Low Rank Gradients and Where to Find Them

Rishi Sonthalia RISHI.SONTHALIA@BC.EDU
Boston College

Michael Murray MJM253@BATH.AC.UK
Bath University

Guido Montúfar GUIDO.MONTUFAR@STAT.UCLA.EDU
University of California, Los Angeles

Abstract

This paper investigates low-rank structure in the gradients of the training loss for two-layer neural networks while relaxing the usual isotropy assumptions on the training data and parameters. We consider a spiked data model in which the bulk can be anisotropic and ill-conditioned, we do not require independent data and weight matrices and we also analyze both the mean-field and neural-tangent-kernel scalings. We show that the gradient with respect to the input weights is approximately low rank and is dominated by two rank-one terms: one aligned with the bulk data–residue, and another aligned with the rank one spike in the input data. We characterize how properties of the training data, the scaling regime and the activation function govern the balance between these two components. Additionally, we also demonstrate that standard regularizers, such as weight decay, input noise and Jacobian penalties, also selectively modulate these components. Experiments on synthetic and real data corroborate our theoretical predictions.

1. Introduction

Feature learning is a critical driver behind the success of deep learning. Despite this, a theoretical characterization of it remains elusive. In order to drive understanding, a line of research [2, 9–11, 20, 30] has emerged studying two-layer networks whose inner weights are trained or updated via one step of gradient descent. In this context, feature learning can be characterized through the emergence of a low-rank structure in the network weights. Ba et al. [2] proved that a ridge estimator trained on such features can outperform random feature models and other kernel methods. However, these prior investigations require idealized conditions, for example isotropic data or weights, which diverge from real-world scenarios where data typically exhibits anisotropy or an ill-conditioned covariance. In addition, the effects of regularization in this context have also been underexplored.

This paper addresses two questions: *1) how do low-rank gradient phenomena arise and behave under more general conditions of anisotropy and ill-conditioning?* and *2) what impact do common regularizers have on feature learning in this context?* Our analysis accommodates spiked data with an anisotropic ill-conditioned bulk. This allows us to explore the effect of the size of the data spike, controlled by a parameter $\nu \geq 0$, as well as spectral decay profiles of the bulk, controlled by a parameter $\alpha \geq 0$. Our central finding is that the gradient of the inner-layer weights is generically well approximated by a **rank-two matrix**. This structure arises from the interplay of two primary rank-one components: S_1, driven by the input bulk and target residue, and S_2, driven by the leading eigenvector of the data covariance. The relative prominence of these components, and consequently the direction of feature learning, is determined by the interplay of data properties, the

© R. Sonthalia, M. Murray & G. Montúfar.

scale of the network parametrization, the choice of loss and activation function as well as the use of regularization. We corroborate our theoretical findings with experiments on both synthetic data (Section 3 and appendix D) and real data (MNIST, CIFAR-10 embeddings).[1] A summary of our key contributions is:

- **Generalized Theory of Low-Rank Gradients:** We provide a theoretical framework (Section 3, Theorems 1 and 3) characterizing the low-rank structure of the gradient under significantly relaxed assumptions on data and weight matrices (anisotropy, ill-conditioning; Section 2).
- **Identification of a Dominant Rank-Two Structure:** We show (Theorems 1 and 3) that the gradient is often better approximated by a rank-two matrix than the rank-one structures identified in prior specialized settings. We provide conditions under which each of these components dominates.
- **Modulation by Activation Function and Regularization:** We show how activation functions and common regularizers selectively modulate the components of the gradient. We reveal that ReLU can suppress the contribution from the residue S_1 (Section 3), while input noise and a Jacobian penalty can promote the residue component and data spike component (Appendix D) respectively.
- **Mean Field (MF) versus Neural Tangent Kernel (NTK) scaling:** We demonstrate differences in dominant spike alignments, $S_1 \sim X_B^T y$ in MF vs. $S_1 \sim X_B^T r$ in NTK, at initialization (Section 3) and the subsequent impact during training.

2. Setup and Assumptions

In this section, we provide the technical details required for analysis. A summary of notation and discussion of examples of when the assumptions hold can be found in Table 1 and Appendix A. We consider shallow networks with d input dimensions, m hidden neurons, and n training data points.

Assumption 1 (Proportional scaling) *Let $\psi_1, \psi_2 \in \mathbb{R}_{>0}$ be fixed constants. We consider m, n as functions of d such that $n/d \to \psi_1 < 1$ and $m/d \to \psi_2$ as $d \to \infty$.*

Data: We consider random input data $x_i \in \mathbb{R}^d$ for $i \in [n]$, sampled i.i.d. These are stored row-wise in a matrix $X \in \mathbb{R}^{n \times d}$. For each x_i, the corresponding label is $y_i \in \mathbb{R}$, and labels stored as $y \in \mathbb{R}^n$.

