Theorem 3.20 in [13]) *Let* $X = (X_1, \ldots, X_d)$ *be a vector of i.i.d. standard Gaussian random variables (i.e.,* X *is a Gaussian vector with zero mean vector and identity covariance matrix). Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be any continuously differentiable function. Then*

$$\mathrm{Var}(f(X)) \le E\left[\|\nabla f(X)\|^2\right].$$

Appendix C. Detection of outliers

Recall that the algorithm described in Algorithm 1 works by first computing the matrix:

$$\hat{C} = \frac{1}{n}\sum_{\mu=1}^{n} y_\mu \mathrm{He}_2(\mathcal{F}[\mathrm{He}_q(x_\mu)]). \tag{C.1}$$

To simplify the our notation, let $F_\mu = \mathcal{F}[\mathrm{He}_q(x_\mu)] \in \mathbb{R}^D$, where $D = B(d,k) = \binom{d+k-1}{k-1} = \Theta_d(d^q)$. With this notation,

$$\hat{C} = \frac{1}{n}\sum_{\mu=1}^{n} y_\mu \left(F_\mu F_\mu^T - I_D\right).$$

We want to study the eigenvectors of this random matrix. For this, we first write:

$$\hat{C} = \mathbb{E}\left[\hat{C}\right] + \left(\hat{C} - \mathbb{E}\left[\hat{C}\right]\right). \tag{C.2}$$

The reader should think of the first term as the signal and the second one as the noise. We begin this section by studying the signal part.

C.1. Studying $\mathbb{E}\left[\hat{C}\right]$

As noted in [58], (and previously in [63, 65], among others) when computing the expectation $\mathbb{E}\left[\hat{C}\right]$, the vectors F_μ behave as if they were isotropic Gaussians in $\mathbb{R}^D$. For this reason, we have:

Lemma 6 *Let $\gamma > 0$, and denote $\lambda_j = Z_\gamma j^{-\gamma}$ and $u_j = A_j^{(1)}$. Then:*

$$\mathbb{E}\left[C^{(1)}\right] = \frac{\nu_1}{\sqrt{2}} A^{(1)} D_\gamma (A^{(1)})^T + \Delta,$$

where $\|\Delta\|_{\mathrm{op}} = o_d(1)$, $A^{(1)} = [u_1, \dots u_{d_1}] \in \mathbb{R}^{D\times d_1}$, $D_\gamma = \mathrm{diag}(\lambda_1, \dots, \lambda_{d_1}) \in \mathbb{R}^{d_1\times d_1}$ and ν_1 denotes the first Hermite coefficient of g.

We postpone the Proof of Theorem 6 to Section E.5.

C.2. Eigenvector Perturbation Formula

As noted before, it will be useful to write:

$$\hat{C} = \mathbb{E}[\hat{C}] + \left(\hat{C} - \mathbb{E}\left[C^{(1)}\right]\right) = A^{(1)} D_\gamma (A^{(1)})^T + \left(\hat{C} - \mathbb{E}\left[C^{(1)}\right]\right) + \Delta. \tag{C.3}$$

Since $\|\Delta\|_{\mathrm{op}} = o_d(1)$, we only care about the first two terms of the decomposition above.

Now, for $z \in \mathbb{C}$, let

$$R_{\bar{C}}(z) = (zI_D - \mathbb{E}[\hat{C}])^{-1}, R_{\hat{C}}(z) = (zI_D - \hat{C})^{-1}. \tag{C.4}$$

Then:

$$R_{\hat{C}}(z) - R_{\bar{C}}(z) = R_{\hat{C}}(z)(\hat{C} - \mathbb{E}[\hat{C}])R_{\bar{C}}(z). \tag{C.5}$$

Denote $\Delta = \hat{C} - \mathbb{E}[\hat{C}]$. Then:

$$zI_D - \hat{C} = zI_D - \mathbb{E}[C] - \Delta = (zI_D - \mathbb{E}[C])(I_D - R_{\bar{C}}(z)\Delta). \tag{C.6}$$

Therefore:

$$R_{\hat{C}}(z) = (I_D - R_{\bar{C}}(z)\Delta)^{-1} R_{\bar{C}}(z). \tag{C.7}$$

If $\|R_{\bar{C}}\Delta\|_{\text{op}} \leq 1$, then we can expand $(I_D - R_{\bar{C}}(z)\Delta)^{-1}$ into its Neumann series:

$$(I_D - R_{\bar{C}}(z)\Delta)^{-1} = \sum_{\ell \geq 0} (R_{\bar{C}}(z)\Delta)^{\ell}. \tag{C.8}$$

Then, going back to Equation (C.7), we can write:

$$R_{\hat{C}}(z) = R_{\bar{C}}(z) + R_{\bar{C}}(z)\Delta R_{\bar{C}}(z) + o(\|\Delta\|_{\text{op}}^2). \tag{C.9}$$

Let $u_k, \hat{u}_k$ be isolated eigenvectors of $\mathbb{E}[\hat{C}], \hat{C}$, respectively. Let γ be a contour around λ_k and $\hat{\lambda}_k$. Then, we can write the projectors $\Pi_k = u_k u_k^T$ and $\hat{\Pi}_k = \hat{u}_k \hat{u}_k$ as:

$$\Pi_k = \frac{1}{2\pi i} \oint_{\gamma} R_{\bar{C}}(z) dz, \quad \hat{\Pi}_k = \frac{1}{2\pi i} \oint_{\gamma} R_{\hat{C}}(z) dz. \tag{C.10}$$

Then, we can integrate Equation (C.9) to get:

$$\hat{\Pi}_k \Pi_k = \Pi_k + \frac{1}{2\pi i} \oint_{\gamma} R_{\bar{C}} \Delta R_{\bar{C}} + o(\|\Delta\|_{\text{op}}^2). \tag{C.11}$$

We can re-write the resolvent $R_{\bar{C}}$ in the following way:

$$R_{\bar{C}}(z) = (zI - \mathbb{E}[\hat{C}])^{-1} = \sum_{j=1}^{d_1} \frac{1}{z - \lambda_j} u_j u_j^T + \frac{1}{z} P_{\text{Ker}}, \tag{C.12}$$

where $P_{\text{Ker}} = P_{\text{Ker}(\mathbb{E}[\hat{C}])}$ denotes the projection into the kernel of $\mathbb{E}[\hat{C}]$. Then:

$$R_{\bar{C}}(z)\Delta R_{\bar{C}}(z) = \left(\sum_{j=1}^{d_1} \frac{1}{z - \lambda_j} u_j u_j^T + \frac{1}{z} P_{\text{Ker}} \right) \Delta \left(\sum_{j=1}^{d_1} \frac{1}{z - \lambda_j} u_j u_j^T + \frac{1}{z} P_{\text{Ker}} \right) \tag{C.13}$$

$$= \sum_{j_1, j_2} \frac{1}{(z - \lambda_{j_1})(z - \lambda_{j_2})} u_{j_1} u_{j_1}^T \Delta u_{j_2} u_{j_2}^T + \sum_{j=1}^{d_1} \frac{1}{z - \lambda_j} u_j u_j^T \Delta P_{\text{Ker}} \tag{C.14}$$

$$+ \sum_{j=1}^{d_1} \frac{1}{z - \lambda_j} u_j u_j^T \Delta P_{\text{Ker}} + \frac{1}{z^2} P_{\text{Ker}} \Delta P_{\text{Ker}}. \tag{C.15}$$

Integrating, we get:

$$\frac{1}{2\pi i}\oint_\gamma R_{\bar{C}}(z)\Delta R_{\bar{C}}(z) = \frac{1}{2\pi i}\oint_\gamma \sum_{j_1,j_2}\frac{1}{(z-\lambda_{j_1})(z-\lambda_{j_2})}u_{j_1}u_{j_1}^T\Delta u_{j_2}u_{j_2}^T + \frac{1}{2\pi i}\oint_\gamma \sum_{j=1}^{d_1}\frac{1}{z-\lambda_j}u_j u_j^T\Delta P_{\text{Ker}} \tag{C.16}$$

$$+\frac{1}{2\pi i}\oint_\gamma \sum_{j=1}^{d_1}\frac{1}{z-\lambda_j}u_j u_j^T\Delta P_{\text{Ker}} + \underbrace{\frac{1}{2\pi i}\oint_\gamma \frac{1}{z^2}P_{\text{Ker}}\Delta P_{\text{Ker}}}_{=0} \tag{C.17}$$

$$= \sum_{j=1,j\neq k}^{d_1}\frac{u_j^T\Delta u_j}{\lambda_k-\lambda_j}\left(u_j u_k^T + u_k u_j^T\right) + \frac{1}{\lambda_k}u_k u_k^T\Delta P_{\text{Ker}} + \frac{1}{\lambda_k}P_{\text{Ker}}\Delta u_k u_k^T. \tag{C.18}$$

Then, replacing this in Equation (C.11):

$$\hat{\Pi}_k = \Pi_k + \sum_{j=1,j\neq k}^{d_1}\frac{u_j^T\Delta u_j}{\lambda_k-\lambda_j}\left(u_j u_k^T + u_k u_j^T\right) + \frac{1}{\lambda_k}u_k u_k^T\Delta P_{\text{Ker}} + \frac{1}{\lambda_k}P_{\text{Ker}}\Delta u_k u_k^T + o(\|\Delta\|_{\text{op}}^2). \tag{C.19}$$

In order to get eigenvectors, we apply this projection to u_k and obtain:

$$\hat{\Pi}_k u_k = u_k + \left(\sum_{j=1,j\neq k}^{d_1}\frac{u_j^T\Delta u_j}{\lambda_k-\lambda_j}\left(u_j u_k^T + u_k u_j^T\right) + \frac{1}{\lambda_k}u_k u_k^T\Delta P_{\text{Ker}} + \frac{1}{\lambda_k}P_{\text{Ker}}\Delta u_k u_k^T + o(\|\Delta\|_{\text{op}}^2)\right)u_k \tag{C.20}$$

$$= u_k + \sum_{j=1,j\neq k}^{d_1}\frac{u_j^T\Delta u_j}{\lambda_k-\lambda_j}u_j + \frac{1}{\lambda_k}P_{\text{Ker}}\Delta u_k + o(\|\Delta\|_{\text{op}}^2). \tag{C.21}$$

Note that the second and third terms are orthogonal to u_k, so

$$\|\hat{\Pi}_k u_k\|^2 = 1 + \left\|\sum_{j=1,j\neq k}^{d_1}\frac{u_j^T\Delta u_j}{\lambda_k-\lambda_j}u_j + \frac{1}{\lambda_k}P_{\text{Ker}}\Delta u_k\right\|^2 + o(\|\Delta\|_{\text{op}}^2), \tag{C.22}$$

and since the term in the middle is bounded by $C\|\Delta\|^2$, we conclude that:

$$\|\hat{\Pi}_k u_k\|^2 = 1 + o(\|\Delta\|^2). \tag{C.23}$$

Then, we can normalize Equation (C.21) and we will have:

$$\hat{u}_k = u_k + \sum_{j=1,j\neq k}^{d_1}\frac{u_j^T\Delta u_j}{\lambda_k-\lambda_j}u_j + \frac{1}{\lambda_k}P_{\text{Ker}}\Delta u_k + o(\|\Delta\|_{\text{op}}^2). \tag{C.24}$$

At last, we replace the eigenvectors u_k by $A_k^{(1)}$, by using Theorem 6, plus the fact that $A^{(1)}$ are almost the eigenvectors of $\mathbb{E}\left[\hat{C}\right]$. To see this, note that by applying a covariance concentration bound [61] for $\hat{A}^{(1)}, \dots, \hat{A}^{(1)}$, the condition $\varepsilon < (k - 2\gamma)$ gives that

$$\left\| \sum_{i=1}^{d_1} \frac{1}{\lambda_i} u_i u^T - \sum_{i=1}^{d_1} \frac{1}{\lambda_i} A_i^{(1)} (A^{(1)})^T \right| \leq d_1^{\gamma} \sqrt{\frac{d_1}{d^q}} = d^{\varepsilon(\frac{1}{2}+2\gamma)-q}, \tag{C.25}$$

which is small by our assumption that $\varepsilon < q/(1-2\gamma)$ to get:

$$\hat{u}_k = A_k^{(1)} + \sum_{j=1, j\neq k}^{d_1} \frac{(A_k^{(1)})^T \Delta (A_j^{(1)})}{\lambda_k - \lambda_j} (A_j^{(1)}) + \frac{1}{\lambda_k} P_{\mathrm{Ker}} \Delta u_k + o(\|\Delta\|_{\mathrm{op}}^2). \tag{C.26}$$

C.3. Analysis of Outliers - Sufficient Sample Complexity

Having Equation (C.24), we can study the eigenvectors of $\hat{C}$. In order for the expansion to be valid, we need $\|\Delta\|_{\mathrm{op}} = \|\hat{C} - \mathbb{E}[\hat{C}]\|_{\mathrm{op}}$ to be small. In the following Lemma, we show that this is indeed the case when $n \gg d^q$.

Lemma 7 *Consider the estimator $\hat{C}$ in Algorithm 1 computed for the Hermite tensor of degree q. Then with high probability*

$$\left\| \hat{C} - \mathbb{E}[\hat{C}] \right\|_{\mathrm{op}} \lesssim \sqrt{\frac{d^q}{n}}. \tag{C.27}$$

Proof The proof proceed the same way as [58] and [65]. By Lemma F.4 in [65], with probability at least $1 - e^{-cd}$,

$$\|F_\mu\|_2^2 \leq C d^q, \tag{C.28}$$

for a universal constant C. By truncating the matrix $\hat{C}$ with indicators $\mathbf{1}_{\|\|F_\mu\|_2^2 \leq C d^q}$, and applying Bernstein's inequality, we get the desired results. ∎

Then, by Theorem 7, we can apply the expansion Equation (C.24) and we can conclude

Lemma 8 *Let $\hat{u}_k$ denote the k-th eigenvector of $\hat{C}$, and u_k the k-th eigenvector of $\mathbb{E}[\hat{C}]$. Denote $\Delta := \hat{C} - \mathbb{E}[\hat{C}]$. Then,*

$$\hat{u}_k = A_k^{(1)} + \sum_{j=1, j\neq k}^{d_1} \frac{(A_k^{(1)})^T \Delta (A_j^{(1)})}{\lambda_k - \lambda_j} (A_j^{(1)}) + \frac{1}{\lambda_k} P_{\mathrm{Ker}} \Delta u_k + o_d(\|\Delta\|_{\mathrm{op}}^2).$$

where $\|\Delta\|_{\mathrm{op}} = O\left(\max\left(\sqrt{\frac{d^q \log(d)}{n}}, \frac{d_1^{-\gamma}}{Z_\gamma} \sqrt{\frac{d_1}{d}} \right) \right)$.

Theorem 8 tells us that we can write the k-th eigenvector of $\hat{C}$ as:

$$\hat{u}_k = A_k^{(1)} + \underbrace{\sum_{j=1, j\neq k}^{d_1} \frac{(A_j^{(1)})^T (\hat{C} - \mathbb{E}[\hat{C}]) A_k^{(1)}}{\lambda_k - \lambda_j} u_j}_{(I)} + \frac{1}{\lambda_k} P_{\mathrm{Ker}(\mathbb{E}[\hat{C}])} (\hat{C} - \mathbb{E}[C^{(1)}]) u_k + \Delta, \tag{C.29}$$

for $\|\Delta\|_2 = o_d(1)$. Let's focus on (I). We have:

$$(I) = \sum_{j=1, j\neq k}^{d_1} \frac{u_j^T(\hat{C} - \mathbb{E}[\hat{C}])u_k}{\lambda_k - \lambda_j} u_j \tag{C.30}$$

$$= \sum_{j=1, j\neq k}^{d_1} \frac{u_j^T(\frac{1}{n}\sum_{\mu=1}^n y_\mu(F_\mu F_\mu^T - I_D) - \mathbb{E}[\hat{C}])u_k}{\lambda_k - \lambda_j} u_j \tag{C.31}$$

$$= \sum_{j=1, j\neq k}^{d_1} \frac{\frac{1}{n}\sum_{\mu=1}^n y_\mu h_{\mu,j} h_{\mu,k}}{\lambda_k - \lambda_j} u_j. \tag{C.32}$$

From Lemma F.4 in [65], there exists a universal constant C such that with high probability $\|F_\mu\|_2^2 \leq Cd^q$. Then, assuming the eigenvectors of $\mathbb{E}[\hat{C}]$ are de-localized (in the sense that $\|u_k\|_\infty \leq Cd^{-q}$, we will have

$$|h_{\mu,j}| = \langle u_j, F_\mu \rangle \leq C. \tag{C.33}$$

Then:

$$\|(I)\|_2 = \left(\sum_{j\neq k} \frac{1}{(\lambda_k - \lambda_j)^2} \left(\frac{1}{n} \sum_{\mu=1}^n y_\mu h_{\mu,j} h_{\mu,k} \right)^2 \right)^{\frac{1}{2}} \tag{C.34}$$

$$\tag{C.35}$$

By Theorem 26, with high probability:

$$\left(\frac{Z_\gamma}{n} y_\mu h_{\mu,k} h_{\mu,j} \right)^2 \lesssim \frac{Z_\gamma^2}{d^q} \min(k,j)^{-2\gamma}. \tag{C.36}$$

Then:

$$\|(I)\|_2 \leq \left(\frac{Z_\gamma^2}{d^q} \sum_{j\neq k} \frac{\min(k,j)^{-2\gamma}}{(\lambda_k - \lambda_j)^2} \right)^{\frac{1}{2}}. \tag{C.37}$$