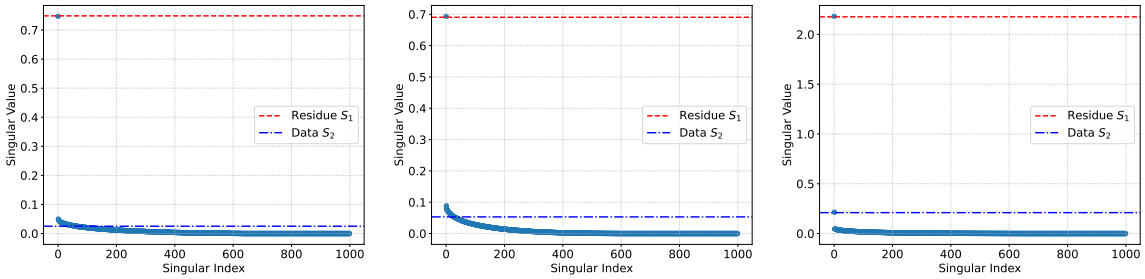
Assumption 2 (Input features distribution) *Let $\hat{\Sigma} \in \mathbb{R}^{d \times d}$ for which there exists an $\alpha \geq 0$ such that the k-th eigenvalue satisfies $\lambda_k(\hat{\Sigma}) = k^{-\alpha}$ for $k = 1, \ldots, d$. Let $q \in \mathbb{S}^{d-1}$ and define $\zeta = n^{\nu}$ for some $\nu \geq 0$. We assume each input data point x_i is sampled i.i.d. from a multivariate Gaussian distribution $N(0, \Sigma)$, where the full covariance $\Sigma \in \mathbb{R}^{d \times d}$ is given by $\Sigma = \hat{\Sigma} + \zeta^2 q q^T$.*

Network: We consider a two-layer neural network with input-output map $f : \mathbb{R}^d \to \mathbb{R}$ defined as $f(x) = \gamma_m a^T \sigma(Wx) \in \mathbb{R}$. The parameter $\gamma_m \in \mathbb{R}_{>0}$ is a non-trainable scaling constant that depends on the network width m.

Assumption 3 (Network parameters) *We assume the following for W, a and γ_m. **Outer weights:** a_j are sampled i.i.d. from $\mathrm{Uniform}(\{-1, 1\})$. **Inner weights**: Rows w_j of W have unit length, $w_j \in \mathbb{S}^{d-1}$. **Scaling parameter:** $\gamma_m = \Theta(1/\sqrt{m})$ (NTK scaling) or $\gamma_m = \Theta(1/m)$ (MF scaling).*

Assumption 4 (Activation function) *The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies the following. **Smoothness:** σ' and σ'', first and second derivatives of σ, exist almost everywhere on $\mathbb{R}$. **Lipschitzness:** σ and σ' are L-Lipschitz for some constant $L > 0$. **Non-trivial expected derivative:** For $x \sim N(0, \Sigma)$ and W , let $\mu_j = \mathbb{E}_x[\sigma'(w_j^T x)]$. We assume $\mu_j = \Omega(1)$ for all j. Let $\mu = [\mu_1, \ldots, \mu_m]^T$. We define $\sigma'_\perp(Wx)_j = \sigma'(w_j^T x) - \mu_j$.*

1. All code is available at the anonymous Github repository: Link



(a) Tanh, BCE, $\nu = \frac{1}{8}$, isotropic W. **(b)** Swish, Hinge, $\nu = \frac{3}{8}$, non-isotropic W. **(c)** Softplus, BCE, $\nu = \frac{3}{8}$, non-isotropic W.

Figure 1: Singular value distribution of the gradient G for varying activation, loss and ν and weight distribution. Red, and blue lines show the singular value of S_1, and S_2 respectively.

Parameter update via GD: Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a function which measures the loss between a label and a prediction. We define the loss given a dataset $(X, y) = (x_i, y_i)_{i \in [n]}$ with respect to the inner-layer weights W as $L(W) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$. We consider an update to W arising from one step of GD with step size $\eta > 0$. We define the *residue*:

$$r = [\partial \ell(f(x_1), y_1)/\partial f(x_1), \ldots, \partial \ell(f(x_n), y_n)/\partial f(x_n)]^T \in \mathbb{R}^n. \tag{1}$$

To motivate this terminology, consider that for the Mean Squared Error (MSE) loss, r corresponds to the vector of residues $[f(x_i) - y_i]_i$. More generally, for many losses r can typically be interpreted as the component of the targets not captured by the predictions of the model (see Appendix A.3). For our results to hold we require the following technical assumption on the residues.