We now focus on the inner sum. Recall that $\lambda_j = z_j Z_\gamma j^{-\gamma}$, and that $z_j \sim \mathrm{Rad}(\frac{1}{2})$. Then the inner sum is upper bounded by the case where $z_j \neq z_k$, that is:

$$\sum_{j\neq k} \frac{\min(k,j)^{-2\gamma}}{(\lambda_k - \lambda_j)^2} \leq \frac{1}{Z_\gamma^2} \sum_{j\neq k} \frac{\min(k,j)^{-2\gamma}}{(k^{-\gamma} - j^{-\gamma})^2}. \tag{C.38}$$

By separating the sum according to $j < k$ and $j > k$:

$$\sum_{j\neq k} \frac{\min(k,j)^{-2\gamma}}{(\lambda_k - \lambda_j)^2} \leq \frac{1}{Z_\gamma^2} \left(\sum_{j<k} \frac{j^{-2\gamma}}{(k^{-\gamma} - j^{-\gamma})^2} + \sum_{j>k} \frac{k^{-2\gamma}}{(k^{-\gamma} - j^{-\gamma})^2} \right) \tag{C.39}$$

$$\leq \frac{1}{Z_\gamma^2} \left(\sum_{j<k} \frac{1}{\left(\left(\frac{k}{j}\right)^{-\gamma} - 1 \right)^2} + \sum_{j>k} \frac{1}{\left(1 - \frac{j}{k}^{-\gamma}\right)^2} \right) \tag{C.40}$$

Note that for $\gamma \in (0,1)$, the function $u \to u^\gamma$ is concave. Then, $u^\gamma \leq 1+\gamma(u-1)$, and therefore $1-u^\gamma > \gamma(1-u)$. On the other hand, for $\gamma > 1$ if $u \in (0,1)$, then $u^\gamma < u$ and hence $1-u^\gamma > 1-u$. Either way, we get the following upper-bound:

$$\sum_{j\neq k} \frac{\min(k,j)^{-2\gamma}}{(\lambda_k-\lambda_j)^2} \lesssim \frac{1}{Z_\gamma^2}\left(\sum_{j<k}\frac{1}{\left(\left(\frac{j}{k}\right)^\gamma-1\right)^2}+\sum_{j>k}\frac{1}{\left(1-\left(\frac{k}{j}\right)^\gamma\right)^2}\right) \tag{C.41}$$

Using this fact:

$$\sum_{j\neq k} \frac{\min(k,j)^{-2\gamma}}{(\lambda_k-\lambda_j)^2} \lesssim \frac{1}{Z_\gamma^2}\left(\sum_{j<k}\frac{1}{\left(1-\frac{j}{k}\right)^2}+\sum_{j>k}\frac{1}{\left(1-\frac{k}{j}\right)^2}\right) \tag{C.42}$$

$$\lesssim \frac{1}{Z_\gamma^2}\left(\sum_{j<k}\frac{k^2}{(k-j)^2}+\sum_{j>k}\frac{j^2}{(j-k)^2}\right) \tag{C.43}$$

$$\lesssim \frac{1}{Z_\gamma^2}\left(k^2+d_1\right) \lesssim \frac{1}{Z_\gamma^2}d_1^2. \tag{C.44}$$

Then, going back to Equation (C.37):

$$\|(I)\|_2 \leq \left(\frac{Z_\gamma^2}{d^q}\frac{1}{Z_\gamma^2}d_1^2\right)^{\frac{1}{2}} \lesssim \left(\frac{d_1^2}{d^q}\right)^{\frac{1}{2}}. \tag{C.45}$$

Since $d_1 = d^\varepsilon$, and $\varepsilon < \frac{q}{2}$, we conclude that $\|I\|_2 = o_d(1)$. From this, we can go back to Equation (C.29), and applying inner product with $A_k^{(1)}$, we will get:

$$\langle \hat{u}_k, A_k^{(1)}\rangle = \|A_k^{(1)}\|^2 + \langle (I), A_k^{(1)}\rangle + o_d(1), \tag{C.46}$$

and by Cauchy-Schwarz and Equation (C.45):

$$\left|\langle (I), A_k^{(1)}\rangle\right| \leq \|\hat{A}_k^{(1)}\|_2 \|(I)\|_2 = o_d(1). \tag{C.47}$$

where we used the fact that the norm of $\hat{A}_k^{(1)}$ concentrates around 1 by Hanson-Wright Inequality ([62], Theorem 6.2.2). Then, we conclude:

$$\langle \hat{u}_k, A_k^{(1)}\rangle = 1 + o_d(1). \tag{C.48}$$

Denote

$$\text{Overlap}(a,b) = \frac{\langle a,b\rangle}{\|a\|\|b\|}. \tag{C.49}$$

After normalizing in Equation (C.48), we get:

$$\text{Overlap}(\hat{u}_k, A_k^{(1)}) = \frac{1}{\sqrt{1+\|I\|^2+\left\|\frac{1}{\lambda_k}P_{\text{Ker}(\mathbb{E}[\hat{C}])}(\hat{C}-\mathbb{E}[C^{(1)}])u_k\right\|^2}} \tag{C.50}$$

By applying Taylor expansion to the function $u \to \frac{1}{\sqrt{1+x^2}}$, we get:

$$1 - \text{Overlap}(\hat{u}_k, A_k^{(1)}) = 1 - O\left(\|I\|^2 + \|II\|^2\right) + O((\|I\|^2 + \|II\|^2)^2). \tag{C.51}$$

By Equation (C.45), we know that $\|I\|^2 = o_d(1)$, with a bound independent of n. Thus, the only thing left to conclude is to show that $\|(II)\|$ goes to zero with n, at a rate $\frac{1}{n}$, and we can conclude the first part of Theorem 1. Computing $\|(II)\|^2$, we get:

$$\|(II)\|_2^2 \leq \frac{1}{\lambda_i^2}\|\hat{C} - \mathbb{E}[\hat{C}]\|_{\text{op}}^2. \tag{C.52}$$

By Theorem 7, with high probability:

$$\|\hat{C} - \mathbb{E}[\hat{C}]\|_{\text{op}} \lesssim \sqrt{\frac{d^q \text{poly}\log(d)}{n}}. \tag{C.53}$$

Then:

$$\|(II)\|_2^2 \lesssim \frac{i^{2\gamma} d^q \text{poly}\log(d)}{Z_\gamma^2 n}. \tag{C.54}$$

Then we conclude that, if $n = \Theta(Z_\gamma^2 i^{2\gamma} d^{k+\delta+\varepsilon})$,

$$\text{Overlap}(\hat{u}_k, \hat{A}_k^{(1)}\rangle = 1 - o_d(1), \tag{C.55}$$

thus, we recover the direction $\hat{A}_k^{(1)}$ with $n = \Theta(Z_\gamma^2 i^{2\gamma} d^{k+\delta+\varepsilon})$. As for the rate, Equation (C.54) tells us that it is controlled by the second term, and decays as $\frac{-1}{n}$, which concludes the first part of Theorem 1.

C.4. Analysis of Outliers - Necessary Sample Complexity

Recall that from Theorem 8:

$$\hat{u}_k = A_k^{(1)} + \sum_{j=1, j\neq k}^{d_1} \frac{(A_j^{(1)})^T(\hat{C} - \mathbb{E}[\hat{C}])A_k^{(1)}}{\lambda_k - \lambda_j} A_j^{(1)} + \frac{1}{\lambda_k} P_{\text{Ker}(\mathbb{E}[\hat{C}])}(\hat{C} - \mathbb{E}[C^{(1)}]) + o(\|\hat{C} - \mathbb{E}[C^{(1)}]\|^2). \tag{C.56}$$

In this section, we will prove that the sample complexity we found in the last section is in fact, necessary. For this, we will focus on the last term:

$$w = \frac{1}{\lambda_k} P_{\text{Ker}}(\hat{C} - \mathbb{E}[C^{(1)}])u_k = \frac{1}{\lambda_k} P_{\text{Ker}} \hat{C} u_k. \tag{C.57}$$

We will prove that with high probability, $\|w\| = \Theta_d(1)$. For this, we will use to preliminary results.

Lemma 9 (Paley-Zigmund Inequality, ([13], Exercise 2.4)) *Let Y be real, positive random variable, and $\theta \in (0, 1)$. Then:*

$$\mathbb{P}\left(Y > \theta \mathbb{E}[Y]\right) \geq (1-\theta)^2 \frac{\mathbb{E}[Y^2]}{\mathbb{E}[Y]^2}. \tag{C.58}$$

The Paley-Zigmund inequality will give us a lower bound as long as we can bound the moments of a particular random variable. In our case, the random variables will be polynomials of Gaussians, so we will use Gaussian hypercontractivity.

Lemma 10 (Gaussian Hypercontractivity, [36] Theorem 5.8) *Let X be a N degree polynomial of m gaussian random variables. Then*

$$\mathbb{E}\left[|X|^p\right]^{\frac{1}{p}} \leq \mathbb{E}C(p,N)\left[|X|^2\right]^{\frac{1}{2}},$$

for all $1 < p < \infty$.

Having Theorem 9 and Theorem 10, we are ready to proceed with the proof of Theorem 1. First, we identify the random variable to which we will apply Paley-Zigmund inequality.

By using the definition of $\hat{C}$, we have:

$$w = \frac{1}{n}\sum_{\mu=1}^{n} \frac{y_\mu}{\lambda_k}\langle u_k, F_\mu\rangle P_{\text{Ker}} F_\mu \tag{C.59}$$

$$= \frac{1}{n}\sum_{\mu=1}^{n} Z_\mu. \tag{C.60}$$

Note that

$$\mathbb{E}[Z_\mu] = o_d(1), \tag{C.61}$$

and

$$\mathbb{E}[\|w\|^2] = \frac{\mathbb{E}[\|Z_\mu\|]^2}{n} + o_d. \tag{C.62}$$

Note that the vector $P_{\text{Ker}}F_\mu$ has polynomials in all of it's coordinates, and the degree of all this polynomials is bounded and independent of d. Moreover, y_μ and $\langle u_k, F_\mu\rangle$ are also polynomials of bounded degrees. This makes the random variable $\|Z_\mu\|^2$ a real polynomial of Gaussians, with bounded degree. If we denote by $q_i(x)$ the polynomial in the i-th coordinate of $P_{\text{Ker}}F_\mu$, then:

$$\mathbb{E}\left[\|Z_\mu\|^4\right] = \mathbb{E}\left[\left(\frac{y_\mu}{\lambda_k}\langle u_k, F_\mu\rangle\right)^4 \|P_{\text{Ker}}F_\mu\|^4\right] \tag{C.63}$$

$$= \mathbb{E}\left[\left(\frac{y_\mu}{\lambda_k}\langle u_k, F_\mu\rangle \sum_{i=1}^{D} q_i(x_\mu)^2\right)^4\right]. \tag{C.64}$$

Then, by Gaussian hypercontractivity Theorem 10:

$$\mathbb{E}\left[\|Z_\mu\|^4\right] = \mathbb{E}\left[\left(\frac{y_\mu}{\lambda_k}\langle u_k, F_\mu\rangle \sum_{i=1}^{D} q_i(x_\mu)^2\right)^4\right] \leq C\mathbb{E}\left[\left(\frac{y_\mu}{\lambda_k}\langle u_k, F_\mu\rangle \sum_{i=1}^{D} q_i(x_\mu)^2\right)^2\right]^2 = C\mathbb{E}\left[\|Z_\mu\|^2\right]^2. \tag{C.65}$$

Then, putting together the Paley-Zigmund inequality Theorem 9 and Equation (C.65) for $\|w\|^2$, we get that for $\theta \in (0,1)$, ,

$$\mathbb{P}\left(\|w\|^2 \geq \theta \frac{\mathbb{E}[\|Z_\mu\|^2]}{n}\right) \geq (1-\theta)^2 \frac{\mathbb{E}[\|w\|^2]^2}{\mathbb{E}[\|w\|^4]} \tag{C.66}$$

$$= (1-\theta)^2 \frac{\mathbb{E}[\|Z\|^2]^2}{\mathbb{E}[\|Z\|^4]} + o_d(1) \tag{C.67}$$

$$\geq C(1-\theta)^2 + o_d(1). \tag{C.68}$$

Taking square roots, with probability at least $C(1-\theta)^2 + o_d(1)$

$$\|w\| \geq \theta \sqrt{\frac{\mathbb{E}[\|Z_\mu\|^2]}{n}}. \tag{C.69}$$

The only thing left is to compute $\mathbb{E}[\|Z_\mu\|^2]$. We have:

$$\mathbb{E}[\|Z_\mu\|^2] = \frac{1}{\lambda_k^2}\mathbb{E}\left[y_\mu^2 \langle u_k, F_\mu\rangle^2 \|P_{\text{Ker}} F_\mu\|^2\right]. \tag{C.70}$$

Denote $G_\mu = y_\mu^2 \langle u_k, F_\mu\rangle^2$. Then:

$$\mathbb{E}[\|Z_\mu\|^2] = \frac{1}{\lambda_k^2}\mathbb{E}\left[G_\mu \|P_{\text{Ker}} F_\mu\|^2\right]. \tag{C.71}$$

Note that $\mathbb{E}[G_\mu] = \Theta(1)$, and we can write $P_{\text{Ker}} = I_D - P_U$, for P_U the projection into the space spanned by $u_1, \ldots, u_{d_1}$. Then:

$$\mathbb{E}[\|Z_\mu\|^2] = \frac{1}{\lambda_k^2}\mathbb{E}\left[G_\mu \|F_\mu\|^2\right] - \frac{1}{\lambda_k^2}\mathbb{E}\left[G_\mu \|P_U F_\mu\|^2\right] \tag{C.72}$$

Now, define the event $\mathcal{A} := \{\|F_\mu\|^2 \geq \frac{D}{2}\}$. Then:

$$\mathbb{E}\left[G_\mu \|F_\mu\|^2\right] \geq \mathbb{E}\left[G_\mu \|F_\mu\|^2 \mathbf{1}_{\mathcal{A}}\right] \tag{C.73}$$

$$\geq \frac{D}{2}\mathbb{E}\left[G_\mu \mathbf{1}_{\mathcal{A}}\right] = \frac{D}{2}\left(\mathbb{E}\left[G_\mu\right] - \mathbb{E}\left[G_\mu \mathbf{1}_{A^c}\right]\right) \tag{C.74}$$

$$\geq \frac{D}{2}\mathbb{E}\left[G_\mu\right] + o_d(1), \tag{C.75}$$

where the last line follows from Theorem 3. Doing the same for the term $\mathbb{E}\left[G_\mu \|P_U F_\mu\|^2\right]$, we get:

$$\mathbb{E}\left[G_\mu \|F_\mu\|^2\right] \geq \frac{D}{\lambda_k^2} + \frac{d_1}{\lambda_k^2} \tag{C.76}$$

Then, taking $n \ll \frac{d^q i^{2\gamma}}{Z_\gamma}$ and going back to Equation (C.69), with probability at least $C(1-\theta)^2$:

$$\|w\| \geq \theta \sqrt{\frac{d^q}{\lambda_k^2 n}} = C\theta d^\delta, \tag{C.77}$$

and taking $\theta = d^{-\delta}$, we get that with probability at least $1 - o_d(1)$ the norm of $\|w\|$ is $\Theta(1)$. Then, since we had:

$$\hat{u}_k = u_k + \sum_{j=1, j\neq k}^{d_1} \frac{u_j^T(\hat{C} - \mathbb{E}[\hat{C}])u_k}{\lambda_k - \lambda_j} u_j + w + o(\|\hat{C} - \mathbb{E}[C^{(1)}]\|^2), \tag{C.78}$$

if $n \ll \frac{d^q i^{2\gamma}}{Z_\gamma}$, there exists a constant probability event such that $\hat{u}_k$ is not aligned with u_k.

Appendix D. Derivation of the rates

In this section, we build on Theorem 1 to derive the recovered-direction count and then the MSE rates stated in the main theorem. In the sharp-threshold approximation, Algorithm 1 either learns a direction $A_i^{(1)}$ or does not learn it.

Since by assumption of Theorem 1, the function g is a polynomial, it suffices to focus on the case where g is the identity, as the other terms will be sub-leading after doing KRR. Since at this point, Algorithm 1 has constructed features which are low-dimensional, we have that the MSE after doing Kernel Ridge Regression is:

$$\text{MSE} = \mathbb{E}\left[\left\|\sum_{j\geq i^\star} j^{-\gamma}\left((h_j^{(1)})^2 - 1\right)\right\|^2\right]. \tag{D.1}$$

By Theorem 1, we learn the $i-th$ direction if

$$n \asymp \frac{d^q i^{2\gamma}}{Z_\gamma^2} \implies i \asymp \left(\frac{Z_\gamma^2 n}{d^q}\right)^{\frac{1}{2\gamma}}. \tag{D.2}$$

Since directions are learned sequentially, the number of learn directions (or the last direction that was recovered) at sample complexity n is:

$$i^\star = \left(\frac{Z_\gamma^2 n}{d^q}\right)^{\frac{1}{2\gamma}}. \tag{D.3}$$

This gives the recovered-direction count used in the main theorem. We now derive the corresponding MSE rates. By Equation (D.1):

$$\text{MSE}(n) = \Theta\left(Z_\gamma^2 \sum_{i\geq i^\star} j^{-2\gamma}\right). \tag{D.4}$$

We now study this sum according to the value of γ. If $\gamma < \frac{1}{2}$, then $Z_\gamma = (d_1^{1-2\gamma})^{-\frac{1}{2}}$, and we get:

$$Z_\gamma^2 \sum_{i \geq i^\star} j^{-2\gamma} = Z_\gamma^2 (Z_\gamma^{-2} - \sum_{i \leq i^\star} i^{-2\gamma}) \tag{D.5}$$

$$= 1 - \frac{1}{d_1^{1-2\gamma}} (i^\star)^{1-2\gamma} \tag{D.6}$$

$$= 1 - \frac{1}{d_1^{1-2\gamma}} \left(\frac{Z_\gamma^2 n}{d^q} \right)^{-1+\frac{1}{2\gamma}} \tag{D.7}$$

$$= 1 - \left(\frac{n}{d_1 d^q} \right)^{(-1+\frac{1}{2\gamma})}, \tag{D.8}$$

which gives the rate for the case where $\gamma < \frac{1}{2}$ and $d^q \ll n \ll d^q d_1$.

On the other hand, for $\gamma > \frac{1}{2}$, $Z_\gamma = \Theta_d(1)$ and

$$\sum_{i \geq i^\star} i^{-2\gamma} = \Theta_d((i^\star)^{1-2\gamma}). \tag{D.9}$$

Then:

$$\mathrm{MSE}(n) = \Theta_d \left(\sum_{i \geq i^\star} i^{-2\gamma} \right) \tag{D.10}$$

$$= \Theta_d((i^\star)^{1-2\gamma}) \tag{D.11}$$

$$= \left(\frac{n}{d^q} \right)^{\frac{1-2\gamma}{2\gamma}} \tag{D.12}$$

$$= \left(\frac{n}{d^q} \right)^{-1+\frac{1}{2\gamma}}, \tag{D.13}$$

and we conclude Theorem 1.

Appendix E. Explicit computations with Weiner Chaos

E.1. Wiener Chaos properties

This section will only overview the necessary concepts we need from Wiener chaos expansions. This results are based on [52], [49] and [65].

For $A \in \mathbb{R}^{B(d,q)}$, we define

$$I_q(A) = \langle A, \mathcal{F}(\mathrm{He}_q(x)) \rangle, \tag{E.1}$$

where for $\beta \in \mathbb{Z}_{\geq 0}^d$ with $|\beta| = q$

$$(\mathrm{He}_q(x))_\beta = \mathrm{He}_\beta(x). \tag{E.2}$$

Lemma 11 (Orthogonality of Different Chaos) *Let $q, q' \in \mathbb{N}$, $A \in (\mathbb{R}^d)^{\odot q}$, $B \in (\mathbb{R}^d)^{\odot q'}$. Then:*

$$\mathbb{E}\left[I_q(A) I_{q'}(B) \right] = \mathbf{1}_{q=q'} \langle A, B \rangle.$$

The space spanned by random variables in the k-th Wiener chaos is denoted by $\mathcal{H}_k$. We will also need the orthogonal projection into the k-Wiener chaos, which we denote by $J_k : L^2 \to \mathcal{H}_k$.

We define the Malliavin derivative $D : \mathrm{dom}(D) \to (L^2)^d$ by

$$DF = (\partial_{x_1} F, \dots, \partial_{x_d} F), \qquad \mathrm{dom}(D) = \left\{ F \in L^2 : \sum_{k=0}^{\infty} k \|J_k(F)\|^2_{L^2(\mathcal{G})} < \infty \right\}$$

where, for smooth functions F with compact support, ∂_{x_j} is the usual partial derivative of $F(\boldsymbol{x}) = F(x_1, \dots, x_d)$ in the variable x_j, and this is extended by completion to $\mathrm{dom}(D)$. We also define the Ornstein-Uhlenbeck infinitesimal generator $L : \mathrm{dom}(L) \to L^2(\mathcal{G})$ by

$$LF = \sum_{k=0}^{\infty} -k\, J_k(F), \qquad \mathrm{dom}(L) = \left\{ F \in L^2(\mathcal{G}) : \sum_{k=0}^{\infty} k^2 \|J_k(F)\|^2_{L^2(\mathcal{G})} < \infty \right\}.$$

and we define its inverse $L^{-1}F = \sum_{k=1}^{\infty} -\frac{1}{k} J_k(F)$ whenever $J_0(F) = \mathbb{E}F(x) = 0$.

We will need the following rules to compute the different terms that appears.

Lemma 12 (Product formula, [49]) *For any $k, \ell \geq 1$, $S \in (\mathbb{R}^d)^{\odot k}$, and $T \in (\mathbb{R}^d)^{\odot \ell}$,*

$$I_k(S) I_\ell(T) = \sum_{r=0}^{\min(k,\ell)} r! \binom{k}{r} \binom{\ell}{r} I_{k+\ell-2r}(S \tilde{\otimes}_r T).$$

We will also need the following rule for computing product of derivatives.

Lemma 13 *For any $k, \ell \geq 1$, $S \in (\mathbb{R}^d)^{\odot k}, T \in (\mathbb{R}^d)^{\odot \ell}$,*

$$(DI_k(S))^{\mathrm{T}} (DI_\ell(T)) = k\ell \sum_{r=1}^{\min(k,\ell)} (r-1)! \binom{k-1}{r-1} \binom{\ell-1}{r-1} I_{k+\ell-2r}(S \tilde{\otimes}_r T)$$

For some functions, specially polynomials, the following Gaussian Integration by Parts Lemma will be very useful.

Lemma 14 (Theorem 2.9.1 in [49], Gaussian Integration by Parts) *Let $F, G \in \mathbb{D}^{1,2}$, and let $g : \mathbb{R} \to \mathbb{R}$ be a C^1 function having a bounded derivative. Then*

$$\mathbb{E}[Fg(G)] = \mathbb{E}[F]\mathbb{E}[g(G)] + \mathbb{E}[g'(G)\langle DG, -DL^{-1}F \rangle_{\mathfrak{H}}].$$

E.2. Computing Expectations with Malliavin Calculus

To compute expectations, we will need:

Remark 15 *As the discussion in [49], Page 31 notes, the conditions under which Theorem 14 hold are not optimal. In particular, it remains true if g is a polynomial.*

Denote the second layer by $h^{(2)} = \sum_{i=1}^{d_1} \lambda_i (I_q(A_i)^2 - 1)$, and let $g : \mathbb{R} \to \mathbb{R}$ be a polynomial. In this section, we will study how to compute expectations of the form:

$$\mathbb{E}\left[g(h^{(2)}) I_q(A_j) I_q(A_k)\right], \tag{E.3}$$

for $j \neq k$. We are interested in how this quantity scales with d, as sharp as possible. In particular, we will avoid using Gaussian Approximations when is not absolutely necessary.

E.3. Linear Case

To get some intuition, we will first study the case $g(u) = u$. We want to compute:

$$E_{\text{lin}} := \mathbb{E}\left[S I_q(A_j) I_q(A_k)\right] = \sum_{i=1}^{d_1} \lambda_i \mathbb{E}\left[(I_q(A_i)^2 - 1) I_q(A_j) I_q(A_k)\right]. \tag{E.4}$$

By Theorem 12:

$$I_q(A_i)^2 - 1 = \sum_{r=0}^{q} c_{q,r} I_{2q-2r}(A_i \tilde{\otimes}_r A_i) - 1 \tag{E.5}$$

$$= \sum_{r=0}^{q-1} c_{q,r} I_{2q-2r}(A_i \tilde{\otimes}_r A_i) + (\|A_i\|_2^2 - 1), \tag{E.6}$$

where $c_{q,r} = r!\binom{q}{r}^2$ Analogously

$$I_q(A_j) I_q(A_k) = \sum_{r=0}^{q} c_{q,r} I_{2q-2r}(A_j \tilde{\otimes}_r A_k) = \sum_{r=0}^{q-1} c_{q,r} I_{2q-2r}(A_j \tilde{\otimes}_r A_k) + q! \langle A_j, A_k \rangle. \tag{E.7}$$

Then, by the orthogonality of different chaos:

$$E_{\text{lin}} = \sum_{i=1}^{d_1} \lambda_i \mathbb{E}\left[(I_q(A_i)^2 - 1) I_q(A_j) I_q(A_k)\right] \tag{E.8}$$

$$= \sum_{i=1}^{d_1} \lambda_i \mathbb{E}\left[\left(\sum_{r=0}^{q-1} c_{q,r} I_{2q-2r}(A_i \tilde{\otimes}_r A_i) + (\|A_i\|_2^2 - 1)\right)\left(\sum_{r=0}^{q-1} c_{q,r} I_{2q-2r}(A_j \tilde{\otimes}_r A_k) + q! \langle A_j, A_k \rangle\right)\right] \tag{E.9}$$

$$= \sum_{i=1}^{d_1} \lambda_i \sum_{r=0}^{q-1} c_{q,r}^2 \langle A_i \tilde{\otimes}_r A_i, A_j \tilde{\otimes}_r A_k \rangle + \sum_{i=1}^{d} \lambda c_{q,q}^2 (\|A_i\|^2 - 1) \langle A_j, A_k \rangle \tag{E.10}$$

$$= \sum_{i=1, i \notin \{j,k\}}^{d_1} \lambda_i c_{q,r}^2 \sum_{r=0}^{q-1} \langle A_i \tilde{\otimes}_r A_i, A_j \tilde{\otimes}_r A_k \rangle + \sum_{i \in \{j,k\}}^{d_1} \lambda_i \sum_{r=0}^{q-1} c_{q,r}^2 \langle A_i \tilde{\otimes}_r A_i, A_j \tilde{\otimes}_r A_k \rangle + \sum_{i=1}^{d} \lambda c_{q,q}^2 (\|A_i\|^2 - 1) \langle A_j, A \tag{E.11}$$

were in the last line we split the sum according to wether $i \in \{j, k\}$ or not. Recall we assume that $A_i \in (\mathbb{R}^d)^{\odot q}$ have independent, centered gaussian entries with variance $\frac{1}{d^q}$. Let:

$$T^r_{i,j,k} = \langle A_i \tilde{\otimes}_r A_i, A_j \tilde{\otimes}_r A_k \rangle. \tag{E.12}$$

For $u, v \in [d]^{q-r}$ and $r \in \{0\} \cup [q-1]$we have:

$$(A_i \tilde{\otimes}_r A_i)_{u,v} = \sum_{\ell \in [d]^r} A_i[u, \ell] A_i[v, \ell], \quad \text{and } (A_j \tilde{\otimes}_r A_k)_{u,v} = \sum_{\ell \in [d]^r} A_j[u, \ell] A_k[v, \ell]. \tag{E.13}$$

Then:

$$T^r_{i,j,k} = \sum_{u,v \in [d]^{q-r}} (A_i \tilde{\otimes}_r A_i)_{u,v} (A_j \tilde{\otimes}_r A_k)_{u,v} \tag{E.14}$$

$$= \sum_{u,v \in [d]^{q-r}} \sum_{\ell_1 \in [d]^r} A_i[u, \ell_1] A_i[v, \ell_1] \sum_{\ell_2 \in [d]^r} A_j[u, \ell_2] A_k[v, \ell_2] \tag{E.15}$$

$$= \sum_{u,v \in [d]^{q-r}} \sum_{\ell_2 \in [d]^r} \left(\sum_{\ell_1 \in [d]^r} A_i[u, \ell_1] A_i[v, \ell_1] \right) A_j[u, \ell_2] A_k[v, \ell_2]. \tag{E.16}$$

If $i \neq j \neq k$, then we have:

$$\mathbb{E}_{A_j} [T_r | A_i, A_k] = 0. \tag{E.17}$$

On the other hand, the conditional variance equals:

$$\mathrm{Var} (T_r | A_i, A_k) = \frac{1}{d^q} \sum_{u,v \in [d]^{q-r}} \left(\sum_{\ell_2 \in [d]^r} \left(\sum_{\ell_1 \in [d]^r} A_i[u, \ell_1] A_i[v, \ell_1] \right) A_k[u, \ell_2] \right)^2. \tag{E.18}$$

Taking expectation with respect to A_k:

$$\mathbb{E}_{A_j} [\mathrm{Var} (T_r | A_i, A_k)] = \frac{d^r}{d^{2q}} \sum_{u,v \in [d]^{q-r}} \left(\sum_{\ell_1 \in [d]^r} A_i[u, \ell_1] A_i[v, \ell_1] \right)^2 = \frac{d^r}{d^{2q}} \| A_i \otimes_r A_i \|_2^2. \tag{E.19}$$

And taking expectation with respect to A_i, we have:

$$\mathbb{E}_{A_i} \left[\| A_i \otimes_r A_i \|_2^2 \right] = \mathbb{E}_{A_i} \left[\sum_{u,v \in [d]^{q-r}} \left(\sum_{\ell_1 \in [d]^r} A_i[u, \ell_1] A_i[v, \ell_1] \right)^2 \right] \tag{E.20}$$

$$= \sum_{u,v \in [d]^{q-r}} \mathbb{E}_{A_i} \left[\left(\sum_{\ell_1 \in [d]^r} A_i[u, \ell_1] A_i[v, \ell_1] \right)^2 \right] \tag{E.21}$$

$$= O \left(\sum_{u \in [d]^{q-r}} \frac{1}{d^q} \right) = O_d(d^{-r}). \tag{E.22}$$

Applying the Law of total variance, using the fact that for distinct i, j, k, $T_{i,j,k}$ is centered:

$$\mathrm{Var}_{A_i,A_j,A_k}(T_{i,j,k}) = O\left(\frac{d^r}{d^{2q}d^r}\right) = O_d\left(\frac{1}{d^{2q}}\right). \tag{E.23}$$

Then, by applying Chebyshev Inequality we get that with high probability with respect to A_i, A_j and A_k:

$$\left|T^r_{i,j,k}\right| = O_d\left(\frac{1}{d^q}\right), \text{ when } i \neq j \neq k. \tag{E.24}$$

We now move to the harder case where $i = j$ (the case $i = k$ is analogous). First, by definition:

$$T^r_{i,i,j} = \sum_{u,v\in[d]^{q-r}} \sum_{\ell_2\in[d]^r} \left(\sum_{\ell_1\in[d]^r} A_i[u,\ell_1]A_i[v,\ell_1]\right) A_i[u,\ell_2]A_j[v,\ell_2]. \tag{E.25}$$

Fixing A_i:

$$\mathbb{E}_{A_j}[T^r_{i,i,j}] = 0. \tag{E.26}$$

The conditional variance given A_i equals:

$$\mathrm{Var}(T^r_{i,i,j}|A_i) = \frac{1}{d^q} \sum_{u,v\in[d]^{q-r}} \sum_{\ell_2\in[d]^r} \left(\left(\sum_{\ell_1\in[d]^r} A_i[u,\ell_1]A_i[v,\ell_1]\right) A_i[u,\ell_2]\right)^2. \tag{E.27}$$

Denote $w(u,v) = \sum_{\ell_1\in[d]^r} A_i[u,\ell_1]A_i[v,\ell_1]$. Then:

$$\sum_{u,v\in[d]^{q-r}} \sum_{\ell_2\in[d]^r} \left(\left(\sum_{\ell_1\in[d]^r} A_i[u,\ell_1]A_i[v,\ell_1]\right) A_i[u,\ell_2]\right)^2 = \sum_{v\in[d]^{q-r}} \sum_{\ell_2\in[d]^r} \sum_{u\in[d]^{q-r}} w(u,v)^2 A_i[u,\ell_2]^2 \tag{E.28}$$

$$\leq \sum_{v\in[d]^{q-r}} \sum_{\ell_2\in[d]^r} \left(\sum_{u\in[d]^{q-r}} w(u,v)^2\right)\left(\sum_{u\in[d]^{q-r}} A_i[u,\ell_2]^2\right) \tag{E.29}$$

where in the last line we applied Cauchy-Schwarz. Now, taking expectation and then applying Cauchy-Schwarz again:

$$\mathbb{E}_{A_i}\left[\mathrm{Var}(T^r_{i,i,j}|A_i)\right] \leq \sum_{v\in[d]^{q-r}} \sum_{\ell_2\in[d]^r} \mathbb{E}_{A_i}\left[\left(\sum_{u\in[d]^{q-r}} w(u,v)^2\right)\left(\sum_{u\in[d]^{q-r}} A_i[u,\ell_2]^2\right)\right] \tag{E.30}$$

$$\leq \sum_{v\in[d]^{q-r}} \sum_{\ell_2\in[d]^r} \mathbb{E}_{A_i}\left[\left(\sum_{u\in[d]^{q-r}} w(u,v)^2\right)^2\right]^{\frac{1}{2}} \mathbb{E}\left[\left(\sum_{u\in[d]^{q-r}} A_i[u,\ell_2]^2\right)^2\right]^{\frac{1}{2}}. \tag{E.31}$$

By the equivalence of norms for polynomials ([36], Theorem 3.50):

$$\mathbb{E}_{A_i}\left[\left(\sum_{u\in[d]^{q-r}} w(u,v)^2\right)^2\right] \le C\mathbb{E}_{A_i}\left[\left(\sum_{u\in[d]^{q-r}} w(u,v)^2\right)\right]^2, \tag{E.32}$$

and the same holds for $\mathbb{E}\left[\left(\sum_{u\in[d]^{q-r}} A_i[u,\ell_2]^2\right)^2\right]$. Hence:

$$\mathbb{E}_{A_i}\left[\mathrm{Var}(T^r_{i,i,j}|A_i)\right] \le C\sum_{v\in[d]^{q-r}}\sum_{\ell_2\in[d]^r}\mathbb{E}_{A_i}\left[\left(\sum_{u\in[d]^{q-r}} w(u,v)^2\right)\right]\mathbb{E}\left[\left(\sum_{u\in[d]^{q-r}} A_i[u,\ell_2]^2\right)\right] \tag{E.33}$$

$$\le C\sum_{\ell_2\in[d]^r}\mathbb{E}_{A_i}\left[\left(\sum_{v\in[d]^{q-r}}\sum_{u\in[d]^{q-r}} w(u,v)^2\right)\right]\mathbb{E}\left[\left(\sum_{u\in[d]^{q-r}} A_i[u,\ell_2]^2\right)\right] \tag{E.34}$$