Assumption 5 (Residue concentration) *Under the proportional scaling regime (Assumption 1), with probability* $1 - o(1)$ *over the training data* (X, y)*, the residue* r *satisfies* $\frac{\|r\|_\infty}{\|r\|_2} = O\left(\frac{\log n}{\sqrt{n}}\right)$.

We emphasize that Assumption 5 is a mild condition: it ensures that no single component of the residue vector disproportionately dominates its overall ℓ_2 norm. Our analysis also depends on the alignment between the residue r and specific structural components of the input data X. From Assumption 2 we have the following decomposition of the input features,

$$X = X_B + X_S = X_B + \zeta z q^T \in \mathbb{R}^{n \times d}, \tag{2}$$

where X_B has rows sampled i.i.d. from $\mathcal{N}(0, \hat{\Sigma})$, $z \sim \mathcal{N}(0, I)$, and q is a unit vector. For sufficiently large ζ, z is approximately the principal eigenvector of XX^T. We will consider the degree of alignment between the residue vector r and the spike component z of the input data. The projection of the residue r onto the principal eigenvector of XX^T is a natural statistic of interest and has been considered in prior works [14, 25]. In Appendix A.5 we provide β estimates for 192 scenarios.

Assumption 6 (Residue alignment) *With probability* $1 - o(1)$, $\left|\frac{1}{\sqrt{n}\|r\|_2} z^T r\right| = \Theta(d^{-\beta/2})$.

3. Spiked Data Leads to a Low-Rank Gradient

We demonstrate that for a spiked data covariance the gradient G is either approximately rank one or rank two, depending primarily on the size of the spike. To demonstrate this we define the

following three rank-one matrices, where $\Xi = \left[z^\top\big((ra^\top)\circ\sigma'_\perp(XW^\top)\big)\right]$:

$$\underbrace{S_1 := \frac{\gamma_m}{n}\left(X_B^\top r\right)(a\circ\mu)^\top}_{\text{Residue Spike}}, \quad \underbrace{S_2 := \frac{\gamma_m\zeta}{n}\, q\Xi}_{\text{Data Spike}}, \quad \underbrace{S_{12} := \gamma_m\zeta\,\frac{z^\top r}{n}\, q\,(a\circ\mu)^\top}_{\text{Interpolant}}.$$

The key contribution of this section is Theorem 1, which characterizes the approximate low rank structure of the gradient for small-to-moderate spike sizes. For large data spikes ($\nu \geq 0.5$), we note that the $\mathcal{C}^2$ smoothness of the activation function and independence between W and X are no longer required and is discussed in Appendix C.

We also note that Theorem 1 generalizes [2, Proposition 2] by covering a broader range of covariance structures, loss functions and initialization scalings[2]. In the small spike setting, $\nu \in [0, 1/4)$, the gradient is approximately rank one and aligns with the residue plus interpolant $S_1 + S_{12}$. By contrast, in the moderate spike setting, $\nu \in [1/4, 1/2)$, the gradient becomes rank two. We empirically verify these our theoretical results (Figures 1) across a range of activation and loss functions under the NTK scaling.

Theorem 1 (Gradient approximation) *Suppose Assumptions 1, 2, 3, 4, 5, 6 are satisfied, X and W are independent, and σ is a $\mathcal{C}^2$ function. Define $E = G - S_1 - S_{12} - S_2$. Then, for all $\nu, \alpha \in \mathbb{R}_{\geq 0}$,*

$$\frac{\|G - S_1 - S_{12}\|_2}{\sqrt{m}\gamma_m\|r\|_\infty} = O\left(\|W\|_2 n^{2\nu-\frac{1}{2}}\right), \quad \frac{\|G - S_1 - S_{12} - S_2\|_2}{\sqrt{m}\gamma_m\|r\|_\infty} = O\left(\|W\|_2 n^{\nu-\frac{1}{2}}\right) \tag{3}$$

with probability $1 - o(1)$ as $d, n, m \to \infty$. Moreover, if $\nu < \frac{1}{2}$ then with the same probability

$$\frac{\|S_1\|_2}{\|E\|_2} = \Omega\left(\frac{n^{\frac{1}{2}-\nu-\frac{\alpha}{2}}}{\log n\|W\|_2}\right), \qquad \frac{\|S_2\|_2}{\|E\|_2} = \Omega\left(\frac{n^\nu}{\log n}\frac{\|(z\circ r)^T\sigma'_\perp(XW^T)\|_2}{\|\sigma'_\perp(XW^T)\|_2}\right), \tag{4}$$