By Equation (E.22):

$$\mathbb{E}\left[\sum_{v\in[d]^{q-r}}\sum_{u\in[d]^{q-r}} w(u,v)^2\right] = \sum_{u\in[d]^{q-r}}\mathbb{E}\left(\sum_{\ell_1\in[d]^r} A_i[u,\ell_1]A_i[v,\ell_1]\right)^2 \tag{E.35}$$

$$= O\left(\frac{1}{d^r}\right), \tag{E.36}$$

and

$$\mathbb{E}\left[\left(\sum_{u\in[d]^{q-r}} A_i[u,\ell_2]^2\right)\right] = O\left(\frac{d^{q-r}}{d^q}\right) = O\left(\frac{1}{d^r}\right). \tag{E.37}$$

Then:

$$\mathbb{E}_{A_i}\left[\mathrm{Var}(T^r_{i,i,j}|A_i)\right] \le \frac{C}{d^q}\sum_{\ell_2\in[d]^r}\frac{1}{d^r}\frac{1}{d^r} = O\left(\frac{1}{d^{q+r}}\right) \tag{E.38}$$

Hence, by the Law of total variance:

$$\mathrm{Var}_{A_i,A_j}\left(T^r_{i,i,j}\right) = O\left(\frac{1}{d^{q+r}}\right), \tag{E.39}$$

and applying Chebyshev Inequality we get that with high probability over A_i, A_j:

$$\left|T^r_{i,i,j}\right| = O\left(\frac{1}{d^{\frac{q+r}{2}}}\right). \tag{E.40}$$

Replacing Equation (E.24) and Equation (E.40) in Equation (E.11):

$$|E_{\text{lin}}| = \sum_{i=1, i\notin\{j,k\}}^{d_1} \lambda_i \sum_{r=0}^{q-1} c_{q,r}^2 T_{i,j,k}^r + \sum_{i\in\{j,k\}}^{d_1} \lambda_i \sum_{r=0}^{q-1} c_{q,r}^2 T_{i,j,k}^r + \sum_{i=1}^{d} \lambda_i c_{q,q}^2 (\|A_i\|^2 - 1)\langle A_j, A_k\rangle \tag{E.41}$$

$$= O\left(\frac{1}{d^q} \sum_{i=1, i\notin\{j,k\}}^{d_1} \lambda_i\right) + O\left(\frac{1}{d^q} \sum_{i\in\{j,k\}}^{d_1} \lambda_i\right) + \sum_{i=1}^{d} \lambda_i c_{q,q}^2 (\|A_i\|^2 - 1)\langle A_j, A_k\rangle. \tag{E.42}$$

For the last term, applying Bernstein's inequality ([62], Theorem 2.9.1) for the cross inner product and Hanson-Wright ([62], Theorem 6.2.2) for the norm, we get that with high probability:

$$(\|A_i\|^2 - 1)\langle A_j, A_k\rangle = O\left(\frac{1}{d^q}\right). \tag{E.43}$$

Finally, recalling that $\lambda_i = Z_\gamma z_i i^{-\gamma}$, with $z_i \sim \text{Rad}(\frac{1}{2})$, we can apply Bernstein's inequality over the Radamacher variables. We get that, with high probability over the z_i:

$$|E_{\text{lin}}| = O\left(\frac{Z_\gamma}{d^q}\sqrt{\sum_{i=1, i\notin\{j,k\}}^{d_1} i^{-2\gamma}}\right) + O\left(\frac{Z_\gamma}{d^{\frac{q}{2}}}\sqrt{\sum_{i\in\{j,k\}}^{d_1} i^{-2\gamma}}\right) + \frac{Z_\gamma}{d^q}\sqrt{\sum_{i=1}^{d} i^{-2\gamma}}. \tag{E.44}$$

By definition $Z_\gamma = (\sum_{i=1}^{d_1} i^{-2\gamma})$, so we finally conclude that with very high probability:

$$|E_{\text{lin}}| = O\left(\frac{Z_\gamma}{d^q}(j^\gamma + k^\gamma))\right). \tag{E.45}$$

Thus, we have proved the following Lemma.

Lemma 16 *Let $S = \sum_{i=1}^{d_1} \lambda_i (I_q(A_i)^2 - 1)$. Then, given $i, j \in [d_1]$ with $i \neq j$:*

$$|\mathbb{E}\left[S I_q(A_j) I_q(A_k)\right]| = O\left(\frac{Z_\gamma}{d^{\frac{q}{2}}}(j^\gamma + k^\gamma))\right).$$

E.4. The non-linear case

Let $g : \mathbb{R} \to \mathbb{R}$ be a polynomial of degree m. Following Theorem 1, we now assume $\gamma > \frac{1}{2}$.

We will now study the order in terms of d of the expectation:

$$\mathbb{E}\left[g(h^{(2)}) I_q(A_j) I_q(A_k)\right], \tag{E.46}$$

where we recall that we denote $S = \sum_{i=1}^{d_1} \lambda_i (I_q(A_i)^2 - 1)$. Since g is a polynomial, we can no longer apply the orthogonality of different chaos in the same way we did for the linear case in Equation (E.11). To compute this, we will use Theorem 14.

E.4.1. FIRST STEP: APPLYING GAUSSIAN INTEGRATION BY PARTS

This section follows the construction made in Chapter 8 in [49], tailored to our setting. The objective is to derive Theorem 17. The reader may skip this subsection on a first reading.

Denote $F_1 = S$, $F_2 = I_q(A_j)I_q(A_k)$, for $j \neq k$. Then, our expectation has the form:

$$\mathbb{E}\left[g(F_1)F_2\right]. \tag{E.47}$$

Applying Theorem 14 once, we get:

$$\mathbb{E}\left[g(F_2)F_1\right] = \underbrace{\mathbb{E}\left[g'(F_1)\langle DF_1, -DL^{-1}(F_2 - \mathbb{E}[F_2])\rangle\right]}_{M:=} + \mathbb{E}[g(F_1)]\mathbb{E}[F_2]. \tag{E.48}$$

We focus on the first term. Denote

$$V_1 = \langle DF_1, -DL^{-1}F_2\rangle. \tag{E.49}$$

Let $v_1 = \mathbb{E}[V_1]$, and denote $\bar{V}_1 = V_1 - v_1$. Then:

$$M = \mathbb{E}\left[g'(F_1)\bar{V}_1\right] + \mathbb{E}[g'(F_1)]v_1. \tag{E.50}$$

Then, applying Theorem 14 again:

$$M = \mathbb{E}\left[g^{(2)}(F_1)\underbrace{\langle DF_1, DL^{-1}\bar{V}_1\rangle}_{V_2}\right] + \mathbb{E}[g'(F_1)]v_1. \tag{E.51}$$

We now denote $v_2 = \mathbb{E}\left[V_2\right]$, and $\bar{V}_2 = V_2 - v_2$. Applying Theorem 14 again:

$$M = \mathbb{E}\left[g^{(2)}(F_1)\bar{V}_2\right] + \mathbb{E}\left[g^{(2)}(F_1)\right]v_2 + \mathbb{E}[g'(F_1)]v_1 \tag{E.52}$$

$$= \mathbb{E}\left[g^{(3)}(F_1)\langle DF_1, DL^{-1}\bar{V}_2\rangle\right] + \mathbb{E}[g^{(2)}(F_1)]v_2 + \sum_{i=1}^{d}\lambda_i v_1. \tag{E.53}$$

Iterating $\deg(g) - 1$ times, we get:

$$\mathbb{E}\left[g(h^{(2)})I_q(A_j)I_q(A_k)\right] = \sum_{r=0}^{\deg(g)-1}\mathbb{E}[g^{(r)}(F_1)]v_r \tag{E.54}$$

where we inductively defined

$$V_{r+1} = \mathbb{E}\left[\langle DF_1, DL^{-1}(V_r - v_r))\rangle\right], \quad v_r = \mathbb{E}\left[V_r\right], \tag{E.55}$$

and $V_0 = I_q(A_j)I_q(A_j)$. Note that the objects $V_1, \dots V_{\deg(g)}$ are exactly the ones that appear in Theorem 31. This can be made precise. We actually have:

$$V_r = \Gamma_{F_2,\underbrace{F_1,\dots,F_1}_{r \text{ times}}}. \tag{E.56}$$

By Theorem 32, one can relate the expectation of this variables to cumulants. To be precise, we have that for $r \in \mathbb{N}$:

$$\kappa_r(F_2, \underbrace{F_1, \dots, F_1}_{r\text{times}}) = \sum_{\sigma \in \mathfrak{S}_{\{2,\dots,r\}}} \mathbb{E}\left[\Gamma_{F_2, F_1, \dots, F_1}(F)\right],$$

and since the expectation is repeated, Equation (E.56) allows us to conclude the relation:

$$\kappa_r(F_2, \underbrace{F_1, \dots, F_1}_{r\text{times}}) = r!\mathbb{E}\left[\Gamma_{F_2, F_1, \dots, F_1}(F)\right] = r!\mathbb{E}\left[V_r\right] = r! v_r. \tag{E.57}$$

Then, we conclude:

Lemma 17 *Let $S = \sum_{i=1}^{d_1} \lambda_i (I_q(h^{(2)})^2 - 1)$, and let $g : \mathbb{R} \to \mathbb{R}$ be a polynomial. Then, given $P, Q \in (\mathbb{R}^d)^{\odot q}$*

$$\mathbb{E}\left[g(h^{(2)}) I_q(P) I_q(Q)\right] = \sum_{r=0}^{\deg(g)} \mathbb{E}\left[g^{(r)}(h^{(2)})\right] \frac{1}{r!} \kappa_r(I_q(P) I_q(Q), \underbrace{S, \dots, S}_{r\text{times}}).$$

Remark 18 *This type of result is a generalization of Eq. 8.5.1 in [49] to the multi-variate case. Note that it works for any symmetric tensor.*

E.4.2. STEP 2: COMPUTING THE ORDER OF THE CUMULANTS

By Theorem 17, our problem is reduced to computing cumulants. In particular, we want to derive bounds for

$$v_r = \frac{1}{r!} \kappa \left(\underbrace{S, S, \dots, S}_{r}, (I_q(A_k) I_q(A_j) - q! \langle A_j, A_k \rangle) \right). \tag{E.58}$$

A key property of cumulants is that they are multi-linear. Since the variable $S = \sum_{i=1}^{d_1} \lambda_i (I_q(A_i)^2 - 1)$, this allows us to exchange this sums with the cumulants. Denote

$$Y_{i_\ell} = (I_q(A_{i_\ell})^2 - 1), G_{i,j} = I_q(A_j) I_q(A_k) - \langle A_j, A_k \rangle.$$

By the multi-linearity of the cumulants:

$$v_r = \frac{1}{r!} \sum_{i_1, \dots, i_r = 1}^{d_1} \lambda_{i_1} \cdots \lambda_{i_r} \kappa\left(G_{j,k}, Y_{i_1}, Y_{i_2}, \dots, Y_{i_r}\right). \tag{E.59}$$

By expanding each Y_ℓ into its chaos decomposition:

$$Y_{i_\ell} = I_q(A_{i_\ell})^2 - 1 = \sum_{L=0}^{q-1} c_{q,L} I_{2q-2L}(A_i \otimes_L A_i) + (\|A_i\|^2 - 1), \tag{E.60}$$

and for $G_{j,k}$:

$$I_q(A_j) I_q(A_k) - \langle A_j, A_k \rangle = \sum_{L=0}^{q-1} c_{q,L} I_{2q-2L}(A_j \tilde{\otimes}_L A_k). \tag{E.61}$$

With this, we can go further with the multi-linearity of the cumulants to obtain:

$$v_r = \frac{1}{r!}\sum_{i_1,\dots,i_r=1}^{d_1}\lambda_{i_1}\cdots\lambda_{i_r}\sum_{s_1,\dots,s_r=0}^{q-1}\sum_{t=0}^{q-1}c_{q,s,t}\kappa\left(I_{2q-2s_1}(A_{i_1}\tilde{\otimes}_{s_1}A_{i_1}),\dots,I_{2q-2s_r}(A_{i_r}\tilde{\otimes}_{s_r}A_{i_r}),I_{2q-2t}(A_j\tilde{\otimes}_t A_k)\right), \tag{E.62}$$

where we ignored the expectation terms in Equation (E.60) since they become negligible. Denote $f_i^s = A_i\tilde{\otimes}_s A_i, i \le r$, and $f_{r+1}^s = A_j\tilde{\otimes}_s A_k$ so that:

$$v_r = \frac{1}{r!}\sum_{i_1,\dots,i_r=1}^{d_1}\lambda_{i_1}\cdots\lambda_{i_r}\sum_{s_1,\dots,s_r=0}^{q-1}\sum_{t=0}^{q-1}c_{q,s,t}\kappa\left(I_{2q-2s_1}(f_{i_1}^{(s_1)}),\dots,I_{2q-2s_r}(f_{i_r}^{(s_r)}),I_{2q-2t}(f_{r+1}^t)\right). \tag{E.63}$$

Now, by Theorem 33, we can write:

$$\kappa\left(I_{2q-2s_1}(f_{i_1}^{(s_1)}),\dots,I_{2q-2s_r}(f_{i_r}^{(s_r)}),I_{2q-2t}(f_{r+1}^t)\right) = \sum_{\sigma\in\mathfrak{S}_{\{2,\dots,|m|\}}}(q_{\lambda_\sigma(|m|)})!\sum_{*}c_{q,l,\sigma}(a_2,\dots,a_{|m|-1}) \tag{E.64}$$

$$\langle(\dots((f_{i_{\lambda(1)}}\widetilde{\otimes}_{r_2}f_{i_{\lambda_\sigma(2)}})\widetilde{\otimes}_{r_3}f_{\lambda_\sigma(3)})\dots)\widetilde{\otimes}_{r_{|m|-1}}f_{\lambda_\sigma(|m|-1)};f_{\lambda_\sigma(|m|)} \tag{E.65}$$

where the second sum runs over combinations of indices having technical conditions (all specified in Theorem 33. We ignored the upper-indices to avoid overloading the notation. The important message of Equation (E.65) is there is a finite set of possible contractions (in particular, a a set of size $O_d(1)$), and we are summing over all of them and all possible permutations of indices. Denote

$$T_\lambda = \langle(\dots((f_{i_{\lambda(1)}}\widetilde{\otimes}_{r_2}f_{i_{\lambda_\sigma(2)}})\widetilde{\otimes}_{r_3}f_{\lambda_\sigma(3)})\dots)\widetilde{\otimes}_{r_{|m|-1}}f_{\lambda_\sigma(|m|-1)};f_{\lambda_\sigma(|m|)}\rangle. \tag{E.66}$$

Thus, we have concluded:

Lemma 19 *Let $h^{(2)} = \sum_{i=1}^{d_1}\lambda_i(I_q(h^{(2)})^2 - 1)$, and let $g:\mathbb{R}\to\mathbb{R}$ be a polynomial. Then:*

$$\mathbb{E}\left[g(h^{(2)})I_q(A_j)I_q(A_k)\right] = \sum_{r=0}^{\deg(g)}\mathbb{E}\left[g^{(r)}(h^{(2)})\right]\frac{1}{r!}\sum_{i_1,\dots,i_r=1}^{d_1}\lambda_{i_1}\cdots\lambda_{i_r}\sum_{s_1,\dots,s_r=0}^{q-1}\sum_{t=0}^{q-1}c_{q,s,t}\sum_{\lambda}T_\lambda.$$

We now make two observations: The first is that v_1 is exactly what we computed for Theorem 16. The second one is that, so far, everything has been exact (i.e we have no error terms in our formula). Since we aim only to compute the order of the expectation, we will now proceed to bound each term, beginning with T_λ

E.4.3. BOUNDING T_λ

Recall that we defined:

$$T_\lambda = \langle(\dots((f_{i_{\lambda(1)}}\widetilde{\otimes}_{r_2}f_{i_{\lambda_\sigma(2)}})\widetilde{\otimes}_{r_3}f_{\lambda_\sigma(3)})\dots)\widetilde{\otimes}_{r_{|m|-1}}f_{\lambda_\sigma(|m|-1)};f_{\lambda_\sigma(|m|)}\rangle. \tag{E.67}$$

In particular, the term T_λ is computed from a particular set of appearances of $A_1,\dots A_{d_1}$. Let $A_{i_1},\dots,A_{i_r}$ denote the tensors involved in the computation of a particular T_λ. We claim the following:

Claim 20 *If $\{i_1, \dots i_r\} \cap \{j, k\} = \emptyset$, then with high probability with respect to $A_1, \dots A_{d_1}$*

$$|T_\lambda| \lesssim \frac{1}{d^q}.$$

Proof Note that symmetrization only change the expectation up to constants, so we can compute the terms without symmetrization. If $\{i_1, \dots i_r\} \cap \{j, k\} = \emptyset$, then the contraction in Equation (E.67) is linear in A_j and A_k, and therefore we can write (without symmetrization):

$$T_\lambda = \sum_{a,b \in [d]^q} w(a, b) A_{j,a} A_{k,b}, \tag{E.68}$$

where $w(a, b)$ sums over all the other contractions. Then, fixing $A_{i_1}, \dots, A_{i_r}$, we note that the expectation of $T_{\lambda((i_1,s_1),\dots,(i_{r+1},t))}$ w.r.t all A is zero. We can also compute the conditional variance to get:

$$\mathrm{Var}(T_\lambda | A_{i_1}, \dots, A_{i_r}) = \frac{1}{d^{2q}} \sum_{a,b \in [d]^q} w(a, b)^2 = \frac{1}{d^{2q}} \|w(a, b)\|^2. \tag{E.69}$$

Since $\mathbb{E}_A \left[\|w(a, b)\|^2\right] = O_d(1)$, we can proceed as with did in the linear case and conclude the lemma by the law of total variance. Having this, we get concentration by applying Chebyshev Inequality and Gaussian Hypercontractivity Theorem 4). ■

Having dealt with the disjoint case in Theorem 20, we now proceed with the harder case where indices $i_1, \dots i_r$ may have a non-empty intersection with $\{j, k\}$. Note that we don't want an sharp bound, but rather a bound that shows that this terms are negligible with respect to the linear part.

Let $(i_1, \dots, i_r$ be such that $\{i_1, \dots, i_r\} \cap \{j, k\} \neq$. Assume that the tensor A_j appears m_j times in the sequence $f_{i_1}, \dots, f_{i_{r+1}}$. Then, necessarily, m_j has to be odd: It appears once in $f_{i_{r+1}}$, and all other appearances will be tensor product of A_j with itself. By the same argument, m_k is also odd. Let $\mathcal{G}_{a_1,\dots,a_\ell}$ denote the σ-algebra generated by $A_{a_1}, \dots, A_{a_l \ell}$. Then from this observation we conclude:

$$\mathbb{E}_A[T_\lambda] = \mathbb{E}[\mathbb{E}[T_\lambda | \mathcal{G}_{[r]\setminus\{j,k\}}]] = 0. \tag{E.70}$$

By computing the conditional variance of T_λ, we will have:

$$\mathrm{Var}(T_\lambda | \mathcal{G}_{[r]\setminus\{j,k\}}) = \mathbb{E}\left[T_\lambda^2 | \mathcal{G}_{[r]\setminus\{j,k\}}\right]. \tag{E.71}$$

Let $F(A_j) = T_\lambda$. Then, by Theorem 5:

$$\mathrm{Var}(T_\lambda^2 | \mathcal{G}_{[r]\setminus\{j\}}) \leq \frac{C}{d^q} \mathbb{E}_{A_j} \left[\|\nabla_{A_j} T_\lambda\|^2 | \mathcal{G}_{[r]\setminus\{j\}}\right]. \tag{E.72}$$

Now, note that T_λ has the following form:

$$T_\lambda = \mathrm{Contraction}(A_{i_1}, \dots, A_{i_r}, A_j, A_k). \tag{E.73}$$

In particular, it is multi-linear in all of this arguments. Let D_H be a directional derivative in direction H. If we denote by $\mathrm{Pos}(j)$ the set of positions such that $A_{i_1} = A_j$, then:

$$D_H T_\lambda = \sum_{p \in \mathrm{Pos}(j)} \mathrm{Contraction}(A_{i_1}, \dots, A_{i_{p-1}}, H, A_{i_{p+1}}, \dots). \tag{E.74}$$

Then, by applying Cauchy-Schwarz:

$$|D_H T_\lambda| \leq C \sum_{p \in \mathrm{Pos}(j)} \|H\|_2 \prod_{\ell \neq p} \|K_\ell\|_2, \tag{E.75}$$

where K_ℓ are all other tensors. Since all these tensors are already contractions, we can apply Cauchy-Schwarz again to get:

$$|D_H T_\lambda| \leq C \sum_{p \in \mathrm{Pos}(j)} \|H\|_2 \|A_j\|^{m_j - 1} \prod_{\ell \neq j} \|A_\ell\|_2 \leq C m_j \|H\|_2 \|A_j\|^{m_j - 1} \prod_{\ell \neq j} \|A_\ell\|_2. \tag{E.76}$$

Taking supremum over the sphere, we conclude:

$$\|\nabla_{A_j} T_\lambda\| \leq C \|A_j\|^{m_j - 1} \prod_{\ell \neq j} \|A_\ell\|_2^{m_\ell}, \tag{E.77}$$

and taking the square:

$$\|\nabla_{A_j} T_\lambda\| \leq C \|A_j\|^{2(m_j - 1)} \prod_{\ell \neq j} \|A_\ell\|_2^{2m_\ell}. \tag{E.78}$$

Finally, by Hanson-Wright we know that $\|A_j\|_2$ is $\Theta_d(1)$ with high probability. Therefore:

$$\|\nabla_{A_j} T_\lambda\| \leq C \prod_{\ell \neq j} \|A_\ell\|_2^{2m_\ell}, \tag{E.79}$$

and in particular:

$$\mathrm{Var}(T_\lambda^2 | \mathcal{G}_{[r] \setminus \{j\}}) \lesssim \frac{1}{d^q} \prod_{\ell \neq j} \|A_\ell\|_2^{2m_\ell}. \tag{E.80}$$

By the Law of total variance, we can take expectation again to conclude:

$$\mathrm{Var}(T_\lambda) \leq \frac{1}{d^q}. \tag{E.81}$$

Then, by applying Chebyshev Inequality and hypercontractivity, we conclude that with high probability over $A_1, \ldots, A_{d_1}$

$$|T_\lambda| \lesssim \frac{1}{d^{\frac{q}{2}}}. \tag{E.82}$$

Let's write all of this in a Lemma.

Lemma 21 *Let T_λ be defined as in Equation (E.67), for a set of tensors $A_{i_1}, \ldots A_{i_r}$. Then:*

1. *If $\{i_1, \ldots, i_r\} \cap \{j, k\} = \emptyset$, then with high probability $|T_\lambda| \lesssim \frac{1}{d^q}$.*
2. *If $\{i_1, \ldots, i_r\} \cap \{j, k\} \neq \emptyset$, then with high probability $|T_\lambda| \lesssim \frac{1}{d^{\frac{q}{2}}}$.*

E.4.4. CONCLUSION

From Theorem 19, we had:

$$\mathbb{E}\left[g(h^{(2)})I_q(A_j)I_q(A_k)\right] = \sum_{r=0}^{\deg(g)} \mathbb{E}\left[g^{(r)}(h^{(2)})\right]\frac{1}{r!}\sum_{i_1,\dots,i_r=1}^{d_1}\lambda_{i_1}\cdots\lambda_{i_r}\sum_{s_1,\dots,s_r=0}^{q-1}\sum_{t=0}^{q-1}c_{q,s,t}\sum_{\lambda}T_\lambda. \tag{E.83}$$

We will separate the linear part from the rest. We write:

$$\mathbb{E}\left[g(h^{(2)})I_q(A_j)I_q(A_k)\right] = E_{\text{linear}} + \underbrace{\sum_{r\geq 2}^{\deg(g)} \mathbb{E}\left[g^{(r)}(h^{(2)})\right]\frac{1}{r!}\sum_{i_1,\dots,i_r=1}^{d_1}\lambda_{i_1}\cdots\lambda_{i_r}\sum_{s_1,\dots,s_r=0}^{q-1}\sum_{t=0}^{q-1}c_{q,s,t}\sum_{\lambda}T_\lambda}_{E_{\text{NL}}}. \tag{E.84}$$

Lets focus on E_{NL}. Replacing Theorem 21, since the sum on the RHS concerns $O_d(1)$ terms:

$$E_{\text{NL}} = O\left(\sum_{r=2}^{\deg(g)} \mathbb{E}\left[g^{(r)}(h^{(2)})\right]\frac{1}{d^{\frac{q}{2}}}\sum_{\substack{i_1,\dots,i_r=1\\ \{i_1,\dots,i_r\}\cap\{j,k\}\neq\emptyset}}^{d_1}\lambda_{i_1}\cdots\lambda_{i_r} + \sum_{r=2}^{\deg(g)} \mathbb{E}\left[g^{(r)}(h^{(2)})\right]\frac{1}{d^{q}}\sum_{\substack{i_1,\dots,i_r=1\\ \{i_1,\dots,i_r\}\cap\{j,k\}=\emptyset}}^{d_1}\lambda_{i_1}\cdots\lambda_{i_r}.\right) \tag{E.85}$$

Since the sum of λ_i is bounded (by the same argument as the linear case), we get:

$$E_{\text{NL}} = O\left(\sum_{r=2}^{\deg(g)} \mathbb{E}\left[g^{(r)}(h^{(2)})\right]\frac{1}{d^{\frac{q}{2}}}\sum_{\substack{i_1,\dots,i_r=1\\ \{i_1,\dots,i_r\}\cap\{j,k\}\neq\emptyset}}^{d_1}\lambda_{i_1}\cdots\lambda_{i_r}\right). \tag{E.86}$$

Now, from Theorem 16, we already computed v_1, so:

$$E_{\text{linear}} = \mathbb{E}\left[g(h^{(2)})\right]\langle A_j, A_k\rangle + \mathbb{E}\left[g'(h^{(2)})\right]O\left(\frac{\max(k,j)^{-\gamma}}{d^{\frac{q}{2}}}\right). \tag{E.87}$$

Analogously to the linear case, we can conclude:

$$E_{\text{linear}} = \mathbb{E}[g(h^{(2)})]\langle A_j, A_k\rangle + \mathbb{E}\left[g'(h^{(2)})\right]O\left(\frac{\max(k,j)^{-\gamma}}{d^{\frac{q}{2}}}\right). \tag{E.88}$$

From Equation (E.86) and he fact that all eigenvalues are in $[0,1]$, and $r\geq 2$, we get that E_{NL} is sub-leading with respect to the linear term. To finish, we need the following Lemma, whose proof we postpone to Section F.

Lemma 22 *Let $g:\mathbb{R}\to\mathbb{R}$ be a polynomial with information exponent 1. Then:*

$$\mathbb{E}[g(h^{(2)})] = \frac{1}{\sqrt{d}} \quad \textit{and}\ \mathbb{E}\left[g'(h^{(2)})\right] = \nu_1 + \frac{C}{\sqrt{d}}, \tag{E.89}$$

where ν_1 is the first Hermite coefficient of g.

Combining this with Theorem 16, we get that the first constant term is sub-leading and we conclude:

Lemma 23 *Let $h^{(2)} = \sum_{i=1}^{d_1} \lambda_i (I_q(A_i)^2 - 1)$, and let g be a polynomial with information exponent 1. Then, given $i, j \in [d_1]$ with $i \neq j$:*

$$\left| \mathbb{E}\left[g(h^{(2)}) I_q(A_j) I_q(A_k) \right] \right| = O\left(\frac{Z_\gamma}{d^q} (j^{-\gamma} + k^{-\gamma})) \right).$$

E.5. Studying $\mathbb{E}[\hat{C}]$

The objective of this section is to prove the following Lemma:

Lemma 24 *Under the assumptions of Theorem 1:*

$$\mathbb{E}\left[C^{(1)} \right] = \frac{\mathbb{E}\left[g'(h^{(2)}) \right]}{\sqrt{2}} A^{(1)} D_\gamma (A^{(1)})^T + \Delta,$$

where $\|\Delta\|_{\text{op}} = o_d(1)$, $A^{(1)} = [u_1, \dots u_{d_1}] \in \mathbb{R}^{D \times d_1}$, and $D_\gamma = \text{diag}(\lambda_1, \dots, \lambda_{d_1}) \in \mathbb{R}^{d_1 \times d_1}$.

The proof is similar to the one in Section E.4. If g is linear, the result is trivial, so we will focus on the non-linear setting, where we assume $0 \leq \gamma < \frac{1}{2}$. The linear setting will follow as a corollary.

Recall that

$$\hat{C} = \frac{1}{n} \sum_{\mu=1}^{n} y_\mu He_2(F_\mu), \tag{E.90}$$

with $F_\mu = \mathcal{F}(\text{He}_q(x_\mu))$, and $\text{He}_q(x_\mu)$ is the degree q Hermite tensor. Then the expectation is given by:

$$\mathbb{E}\left[\hat{C} \right] = \mathbb{E}\left[y_\mu He_2(F_\mu) \right]. \tag{E.91}$$

We want to prove that this expectation is close in operator norm to:

$$\mathbb{E}[g'(h^2)] \sum_{i=1}^{d_1} \lambda_i A_i^{(1)} (A_i^{(1)})^T. \tag{E.92}$$

Let $B(d, q$ be the dimension of the vectors $A_i^{(1)}$. Then:

$$\|\mathbb{E}\left[\hat{C} \right] - \frac{\mathbb{E}[g'(h^2)]}{\sqrt{2}} \sum_{i=1}^{d_1} \lambda_i A_i^{(1)} (A_i^{(1)})^T \|_{\text{op}} = \max_{B \in \mathbb{R}^{B(d,q)}, \|B\|_2 = 1} \left| B^T \mathbb{E}\left[\hat{C} \right] B - \frac{\mathbb{E}[g'(h^2)]}{\sqrt{2}} \sum_{i=1}^{d} \lambda_i \langle A_i, B \rangle^2 \right|. \tag{E.93}$$

Denote $Y_B := I_q(B)^2 - 1$. Then:

$$B^T \text{He}_2(F_\mu) B = \frac{1}{\sqrt{2}} (\langle B, F_\mu \rangle^2 - \underbrace{1}_{= \|B\|^2}) = \frac{1}{\sqrt{2}} (I_q(B)^2 - 1) = \frac{1}{\sqrt{2}} F_B. \tag{E.94}$$

Therefore:

$$B^T \mathbb{E}\left[\hat{C} \right] B = \frac{1}{\sqrt{2}} \mathbb{E}\left[g(h^{(2)}) Y_B \right]. \tag{E.95}$$

Then, going back to Equation (E.93):

$$\|\mathbb{E}\left[\hat{C}\right]-\mathbb{E}[g'(h^2)]\sum_{i=1}^{d_1}\lambda_i A_i^{(1)}(A_i^{(1)})^T\|_{\text{op}} = \frac{1}{\sqrt{2}}\max_{B\in\mathbb{R}^{B(d,q)},\|B\|_2=1}\left|\mathbb{E}\left[g(h^{(2)})Y_B\right]-\sum_{i=1}^{d}\lambda_i\langle A_i,B\rangle^2\right|. \tag{E.96}$$

Let

$$\Delta := \max_{B\in\mathbb{R}^{B(d,q)},\|B\|_2=1}\left|\mathbb{E}\left[g(h^{(2)})Y_B\right]-\sum_{i=1}^{d}\lambda_i\langle A_i,B\rangle^2\right|. \tag{E.97}$$

If we conclude that Δ is $o_d(1)$, we complete the proof. Our objective is hence to prove this. We will do this in a three steps.

In the following, we will extensively use the Wiener decomposition of Y_B and similar random variables, which is computed by Theorem 12. For any symmetric tensor $C\in(\mathbb{R}^d)^{\odot q}$:

$$(I_q(C)^2-1)=\sum_{r=0}^{q-1}c_{q,r}I_{2q-2r}(C\tilde{\otimes}_r C). \tag{E.98}$$

E.5.1. STEP 1: INTEGRATION BY PARTS

We will first try to write the term $\mathbb{E}\left[g(h^{(2)})Y_B\right]$ in Equation (E.97) in the form:

$$\mathbb{E}\left[g(h^{(2)})Y_B\right]=\sum_{i=1}^{d}\lambda_i\langle A_i,B\rangle^2+\text{other term}. \tag{E.99}$$

The tool for this is Theorem 14, Integration by Parts. We will apply it twice. On a first iteration, since Y_B is centered, we have:

$$\mathbb{E}\left[g(h^{(2)})Y_B\right]=\mathbb{E}\left[g'(h^{(2)})\langle Dh^{(2)},DL^{-1}Y_B\rangle\right]. \tag{E.100}$$

Now we apply integration by parts again and obtain:

$$\mathbb{E}\left[g(h^{(2)})Y_B\right]=\mathbb{E}\left[g'(h^{(2)})\right]\mathbb{E}\left[\langle Dh^{(2)},DL^{-1}Y_B\rangle\right] \tag{E.101}$$

$$+\mathbb{E}\left[g^{(2)}(h^{(2)})\langle Dh^{(2)},D\left(\langle Dh^{(2)},DL^{-1}Y_B\rangle\right)\rangle\right]. \tag{E.102}$$

Now, by the definition of $h^{(2)}$, and the linearity of the derivative:

$$\mathbb{E}\left[\langle Dh^{(2)},DL^{-1}Y_B\rangle\right]=\sum_{i=1}^{d}\lambda_i\mathbb{E}\left[\langle D(I_q(A_i)^2-1),DL^{-1}Y_B\rangle\right]. \tag{E.103}$$

Define $Y_i=(I_q(A_i)^2-1)$. Then, by doing inverse Gaussian integration by parts:

$$\mathbb{E}\left[\langle Dh^{(2)},DL^{-1}Y_B\rangle\right]=\sum_{i=1}^{d}\lambda_i\mathbb{E}\left[Y_iY_B\right], \tag{E.104}$$

and by applying Equation (E.98), and the orthogonality of Wiener Chaos Theorem 11:

$$\mathbb{E}\left[\langle Dh^{(2)}, DL^{-1}Y_B\rangle\right] = \sum_{i=1}^{d}\lambda_i\sum_{r=0}^{q-1}c_{q,r}\langle A_i\tilde{\otimes}_r A_i, B\tilde{\otimes}_r B\rangle \tag{E.105}$$

$$= \sum_{i=1}^{d}\lambda_i c_{q,1}\langle A_i\tilde{\otimes}A_i, B\tilde{\otimes}B\rangle + \sum_{i=1}^{d}\lambda_i\sum_{r=1}^{q-1}c_{q,r}\langle A_i\tilde{\otimes}_r A_i, B\tilde{\otimes}_r B\rangle. \tag{E.106}$$

By Theorem 30, we have that for all $r \in [q-1]$:

$$\mathbb{E}_{A_i}\left[\|A_i^{(1)}\otimes_r A_i^{(1)}\|_F^2\right] = O\left(\frac{1}{d^{-r}}\right). \tag{E.107}$$

Then with high probability:

$$\|A_i^{(1)}\otimes_r A_i^{(1)}\|_F^2 \lesssim \frac{1}{d^{-r}} \tag{E.108}$$

On the other hand, using the fact that $\lambda_i = z_i Z_\gamma i^{-\gamma}$, with $z_i \sim \mathrm{Rad}(\frac{1}{2})$, for fixed $A's$, we can apply Bernstein's Inequality [62] to get:

$$\left|\sum_{i=1}^{d}\lambda_i\sum_{r=1}^{q-1}c_{q,r}\langle A_i\tilde{\otimes}_r A_i, B\tilde{\otimes}_r B\rangle\right| \lesssim \sqrt{\sum_{i=1}^{d}Z_\gamma^2 i^{-2\gamma}\left(\sum_{r=1}^{q-1}c_{q,r}\langle A_i\tilde{\otimes}_r A_i, B\tilde{\otimes}_r B\rangle\right)^2} \tag{E.109}$$

$$\leq \sqrt{\sum_{i=1}^{d}Z_\gamma^2 i^{-2\gamma}\sum_{r=1}^{q-1}c_{q,r}\langle A_i\tilde{\otimes}_r A_i, B\tilde{\otimes}_r B\rangle^2} \tag{E.110}$$

$$\leq \sqrt{\sum_{i=1}^{d}Z_\gamma^2 i^{-2\gamma}\sum_{r=1}^{q-1}c_{q,r}\|A_i\tilde{\otimes}_r A_i\|_F^2\|B\tilde{\otimes}_r B\|_F^2}, \tag{E.111}$$

where in the last line we applied Cauchy Schwartz. Since $\|B\|_2^2$ by definition, we have $\|B\tilde{\otimes}_r B\|_F^2 \lesssim 1$. Then:

$$\left|\sum_{i=1}^{d}\lambda_i\sum_{r=1}^{q-1}c_{q,r}\langle A_i\tilde{\otimes}_r A_i, B\tilde{\otimes}_r B\rangle\right| \lesssim \sqrt{\sum_{i=1}^{d}Z_\gamma^2 i^{-2\gamma}\sum_{r=1}^{q-1}c_{q,r}\|A_i\tilde{\otimes}_r A_i\|_F^2}, \tag{E.112}$$

and replacing Equation (E.108), we conclude:

$$\left|\sum_{i=1}^{d}\lambda_i\sum_{r=1}^{q-1}c_{q,r}\langle A_i\tilde{\otimes}_r A_i, B\tilde{\otimes}_r B\rangle\right| \lesssim \frac{1}{\sqrt{d}}. \tag{E.113}$$

Replacing in Equation (E.106), we conclude:

$$\mathbb{E}\left[\langle Dh^{(2)}, DL^{-1}Y_B\rangle\right] = \sum_{i=1}^{d}\lambda_i c_{q,1}\langle A_i\tilde{\otimes}A_i, B\tilde{\otimes}B\rangle + \Delta_1, \tag{E.114}$$

with $|\Delta_1| \lesssim \frac{1}{d}$. Replacing in Equation (E.102):

$$\mathbb{E}\left[g(h^{(2)})Y_B\right] = \sum_{i=1}^{d} \lambda_i c_{q,1} \langle A_i \tilde{\otimes} A_i, B \tilde{\otimes} B\rangle + \Delta_1 \tag{E.115}$$

$$+ \mathbb{E}\left[g^{(2)}(h^{(2)})\langle Dh^{(2)}, D\left(\langle Dh^{(2)}, DL^{-1}Y_B\rangle\right)\rangle\right] + \Delta_1, \tag{E.116}$$

with $|\Delta_1| \lesssim \frac{1}{d}$. Then, by Equation (E.97), we conclude:

$$\Delta := \max_{B\in\mathbb{R}^{B(d,q)}, \|B\|_2=1} \left|\mathbb{E}\left[g^{(2)}(h^{(2)})\langle Dh^{(2)}, D\left(\langle Dh^{(2)}, DL^{-1}Y_B\rangle\right)\rangle\right]\right| + o_d(1). \tag{E.117}$$

We can now proceed to step 2.

E.5.2. STEP 2: COMPUTATION OF THE KERNELS

So far, we have reduced our problem to bounding

$$\left|\mathbb{E}\left[g^{(2)}(h^{(2)})\langle Dh^{(2)}, D\left(\langle Dh^{(2)}, DL^{-1}Y_B\rangle\right)\rangle\right]\right| \tag{E.118}$$

uniformly for $\|B\| = 1$. Since g is a polynomial, $g^{(2)}$ is also a polynomial. At the same time, $h^{(2)}$ has finite variance. Then we have:

$$\mathbb{E}\left[g^{(2)}(h^{(2)})^2\right]^{\frac{1}{2}} \lesssim 1. \tag{E.119}$$

Then, by Cauchy Schwartz:

$$\Delta \lesssim \max_{B\in\mathbb{R}^{B(d,q)}, \|B\|_2=1} \mathbb{E}\left[\left\langle Dh^{(2)}, D\left(\langle Dh^{(2)}, DL^{-1}Y_B\rangle\right)\right\rangle^2\right]^{\frac{1}{2}} + o_d(1) \tag{E.120}$$

$$\lesssim \max_{B\in\mathbb{R}^{B(d,q)}, \|B\|_2=1} \mathbb{E}\left[\left(\sum_{i,j=1}^{d_1} \lambda_i\lambda_j \left\langle DY_i, D\left(\langle DY_j, DL^{-1}Y_B\rangle\right)\right\rangle\right)^2\right]^{\frac{1}{2}} + o_d(1) \tag{E.121}$$

Let

$$V^{i_1,i_2} = \langle DY_{i_2}, D\left(\langle DY_{i_1}, DL^{-1}Y_B\rangle\right)\rangle \tag{E.122}$$

We will now compute the Chaos expansion of V. We begin with the nested derivative. Define $T_i^r = A_i \tilde{\otimes}_r A_i$, and $T_B^r = B\tilde{\otimes}_r B$. From Equation (E.60) and the derivative computation rule Theorem 13:

$$\langle DY_{i_1}, DL^{-1}Y_B\rangle = \sum_{r_1,r_2=0}^{q-1} c_{q,r_1,r_2} \langle DI_{2q-2r_1}(T_{i_1}^{r_2}), DI_{2q-2r_2}(T_B^{r_2})\rangle \tag{E.123}$$

$$= \sum_{r_1,r_2=0}^{q-1} c_{q,r_1,r_2} \sum_{r_3=1}^{2q-2\max(r_1,r_2)} I_{4q-2(r_1+r_2+r_3)}(T_{i_1}^{r_2} \tilde{\otimes}_{r_3} T_B^{r_1}). \tag{E.124}$$

Replacing in Equation (E.122):

$$V^{i_1,i_2} = \langle DY_{i_2}, DL^{-1}\left(\langle DY_{i_1}, DL^{-1}Y_B\rangle\right)\rangle \tag{E.125}$$

$$= \sum_{r_4=0}^{q-1} c_{q,r_4}\langle DI_{2q-2r_4}(T_{i_2}^{r_4}), D\left(\sum_{r_1,r_2=0}^{q-1} c_{q,r_1,r_2} \sum_{r_3=1}^{2q-2\max(r_1,r_2)} I_{4q-2(r_1+r_2+r_3)}(T_{i_1}^{r_2}\tilde{\otimes}_{r_3}T_B^{r_1})\right) \tag{E.126}$$

$$= \sum_{r_1,r_2,r_4=0}^{q-1} \sum_{r_3=1}^{2q-2\max(r_1,r_2)} c_{q,r}\langle DI_{2q-2r_4}(T_{i_2}^{r_4}), DI_{4q-2(r_1+r_2+r_3)}(T_{i_1}^{r_2}\tilde{\otimes}_{r_3}T_B^{r_1})\rangle \tag{E.127}$$

$$= \sum_{r_1,r_2,r_4=0}^{q-1} \sum_{r_3=1}^{2q-2\max(r_1,r_2)} \sum_{r_5=1}^{6q-2\max(r_4,(r_1+r_2+r_4)} c_{q,r} I_{6q-2(\sum_{\ell=1}^5 r_i)}\left(T_{i_2}^{r_4}\tilde{\otimes}_{r_5}\left(T_{i_1}^{r_2}\tilde{\otimes}_{r_3}T_B^{r_1}\right)\right). \tag{E.128}$$

We can now go to Step 3.

E.5.3. STEP 3: CONTRACTION BOUNDS

Ignoring the sum for the moment. The tensors involved in this computation are of the form:

$$T_{i_2}^{r_4}\tilde{\otimes}_{r_5}\left(T_{i_1}^{r_2}\tilde{\otimes}_{r_3}T_B^{r_1}\right) \tag{E.129}$$

Now, there are three possible cases:

1. **Case 1:** The contraction r_5 contains a contraction of size greater than 1 between $T_{i_2}^{r_4}$ and $T_{i_1}^{r_1}$.
2. **Case 2**: The contraction r_5 only contracts $T_{i_2}^{r_4}$ and $T_B^{r_1}$, but $\max(r_4, r_1) \geq 1$.
3. **Case 2**: The contraction r_5 only contracts $T_{i_2}^{r_4}$ and $T_B^{r_1}$, but $r_4 = r_1 = 0$.

We write:

$$V^{i_1,i_1} = V_1^{i_1,i_1} + V_2^{i_1,i_1} + V_3^{i_1,i_1}, \tag{E.130}$$

where $V_1^{i_1,i_1}$ counts only the indices in **Case 1**, $V_2^{i_1,i_1}$ counts only the indices in **Case 2**, and $V_3^{i_1,i_1}$ counts only the indices in **Case 3**. We now study each term separately.

Case 1: Assume there is a contraction between $T_{i_2}^{r_4}$ and $T_{i_1}^{r_1}$ of t_{i_1,i_2} indices. Then, by Cauchy-Schwarz:

$$\|T_{i_2}^{r_4}\tilde{\otimes}_{r_5}\left(T_{i_1}^{r_2}\tilde{\otimes}_{r_3}T_B^{r_1}\right)\|_F \leq \|T_{i_2}^{r_4}\tilde{\otimes}_{t_{i_1,i_2}}T_{i_1}^{r_2}\|\|T_B^{r_1}\|_F. \tag{E.131}$$

Since $\|B\|^2 = 1$, we can bound $\|T_B^{r_1}\|_F$ by a constant and obtain:

$$\|T_{i_2}^{r_4}\tilde{\otimes}_{r_5}\left(T_{i_1}^{r_2}\tilde{\otimes}_{r_3}T_B^{r_1}\right)\|_F \leq C\|T_{i_2}^{r_4}\tilde{\otimes}_{t_{i_1,i_2}}T_{i_1}^{r_2}\|_F. \tag{E.132}$$

By Theorem 36 and Theorem 35, we have:

$$\mathbb{E}\left[\|T_{i_2}^{r_4}\tilde{\otimes}_{t_{i_1,i_2}}T_{i_1}^{r_2}\|^2\right] = O(\frac{1}{d}). \tag{E.133}$$

Then, since $\|T_{i_2}^{r_4}\tilde{\otimes}_{t_{i_1,i_2}}T_{i_1}^{r_2}\|^2$ is a polynomial, we can apply Chebyshev Inequality to obtain that with high probability:

$$\|T_{i_2}^{r_4}\tilde{\otimes}_{t_{i_1,i_2}}T_{i_1}^{r_2}\| \lesssim \frac{1}{\sqrt{d}}. \tag{E.134}$$

Then:

$$\mathbb{E}[(V_1^{i_1,i_1})^2]^{\frac{1}{2}} \lesssim \frac{1}{\sqrt{d}}. \tag{E.135}$$

Case 2: If the contraction r_5 only contracts $T_{i_2}^{r_4}$ and $T_B^{r_1}$, but $\max(r_4, r_1) \geq 1$. Then, applying Equation (E.131) we get:

$$\|T_{i_2}^{r_4}\tilde{\otimes}_{r_5}\left(T_{i_1}^{r_2}\tilde{\otimes}_{r_3}T_B^{r_1}\right)\|_F \leq \|T_{i_2}^{r_4}\|_F\|T_{i_1}^{r_2}\|_F\|T_B^{r_1}\|_F, \tag{E.136}$$

By Lemma A.3 from [58], we have:

$$\mathbb{E}\left[\|T_i^r\|^2\right] = O\left(\frac{1}{d^{-r}}\right), \tag{E.137}$$

for $r \in [q-1]$. Then, by applying hypercontractivity and Chebyshev inequality, with high probability:

$$\|T_i^r\|_F \lesssim d^{-\frac{r}{2}}, \tag{E.138}$$

with high probability, so from the fact that $\max(r_2, r_4) \geq 1$, we conclude that with high probability:

$$\|T_{i_2}^{r_4}\tilde{\otimes}_{r_5}\left(T_{i_1}^{r_2}\tilde{\otimes}_{r_3}T_B^{r_1}\right)\|_F \lesssim \frac{1}{\sqrt{d}}. \tag{E.139}$$

Then:

$$\mathbb{E}[(V_2^{i_1,i_1})^2]^{\frac{1}{2}} \lesssim \frac{1}{\sqrt{d}}. \tag{E.140}$$

Note that we can write Equation (E.121) as:

$$\Delta \lesssim \max_{B\in\mathbb{R}^{B(d,q)},\|B\|_2=1} \mathbb{E}\left[\left(\sum_{i,j=1}^{d_1}\lambda_i\lambda_j V^{i_1,i_2}\right)^2\right]^{\frac{1}{2}} + o_d(1) \tag{E.141}$$

$$\lesssim \max_{B\in\mathbb{R}^{B(d,q)},\|B\|_2=1} \mathbb{E}\left[\left(\sum_{i,j=1}^{d_1}\lambda_i\lambda_j (V_1^{i,j}+V_2^{i,j}+V_3^{i,j})\right)^2\right]^{\frac{1}{2}} + o_d(1). \tag{E.142}$$

Replacing Equation (E.140) and Equation (E.135) in Equation (E.142), and using the fact that for $\gamma < \frac{1}{2}$,

$$\left(\sum_{i,j=1}^{d_1}\lambda_i^2\lambda_j^2\right) = \Theta(1), \tag{E.143}$$

we get:

$$\Delta \lesssim \max_{B\in\mathbb{R}^{B(d,q)},\|B\|_2=1} \mathbb{E}\left[\left(\sum_{i,j=1}^{d_1}\lambda_i\lambda_j V_3^{i,j})\right)^2\right]^{\frac{1}{2}} + o_d(1). \tag{E.144}$$

We can now proceed with the last step.

E.5.4. STEP 4: STUDYING THE LAST TERM

In **Case 3**, the contraction r_5 only contracts $T_{i_2}^{r_4}$ and $T_B^{r_1}$, and moreover $r_4 = r_1 = 0$, so we cannot apply the results for controlling the norms of tensors. Note that if $r = 0$, we have:

$$T_i^r = A_i^{(1)} \tilde{\otimes} A_i^{(1)}. \tag{E.145}$$

Then;

$$V_3^{i_1,i_2} = \sum_{r_1}^{q-1} \sum_{r_3=1}^{2q-2r_1} \sum_{r_5=1}^{6q-2\max(r_4,r_1)} c_{q,r} I_{6q-2(\sum_{\ell=1}^5 r_i)} \left(T_{i_2}^0 \tilde{\otimes}_{r_5}^B \left(T_{i_1}^0 \tilde{\otimes}_{r_3} T_B^{r_1} \right) \right), \tag{E.146}$$

where we used the upper index $\otimes^B$ to denotes that the r_5 contractions are only between elements of $T_{i_2}^0$ and $T_B^{r_1}$. Then:

$$\sum_{i,j=1}^{d_1} \lambda_i \lambda_j V_3^{i,j} = \sum_{i,j=1}^{d_1} \lambda_i \lambda_j \sum_{r_1}^{q-1} \sum_{r_3=1}^{2q-2r_1} \sum_{r_5=1}^{6q-2\max(r_4,r_1)} c_{q,r} I_{6q-2(\sum_{\ell=1}^5 r_i)} \left(T_i^0 \tilde{\otimes}_{r_5}^B \left(T_j^0 \tilde{\otimes}_{r_3} T_B^{r_1} \right) \right) \tag{E.147}$$

$$= \sum_{r_1}^{q-1} \sum_{r_3=1}^{2q-2r_1} \sum_{r_5=1}^{6q-2\max(r_4,r_1)} c_{q,r} I_{6q-2(\sum_{\ell=1}^5 r_i)} \left(\sum_{i,j=1}^{d_1} \lambda_i \lambda_j T_i^0 \tilde{\otimes}_{r_5}^B \left(T_j^0 \tilde{\otimes}_{r_3} T_B^{r_1} \right) \right). \tag{E.148}$$

Then, going to Equation (E.144), we can apply Cauchy Schwartz to get:

$$\Delta \lesssim \max_{B \in \mathbb{R}^{B(d,q)}, \|B\|_2 = 1} \mathbb{E} \left[\left(\sum_{r_1}^{q-1} \sum_{r_3=1}^{2q-2r_1} \sum_{r_5=1}^{6q-2\max(r_4,r_1)} c_{q,r} I_{6q-2(\sum_{\ell=1}^5 r_i)} \left(\sum_{i,j=1}^{d_1} \lambda_i \lambda_j T_i^0 \tilde{\otimes}_{r_5}^B \left(T_j^0 \tilde{\otimes}_{r_3} T_B^{r_1} \right) \right) \right)^2 \right]^{\frac{1}{2}} + o_d(1 \tag{E.149}$$

$$\lesssim \max_{B \in \mathbb{R}^{B(d,q)}, \|B\|_2 = 1} \sum_{r_1}^{q-1} \sum_{r_3=1}^{2q-2r_1} \sum_{r_5=1}^{6q-2\max(r_4,r_1)} c_{q,r} \mathbb{E} \left[I_{6q-2(\sum_{\ell=1}^5 r_i)} \left(\sum_{i,j=1}^{d_1} \lambda_i \lambda_j T_i^0 \tilde{\otimes}_{r_5}^B \left(T_j^0 \tilde{\otimes}_{r_3} T_B^{r_1} \right) \right)^2 \right]^{\frac{1}{2}} + o_d(1) \tag{E.150}$$

$$\lesssim \max_{B \in \mathbb{R}^{B(d,q)}, \|B\|_2 = 1} \sum_{r_1}^{q-1} \sum_{r_3=1}^{2q-2r_1} \sum_{r_5=1}^{6q-2\max(r_4,r_1)} c_{q,r} \mathbb{E} \left[I_{6q-2(\sum_{\ell=1}^5 r_i)} \left(\sum_{i,j=1}^{d_1} \lambda_i \lambda_j T_i^0 \tilde{\otimes}_{r_5}^B \left(T_j^0 \tilde{\otimes}_{r_3} T_B^{r_1} \right) \right)^2 \right]^{\frac{1}{2}} + o_d(1) \tag{E.151}$$

For a fixed **Case 3** pattern of contractions α, define

$$K_{B,\alpha} = \sum_{i,j=1}^{d_1} \lambda_i \lambda_j \, T_i^0 \tilde{\otimes}_{r_5}^{B,\alpha} \left(T_j^0 \tilde{\otimes}_{r_3}^{\alpha} T_B^{r_1} \right). \tag{E.152}$$

Where we recall that the upper index B, α means that the r_5 contractions are only between T_i^0 and the $T_B^{r_1}$. Then, computing the expectation

$$\mathbb{E}\left[I_\alpha\left(\sum_{i,j=1}^{d_1}\lambda_i\lambda_j T_i^0 \tilde{\otimes}_{r_5}^{B}\left(T_j^0\tilde{\otimes}_{r_3}T_B^{r_1}\right)\right)^2\right]^{\frac{1}{2}} \leq C\|K_{B,\alpha}\|_F \tag{E.153}$$

Recall that $T_i^0 = A_i^{(1)}\otimes A_i^{(1)}$. Then, each kernel $K_{B,\alpha}$ has the form:

$$K_{B,\alpha} = \sum_{i,j=1}^{d_1}\lambda_i\lambda_j\,(A_i^{(1)}\otimes A_i^{(1)})\tilde{\otimes}_{r_5}^{B,\alpha}\left((A_i^{(1)}\otimes A_i^{(1)})^0\tilde{\otimes}_{r_3}^{\alpha}T_B^{r_1}\right) = (\sum_{i=1}^{d_1}\lambda_i A_i^{(1)}\otimes A_i^{(1)})\tilde{\otimes}_{r_5}^{B,\alpha}\left(\sum_{i=1}^{d_1}\lambda_i(A_i^{(1)}\otimes A_i^{(1)})^0\tilde{\otimes}_{r_3}^{\alpha}T \right. \tag{E.154}$$

where in the last equality we used bi-linearity of the conteractions. From here, for each pattern α, since the different components of the first and second tensor don't interact, there exists deterministic flattening maps such that, considering $K_{B,\alpha}$ as an operator in a Hilbert Space:

$$K_{B,\alpha} = \left(\sum_{i=1}^{d_1}\lambda_i\mathcal{F}_1\left[A_i^{(1)}\otimes A_i^{(1)}\right]\right)\circ\left(\sum_{i}^{d_1}\lambda_i\mathcal{F}_2\left[A_i^{(1)}\otimes A_i^{(1)}\right]\right)\circ\mathcal{F}_3[T^r B] \tag{E.155}$$

Then:

$$\|K_{B,\alpha}\|_F \lesssim \left\|\sum_{i=1}^{d_1}\lambda_i\mathcal{F}_1\left[A_i^{(1)}\otimes A_i^{(1)}\right]\right\|_{\text{op}}\left\|\sum_{i=1}^{d_1}\lambda_i\mathcal{F}_2\left[A_i^{(1)}\otimes A_i^{(1)}\right]\right\|_{\text{op}}\|\mathcal{F}_4[T^r B]\|_F, \tag{E.156}$$

and since $\|B\| = 1$, we have:

$$\|K_{B,\alpha}\|_F \lesssim \left\|\sum_{i=1}^{d_1}\lambda_i\mathcal{F}_1\left[A_i^{(1)}\otimes A_i^{(1)}\right]\right\|_{\text{op}}\left\|\sum_{i=1}^{d_1}\lambda_i\mathcal{F}_2\left[A_i^{(1)}\otimes A_i^{(1)}\right]\right\|_{\text{op}}. \tag{E.157}$$

Since the maps $\mathcal{F}_1, \mathcal{F}_2$ are linear and depend on a finite number of re-arrangements inside the tensors, we have:

$$\|K_{B,\alpha}\|_F \lesssim \left\|\sum_{i=1}^{d_1}\lambda_i A_i^{(1)}\otimes A_i^{(1)}\right\|_{\text{op}}\left\|\sum_{i=1}^{d_1}\lambda_i A_i^{(1)}\otimes A_i^{(1)}\right\|_{\text{op}}. \tag{E.158}$$

Applying a standard matrix concentration bound [61], Corollary 5.35, with high probability:

$$\left\|\sum_{i=1}^{d_1}A_i^{(1)}\otimes A_i^{(1)}\right\|_{\text{op}} \lesssim \frac{1}{\sqrt{d^q}}(\sqrt{d^q}+\sqrt{d_1}). \tag{E.159}$$

Then applying triangular inequality for the matrix $\sum_{i=1}^{d_1}\lambda_i A_i^{(1)}\otimes A_i^{(1)}$, with high probability:

$$\left\|\sum_{i=1}^{d_1}\lambda_i A_i^{(1)}\otimes A_i^{(1)}\right\|_{\text{op}} \lesssim \max|\lambda_i|, \tag{E.160}$$

and since $\gamma < \frac{1}{2}$, $\max|\lambda_i| = |\lambda_1| = Z_\gamma = \Theta_d(d^{\frac{1-2\alpha}{2}})$. Replacing this bound in Equation (E.158), with high probability:

$$\|K_{B,\alpha}\|_F \lesssim \frac{1}{d^{1-2\gamma}}, \tag{E.161}$$

for all patterns α in **Case 3**. Then, replacing in Equation (E.151), we get:

$$\|\Delta\|_{\text{op}} \lesssim \frac{1}{d^{1-2\gamma}}, \tag{E.162}$$

with high probability. Therefore, we conclude that:

$$\|\mathbb{E}\left[\hat{C}\right] - \mathbb{E}[g'(h^2)]\sum_{i=1}^{d_1}\lambda_i A_i^{(1)}(A_i^{(1)})^T\|_{\text{op}} \lesssim \Delta \lesssim \frac{1}{d^{1-2\gamma}}, \tag{E.163}$$

which concludes the Proof.

E.5.5. PROOF OF THEOREM 6

Putting Theorem 24 and Theorem 22 together, we can conclude:

Corollary 25 *Under the assumptions of Theorem 1:*

$$\mathbb{E}\left[C^{(1)}\right] = \frac{\nu_1}{\sqrt{2}}A^{(1)}D_\gamma(A^{(1)})^T + \Delta,$$

where $\|\Delta\|_{\text{op}} = o_d(1)$, $A^{(1)} = [u_1, \dots u_{d_1}] \in \mathbb{R}^{D\times d_1}$, $D_\gamma = \text{diag}(\lambda_1, \dots, \lambda_{d_1}) \in \mathbb{R}^{d_1\times d_1}$, *and* ν_1 *is the first Hermite coefficient of* g.

Appendix F. Deferred Proofs

Lemma 26 *Let* $k, j \in [d_1]$, *with* $k \neq j$. *Assume* $n = \omega_d(\frac{\min(k,j)^{2\gamma}}{Z_\gamma^2}d^q)$. *Then with high probability*

$$\left(\frac{Z_\gamma}{n}y_\mu h_{\mu,k}h_{\mu,j}\right)^2 \lesssim \frac{Z_\gamma^2}{d^q}\min(k,j)^{-2\gamma}.$$

Proof We want to concentrate

$$E^2 = \left(\frac{Z_\gamma}{n}y_\mu h_{\mu,k}h_{\mu,j}\right)^2. \tag{F.1}$$

For this, we begin by decomposing:

$$E^2 = \left(\frac{1}{n}\sum_{\mu=1}^{n}\left(y_\mu h_{\mu,k}h_{\mu,j} - \mathbb{E}\left[y_\mu h_{\mu,k}h_{\mu,j}\right]\right) + \mathbb{E}\left[y_\mu h_{\mu,k}h_{\mu,k}\right]\right)^2 \tag{F.2}$$

$$\lesssim \underbrace{\left(\frac{1}{n}\sum_{\mu=1}^{n}\left(y_\mu h_{\mu,k}h_{\mu,j} - \mathbb{E}\left[y_\mu h_{\mu,k}h_{\mu,j}\right]\right)\right)^2}_{(I)} + \underbrace{\mathbb{E}\left[y_\mu h_{\mu,k}h_{\mu,k}\right]^2}_{(II)}. \tag{F.3}$$

We begin by concentrating (I). For this, we note that since all terms are centered, independent random variables:

$$\mathbb{E}\left[(I)\right] = \frac{1}{n^2}\sum_{\mu=1}^{n}\mathbb{E}\left[\left(y_\mu h_{\mu,k}h_{\mu,j} - \mathbb{E}\left[y_\mu h_{\mu,k}h_{\mu,j}\right]\right)^2\right]. \tag{F.4}$$

Note that, since $\mathrm{Var}(y_\mu) \lesssim 1$, we have:

$$\mathbb{E}\left[\left(y_\mu h_{\mu,k}h_{\mu,j} - \mathbb{E}\left[y_\mu h_{\mu,k}h_{\mu,j}\right]\right)^2\right] \lesssim \mathbb{E}\left[\left(y_\mu h_{\mu,k}h_{\mu,j}\right)^2\right] \lesssim 1. \tag{F.5}$$

Then

$$\mathbb{E}\left[(I)\right] \lesssim \frac{1}{n}, \tag{F.6}$$

and since (I) is positive, by Markov inequality we conclude that with high probability:

$$\left(\frac{Z_\gamma}{n}y_\mu h_{\mu,k}h_{\mu,j}\right)^2 \lesssim \frac{1}{n} + \mathbb{E}\left[y_\mu h_{\mu,k}h_{\mu,k}\right]^2. \tag{F.7}$$

We now turn to $(II) = \mathbb{E}\left[y_\mu h_{\mu,k}h_{\mu,k}\right]^2$. Recall that, by definition:

$$y_\mu = g\left(\sum_{p=1}^{d_1}\lambda_p((h_{\mu,p}^{(1)})^2 - 1)\right), \tag{F.8}$$

for some polynomial $g : \mathbb{R} \to \mathbb{R}$, with $\mathbb{E}[g'(h^{(2)})] \neq 0$. Then, by Theorem 23

$$\left|\mathbb{E}\left[y_\mu h_{\mu,k}h_{\mu,k}\right]\right| = O\left(\frac{Z_\gamma}{d^{\frac{q}{2}}}\min(k,j)^{-\gamma}\right). \tag{F.9}$$

Taking the square:

$$\mathbb{E}\left[y_\mu h_{\mu,k}h_{\mu,k}\right]^2 \lesssim \frac{Z_\gamma^2}{d^q}\min(k,j)^{-2\gamma}. \tag{F.10}$$

By putting together Equation (F.7) and Equation (F.10):

$$\left(\frac{Z_\gamma}{n}y_\mu h_{\mu,k}h_{\mu,j}\right)^2 \lesssim \frac{1}{n} + \frac{Z_\gamma^2}{d^q}\min(k,j)^{-2\gamma}. \tag{F.11}$$

and since $n = \omega_d(\frac{\min(k,j)^{2\gamma}}{Z_\gamma^2}d^q)$, we conclude that with high probability:

$$\left(\frac{Z_\gamma}{n}y_\mu h_{\mu,k}h_{\mu,j}\right)^2 \lesssim \frac{Z_\gamma^2}{d^q}\min(k,j)^{-2\gamma}. \tag{F.12}$$

■

F.1. Gaussian Universality

In the following, let $\mathcal{W}_1$ denote the 1-Wasserstein distance on $\mathcal{P}_1(\mathbb{R}^r)$. That is, for $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^r)$, $\mathcal{W}_1$ defines the metric:

$$\mathcal{W}_1(\mu,\nu) = \sup_{h:\mathbb{R}^r\to\mathbb{R},\mathrm{Lip}(h)\leq 1} \left|\mathbb{E}_{G\sim\mu}[h(G)] - \mathbb{E}_{Z\sim\nu}[h(Z)]\right|. \tag{F.13}$$

We will need the following result bounding the Wasserstein distance to Gaussians.

Lemma 27 (Theorem 5.1.3 in [49]) *Let $F \in \mathbb{D}^{1,2}$ with $E[F] = 0$ and $E[F^2] = \sigma^2 > 0$, and let $N \sim \mathcal{N}(0,\sigma^2)$. Then*

$$d_{\mathrm{W}}(F,N) \leq \frac{\sqrt{2}}{\sigma\sqrt{\pi}} E\left[\left|\sigma^2 - \langle DF, -DL^{-1}F\rangle_{\mathfrak{F}}\right|\right].$$

Lemma 28 *Let $r \in \mathbb{N}$, and let $A_1^{(1)},\dots,A_r^{(1)} \in (\mathbb{R}^d)^{\otimes q}$ be symmetric tensors of order q such as the ones specified in Section 3. Let $x \sim \mathcal{N}(0, I_d)$, and let $H_k(x)$ denote the degree k Hermite tensor of x. Define $h^{(2)}$ as in Section 3. Then, then there exists a constant $C_k < \infty$, depending only on k, such that:*

$$\mathcal{W}_1\left(h^{(2)}, N\right) \leq C_k \frac{1}{\sqrt{d}}, \tag{F.14}$$

where $N \sim \mathcal{N}(\mathbb{E}[h^{(2)}], \mathrm{Var}(\mathrm{h}^{(2)}))$.

Proof The proof is similar to Lemma A.1 in [58], but instead of using the multi-variate Gaussian approximation Lemma, we use the one-dimensional version Theorem 27. We begin by computing the Wiener expansion of $h^{(2)}$. We have:

$$h^{(2)} = \sum_{i=1}^{d_1} \lambda_i \left(I_q(A_i^{(1)})^2 - 1\right) = \sum_{i=1}^{d_1} \lambda_i (I_q(A_i)^2 - \|A_i\|^2) + \sum_{i=1}^{d} \lambda_i(\|A_i\|^2 - 1). \tag{F.15}$$

By the product formula Theorem 12, and the fact that the first term is centered:

$$h^{(2)} = \sum_{i=1}^{d_1} \lambda_i \sum_{r=0}^{q-1} c_{q,r} I_{2q-2r}(A_i \tilde{\otimes}_r A_i) + \sum_{i=1}^{d} \lambda_i(\|A_i\|^2 - 1) \tag{F.16}$$

$$= \sum_{r=0}^{q-1} I_{2q-2r}(c_{q,r} \sum_{i=1}^{d_1} \lambda_i A_i \tilde{\otimes}_r A_i) + \sum_{i=1}^{d} \lambda_i(\|A_i\|^2 - 1). \tag{F.17}$$

From here, we have:

$$\mathbb{E}\left[h^{(2)}\right] = \sum_{i=1}^{d} \lambda_i(\|A_i\|^2 - 1), \text{ and} \mathbb{E}\left[(h^{(2)})^2\right] = \sum_{r=0}^{q-1} c_{q,r}^2 \| \sum_{i=1}^{d_1} \lambda_i A_i \tilde{\otimes}_r A_i\|_F^2. \tag{F.18}$$

Now, we want to apply Theorem 27. By centering, we have to control $\langle Dh^{\bar{(2)}}, DL^{-1}h^{\bar{(2)}}\rangle$, for $h^{\bar{(2)}} = h^{(2)} - \mathbb{E}[h^{(2)}]$. Denote $B_r = \sum_{i=1}^{d_1} c_{q,r}\lambda_i A_i \tilde{\otimes}_r A_i$. By the linearity of the derivative:

$$\langle Dh^{\bar{(2)}}, DL^{-1}h^{\bar{(2)}}\rangle = \sum_{r,r'=0}^{q-1} \langle DI_{2q-2r}(B_r), DI_{2q-2r'}(B_{r'})\rangle, \tag{F.19}$$

and by applying Theorem 13:

$$\langle Dh^{\bar{(2)}}, DL^{-1}h^{\bar{(2)}}\rangle = \sum_{r,r'=0}^{q-1}\sum_{p=1}^{2q-2\max(r_1,r_2)} c_{q,r,r',p} I_{4q-2r-2r'-2p}(B_r \tilde{\otimes}_p B_{r'}). \tag{F.20}$$

Now, let

$$K_s = \sum_{r,r'=0}^{q-1}\sum_{p=1}^{2q-2\max(r_1,r_2)} c_{q,r,r',p}\mathbf{1}_{4q-2r-2r'-2p=s} B_r \tilde{\otimes}_p B_{r'}. \tag{F.21}$$

Then:

$$\langle Dh^{\bar{(2)}}, DL^{-1}h^{\bar{(2)}}\rangle = \sum_{s\geq 0} I_s(K_s). \tag{F.22}$$

Then, we have:

$$\mathbb{E}\left[\left(\mathrm{Var}(h^{(2)}) - \langle Dh^{(2)}, DL^{-1}h^{(2)}\rangle\right)^2\right] = \sum_{s\geq 1}\|K_s\|_F^2. \tag{F.23}$$

Then to conclude, we need to show that this norms are negligible for large d. Let $s \geq 0$. We have:

$$\|K_s\|_F^2 = \|\sum_{r,r'=0}^{q-1}\sum_{p=1}^{2q-2\max(r_1,r_2)} c_{q,r,r',p}\mathbf{1}_{4q-2r-2r'-2p=s}\|B_r \tilde{\otimes}_p B_{r'}\|_F^2 \tag{F.24}$$

$$\leq C\sum_{r,r'=0}^{q-1}\sum_{p=1}^{2q-2\max(r_1,r_2)} \mathbf{1}_{4q-2r-2r'-2p=s}\left\|B_r \tilde{\otimes}_p B_{r'}\right\|_F^2. \tag{F.25}$$

Recall $B_r = \sum_{i=1}^{d_1} c_{q,r}\lambda_i A_i \tilde{\otimes}_r A_i$. Denote $T_i^r = A_i \tilde{\otimes} r A_i$. Then:

$$B_r \tilde{\otimes}_p B_{r'} = \sum_{i,j}\lambda_i\lambda_j T_i^r \tilde{\otimes}_p T_j^r, \tag{F.26}$$

and:

$$\|K_s\|_F^2 \leq C\sum_{r,r'=0}^{q-1}\sum_{p=1}^{2q-2\max(r_1,r_2)} \mathbf{1}_{4q-2r-2r'-2p=s}\sum_{i,j}|\lambda_i|^2|\lambda_j|^2\|T_i^r \tilde{\otimes}_p T_j^r\|_F^2, \tag{F.27}$$

so everything reduces to estimating the norms $\|T_i^r \tilde{\otimes}_p T_j^r\|_F$.

Ignoring symmetrization, which only changes things up to constants, given $a, b \in [d]^{q-r}$, we can apply Theorem 35 for $i = j$ and Theorem 36 for $i \neq j$. From which we will obtain:

$$\mathbb{E}\left[\|T_i^r \tilde{\otimes}_p T_j^r\|_F^2\right] = O\left(\frac{1}{d}\right) \forall i, j \in [d_1] \tag{F.28}$$

Since $\|T_i^r \tilde{\otimes}_p T_j^r\|_F^2$ is a polynomial of Gaussians, we can use Theorem 4 to control its moments, and applying Chebyshev we will get:

$$\|T_i^r \tilde{\otimes}_p T_j^r\|_F^2 \lesssim \frac{1}{\sqrt{d}}, \tag{F.29}$$

with high probability over $A_1^{(1)}, \dots, A_{d_1}^{(1)}$. Replacing in Equation (F.27):

$$\|K_s\|_F^2 \lesssim \frac{1}{\sqrt{d}} C \sum_{r,r'=0}^{q-1} \sum_{p=1}^{2q-2\max(r_1,r_2)} \mathbf{1}_{4q-2r-2r'-2p=s} \sum_{i,j} |\lambda_i|^2 |\lambda_j|^2. \tag{F.30}$$

Note that, since $\gamma < \frac{1}{2}$

$$\sum_{i,j} |\lambda_i|^2 |\lambda_j|^2 = Z_\gamma^4 \left(\sum_{i=1}^{d_1} i^{-2\alpha} \right)^2 = \Theta_d \left(\frac{d^{2(1-2\alpha)}}{d^{4\frac{(1-2\alpha)}{2}}} \right) = \Theta_d(1). \tag{F.31}$$

From this, we can conclude:

$$\|K_s\|_F^2 \lesssim \frac{1}{\sqrt{d}} \forall s, \tag{F.32}$$

with high probability. Finally, this allows us to conclude:

$$\mathbb{E}\left[\left(\mathrm{Var}(h^{(2)}) - \langle Dh^{(2)}, DL^{-1}h^{(2)} \rangle \right)^2 \right] = O\left(\frac{1}{\sqrt{d}} \right), \tag{F.33}$$

with high probability over $A_1^{(1)}, \dots, A_{d_1}^{(1)}$. ■

With this, conclude the following Corollary, stated as Theorem 22 in Section E.

Corollary 29 *Let $g : \mathbb{R} \to \mathbb{R}$ be a polynomial with information exponent 1. Assume $\gamma < \frac{1}{2}$. Then:*

$$\mathbb{E}[g(h^{(2)})] = \frac{1}{\sqrt{d}} \quad \textit{and } \mathbb{E}\left[g'(h^{(2)}) \right] = \nu_1 + \frac{C}{\sqrt{d}}, \tag{F.34}$$

where ν_1 is the first Hermite coefficient of g.

Proof [Sketch of the Proof] The idea of the proof is to approximate g be a Lipschitz function which has bounded support and apply Theorem 28. ■

Lemma 30 (Lemma A.3 in [58]) *Let A, B be independent tensors in $(\mathbb{R}^d)^{\otimes k}$ with i.i.d. entries $\mathcal{N}(0, d^{-k})$. Then for each $s \in \{1, \dots, k\}$,*

$$\mathbb{E}\|A \otimes_s B\|_F^2 = \Theta(d^{-s}).$$

While for the self-contractions, $E\|A \otimes_s A\|_F^2 = \Theta(d^{-s})$ for $s \in \{1, \dots, k-1\}$ and $E\|A \otimes_k A\|_F^2 = 1$.

Appendix G. Technical Lemmas

G.1. Cumulants

Definition 31 (Definition 8.2.1 in [49]) *Let $F = (F_1, \dots, F_N)$ be an $\mathbb{R}^N$-valued random vector with $F_i \in \mathbb{D}^{1,2}$ for each i. Let $l_1, l_2, \dots$ be a sequence taking values in the multi-index set $\{e_1, \dots, e_N\}$. We set $\Gamma_{l_1}(F) = F^{l_1} = F_j$, where j is such that $l_1 = e_j$. If the random variable $\Gamma_{l_1,\dots,l_k}(F)$ is a well-defined element of $L^2(\Omega)$ for some $k \geq 1$, we set*

$$\Gamma_{l_1,\dots,l_{k+1}}(F) = \langle DF^{l_{k+1}}, -DL^{-1}\Gamma_{l_1,\dots,l_k}(F)\rangle_{\mathfrak{H}}.$$

Lemma 32 (Theorem 8.2.5 in [49]) *Let $m = (m_1, \dots, m_d) \in \mathbb{N}^d \setminus \{0\}$ be a multi-index. Write $m = l_1 + \ldots + l_{|m|}$ where the multi-indices $l_i \in \{e_1, \dots, e_d\}$, $i = 1, \dots, |m|$, are unique in the sense of Lemma 8.1.1. Suppose that the random vector $F = (F_1, \dots, F_d)$ is such that $F_i \in \mathbb{D}^{|m|,2^{|m|}}$ for each i . Then*

$$\kappa_m(F) = \sum_{\sigma \in \mathfrak{S}_{\{2,\dots,|m|\}}} \mathbb{E}\left[\Gamma_{l_1, l_{\sigma(2)},\dots,l_{\sigma(|m|)}}(F)\right].$$

Lemma 33 (Theorem 8.3.1 in [49]) *Let $m \in \mathbb{N}^d \setminus \{0\}$ be a multi-index such that $|m| \geq 3$. Write $m = l_1 + \ldots + l_{|m|}$ with $l_i \in \{e_1, \dots, e_d\}$ for each i (see Lemma 8.1.1). Consider an $\mathbb{R}^d$ -valued random vector of the form*

$$F = (F_1, \dots, F_d) = (I_{q_1}(f_1), \dots, I_{q_d}(f_d)),$$

where each f_i belongs to $\mathfrak{H}^{\odot q_i}$. When $l_k = e_j$, we set $\lambda_k = j$, so that $F^{l_k} = F_{\lambda_k}$ for all $k = 1, \dots, |m|$. Then

$$\kappa_m(F) = \sum_{\sigma \in \mathfrak{S}_{\{2,\dots,|m|\}}} (q_{\lambda_\sigma(|m|)})! \sum_{*} c_{q,l,\sigma}(r_2, \dots, r_{|m|-1})$$

$$\times \langle (\dots((f_{\lambda_1} \widetilde{\otimes}_{r_2} f_{\lambda_\sigma(2)}) \widetilde{\otimes}_{r_3} f_{\lambda_\sigma(3)}) \dots) \widetilde{\otimes}_{r_{|m|-1}} f_{\lambda_\sigma(|m|-1)}; f_{\lambda_\sigma(|m|)} \rangle_{\mathfrak{H}^{\otimes q_{\lambda_\sigma(|m|)}}},$$

where the second sum $\sum_*$ runs over all collections of integers $r_2, \dots, r_{|m|-1}$ such that:
(i) $1 \leq r_i \leq q_{\lambda_\sigma(i)}$ *for all* $i = 2, \dots, |m| - 1$;
(ii) $r_2 + \ldots + r_{|m|-1} = \frac{q_{\lambda_1} + q_{\lambda_\sigma(2)} + \ldots + q_{\lambda_\sigma(|m|-1)} - q_{\lambda_\sigma(|m|)}}{2}$;
(iii) $r_2 < \frac{q_{\lambda_1} + q_{\lambda_\sigma(2)}}{2}, \dots, r_2 + \ldots + r_{|m|-2} < \frac{q_{\lambda_1} + q_{\lambda_\sigma(2)} + \ldots + q_{\lambda_\sigma(|m|-2)}}{2}$;
(iv) $r_2 \leq q_{\lambda_1}, r_3 \leq q_{\lambda_1} + q_{\lambda_\sigma(2)} - 2r_2, \dots, r_{|m|-1} \leq q_{\lambda_1} + q_{\lambda_\sigma(2)} + \ldots + q_{\lambda_\sigma(|m|-2)} - 2r_2 - \ldots - 2r_{|m|-2}$;

G.2. Gaussian Tensors

Lemma 34 (Theorem 3.1.1 in [25]) *For natural numbers $n \geq m$, let $\{X_i\}_{i=1}^n$ be n independent random variables with values in a measurable space (S, δ), and let $\{X_i^k\}_{i=1}^n$, $k = 1, \ldots, m$, be m independent copies of this sequence. Let B be a separable Banach space and, for each $(i_1, \ldots, i_m) \in I_n^m$, let $h_{i_1 \ldots i_m} : S^m \mapsto B$ be measurable functions such that $\mathbb{E}\big(\|h_{i_1 \ldots i_m}(X_{i_1}, \ldots, X_{i_m})\|\big) < \infty$. Let $\Phi : [0, \infty) \to [0, \infty)$ be a convex non-decreasing function such that $\mathbb{E}\Phi\big(\|h_{i_1 \ldots i_m}(X_{i_1}, \ldots, X_{i_m})\|\big) < \infty$ for all $(i_1, \ldots, i_m) \in I_n^m$. Then,*

$$\mathbb{E}\Phi\Big(\Big\|\sum_{I_n^m} h_{i_1 \ldots i_m}(X_{i_1}, \ldots, X_{i_m})\Big\|\Big) \leq \mathbb{E}\Phi\Big(C_m\Big\|\sum_{I_n^m} h_{i_1 \ldots i_m}(X_{i_1}^1, \ldots, X_{i_m}^m)\Big\|\Big).$$

Lemma 35 *Let $A \in (\mathbb{R}^d)^{\odot q}$ be a symmetric tensor with independent gaussian centered entries with variance d^{-q}. Let $r, r' \in [q-1] \cup \{0\}$, and let $p \in [2q - 2\max(r_1, r_2)]$. Then:*

$$\mathbb{E}\left[\|(A \otimes_{r_1} A) \otimes_{r_3} (A \otimes_{r_2} A)\|^2\right] = O\left(\frac{1}{d}\right).$$

Proof Let $M = (r_1 + r_2 + r_3)$, and $N = 4q - 2(r_1 + r_2 + r_3) = 4q - 2M$. Let $F : (\mathbb{R}^d)^{\otimes q} \to (\mathbb{R}^d)^{\otimes N}$ be defines by:

$$F(A)_x = \sum_{z \in \mathbb{R}^N} \prod_{\ell=1}^4 A_{I_\ell(x,z)}, \tag{G.1}$$

where each $I_\ell(x, z) \in [d]^q$ is a q-tuple. Given a pair (x, z), let $\pi(x, z)$ define a partition of set $[4]$ by

$$a \sim_{\pi(x,z)} b \iff I_a(x, z) = I_b(x, z). \tag{G.2}$$

Denote by $\hat{\pi}$ the minimal partition, that is $\hat{\pi} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$. In this partition, we have $I_1(x, z) = I_2(x, z) = I_3(x, z) = I_4(x, z)$. Let

$$F^{\text{distinct}}(A) = \sum_{z \in \mathbb{R}^N} \mathbf{1}_{\pi(x,z) = \hat{\pi}} \prod_{\ell=1}^4 A_{I_\ell(x,z)}, \tag{G.3}$$

and let

$$F^{\text{equal}}(A) = F(A) - F^{\text{distinct}}(A) = \sum_{z \in [d]^N} \mathbf{1}_{\pi(x,z) \neq \hat{\pi}} \prod_{\ell=1}^4 A_{I_\ell(x,z)}. \tag{G.4}$$

Let $\Phi(x) = \|x\|^2$. Then we have:

$$\phi(F(A)) \leq \Phi(F^{\text{equal}}(A)) + \phi(F^{\text{distinct}}(A)) \tag{G.5}$$

Since Φ is convex, we can apply Theorem 34 to obtain:

$$\mathbb{E}\left[\phi(F^{\text{distinct}}(A))\right] \leq C\mathbb{E}\left[\phi(F^{\text{distinct}}(A^1, A^2, A^3, A^4))\right], \tag{G.6}$$

where $A^\ell, \ell \in [4]$ are independent copies of A. Then:

$$\mathbb{E}\left[\phi(F^{\text{distinct}}(A,B,C,D))\right] = \sum_{x\in[d]^M}\sum_{z,z'\in[d]^N}\prod_{a=1}^{4}\mathbb{E}\left[A^{(a)}_{I_a(x,z)}A^{(a)}_{I_a(x,z')}\right] \tag{G.7}$$

$$=\lesssim \frac{1}{d^{4q}}d^{M+N} = \frac{1}{d^{r_1+r_2+r_3}}. \tag{G.8}$$

Since $r_3 > 0$, this term is at most $O(\frac{1}{d})$. We now move to F^{equal}. Given $\ell, \ell' \in [4]$, define the sets:

$$J_{ab} = \{(x,z) : I_\ell(x,z) = I_{\ell'}(x,z)\}, \tag{G.9}$$

and

$$J_{\text{equal}} = \{(x,z) : \pi(x,z) \neq \hat{\pi}\}. \tag{G.10}$$

Then, by definition of $\hat{\pi}$:

$$J_{\text{equal}} \subseteq \bigcup_{a<b} J_{ab}. \tag{G.11}$$

We claim that $|J_{a,b}| \leq Cd^{M+N-1}$. To see this, note that a pair (x,z) has $M+N$ degrees of freedom (M from $x \in [d]^M$, and N from $z \in [d]^N$). Moreover, denote by $m_{a,b}$ the number of shared coordinates between $I_a(x,z)$ and $I_b(x,z)$. Then, we have $M+N-m_{a,b}$ degrees of freedom. Let's bound this quantity.

If we look at the function F as $F(A^1,A^2,A^3,A^4)$, then we have that $m_{1,2} = r_1$ and $m_{3,4} = r_2$. The final r_3 contraction further identifies coordinates as a partition:

$$r_3 = s_{13} + s_{14} + s_{23} + s_{24}, \tag{G.12}$$

where s_{ab} denotes the number of coordinates that are shared after the contraction r_3. Since $r_3 > 1$, we have that one of this terms has to be at least 1.

Note that, since initially the different blocks of the r_3 contraction were not sharing coordinates, we have that $m_{13} = s_{1,3}$, $m_{1,4} = s_{1,4}$, $m_{2,3} = s_{2,3}$ and $m_{2,4} = s_{2,4}$. Since $r_1, r_2 \in [q-1] \cup \{0\}$, we have that $\max(m_{12}, m_{23}) \leq q-1$. Moreover, a cross contraction $s_{a,b}$ cannot be more than the indices that were already contracted in a or b, so

$$s_{a,b} \leq q - \min(r_1, r_2) \leq q-1. \tag{G.13}$$

Then we have that $m_{a,b} \leq q-1$ for all $a,b \in [4]$, and consequently, $q - m_{a,b} \geq 1$, and therefore, we cannot have more that $N+M-1$ degrees of freedom, which by definition is at most $4q-1$ degrees of freedom. As a consequence, $|J_{a,b}| \leq Cd^{M+N-1}$.

With this, we can finally bound $\|F^{\text{equal}}\|^2$. We have:

$$\mathbb{E}\left[\|F^{\text{equal}}\|^2\right] = \sum_{x\in[d]^M}\sum_{z,z\in[d]^N}\mathbf{1}_{\pi(x,z)\neq\hat{\pi}}\mathbb{E}\left[\prod_{\ell=1}^{4}A_{I_\ell(x,z)}\prod_{\ell=1}^{4}A_{I_\ell(x,z')}\right] \tag{G.14}$$

$$\leq \frac{1}{d^{4q}}\left|J_{\text{equal}}\right| \tag{G.15}$$

$$\leq \frac{1}{d^{4q}}\sum_{a<b,a,b\in[4]}|J_{a,b}|, \tag{G.16}$$

but as we just saw, $|J_{a,b}| \leq d^{M+N-1}$, so we conclude:

$$EE\left[\|F^{\text{equal}}\|^2\right] \leq \frac{C}{d}, \tag{G.17}$$

and we conclude the proof. ■

Lemma 36 *Let $A, B \in (\mathbb{R}^d)^{\odot q}$ be two distinct symmetric tensors with independent gaussian centered entries with variance d^{-q}. Let $r, r' \in [q-1] \cup \{0\}$, and let $p \in [2q - 2\max(r_1, r_2)]$. Then:*

$$\mathbb{E}\left[\|(A \otimes_{r_1} A) \otimes_{r_3} (A \otimes_{r_2} A)\|^2\right] = O\left(\frac{1}{d}\right).$$

Proof The proof is analogous to one of Theorem 35. ■