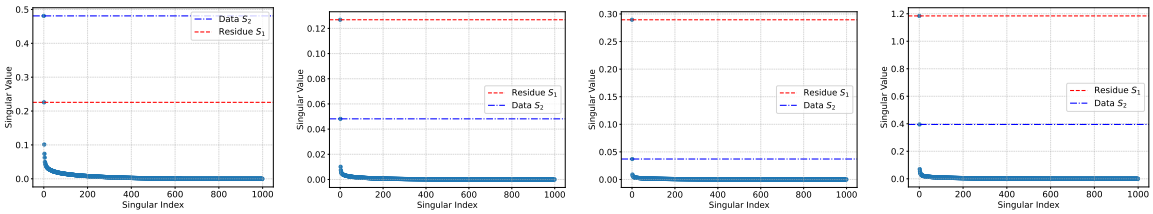
$$\frac{\|S_{12}\|_2}{\|E\|_2} = \Omega\left(\frac{n^{\frac{1}{2}-\frac{\beta}{2}}}{\log n\|W\|_2}\right), \qquad \Omega(n^{\nu-\frac{\beta}{2}}) \leq \frac{\|S_{12}\|_2}{\|S_1\|_2} \leq O(n^{\nu-\frac{\beta}{2}+\frac{\alpha}{2}}). \tag{5}$$

Observe that for $\nu < 1/4$, if $\|W\|_2 \log n = o(n^{\frac{1}{2}-\nu-\frac{\alpha}{2}})$ then G is approximately equal to the rank-one matrix $S_1 + S_{12}$. Further, if $\beta > 2\nu + \alpha$ then the gradient is dominated by S_1 and the spike is aligned with the data-residue term $X_B^T r$. However, if $\beta < 2\nu$ then the gradient term is dominated by S_{12}, which is aligned with the data spike q. In addition, for $\nu \in [1/4, 1/2)$, if $\|W\|_2 \log n = o(n^{\frac{1}{2}-\nu-\frac{\alpha}{2}})$ and $n^\nu = \omega\left(\log n \frac{\|\sigma'_\perp(XW^T)\|_2}{\|(z\circ r)^T\sigma'_\perp(XW^T)\|_2}\right)$, then the gradient is approximately the rank-two matrix $S_1 + S_{12} + S_2$. Note this is distinct from prior works [2, 3, 11, 30] where the gradient is only ever approximately rank one.

Theorem 1 requires the activation to be $\mathcal{C}^2$. As detailed in the proof, this is needed to establish that $\|\sigma'_\perp(XW^T)\|_2 \leq O(\|W\|n^{\nu+\frac{1}{2}})$. Indeed, when $\|W\|_2 = \Theta(1)$ and $\nu < \frac{1}{2}$ we have $\|\sigma'_\perp(XW^T)\|_2 = o(n)$, which is key for S_1 to separate from the bulk spectrum. However, ReLU is not $\mathcal{C}^2$. To understand, the effect of using ReLU we provide Proposition 2.

Proposition 2 (ReLU gradient) *If $2\nu > 1 - \alpha$, and the row of W are i.i.d. from the unit sphere, then with probability $1 - o(1)$ we have that $\sigma'_\perp(XW^T) = \frac{1}{2}\operatorname{sign}(z_i)\operatorname{sign}(Wq)^T$.*

2. [2, Proposition 2] requires $\nu = \alpha = 0$, isotropic data and MSE loss with MF scaling.



(a) ReLU **(b)** Swish **(c)** ELU **(d)** Softplus

Figure 2: ReLU suppresses the residue spike (S_1) compared to smooth activations. Fixed parameters: $\nu = 1/8$, $\alpha = 5/9$, $n = 750$, $d = 1000$, and $m = 1250$.

From Proposition 2 we see that for ReLU the operator norm of $\sigma'_\perp(XW^T)$ is $\Theta(n)$. This is a significant increase compared to the $o(n)$ scaling for $\mathcal{C}^2$ activations and suggests that the norm of E and S_2 are larger for ReLU. The increased size of E, S_2 results in the relative suppression of the contribution of S_1 and an enhancement of the contribution of S_2 to the spectrum of the gradient. We empirically verify this phenomenon in Figure 2 where we compare ReLU to its $\mathcal{C}^2$ activations ELU, Swish, and Softplus. We see that, for ReLU the relative residue contribution (S_1) is significantly smaller when compared with its smooth approximations.

Impact of the Scale Parameter: MF vs. NTK Scaling We consider the implications of our results for the two scaling regimes and highlight three important distinctions. As with prior work, we consider the large step-size regime. Specifically, we use a step size of γ_m^{-1}. To avoid exploding gradients deploy *Weight Normalization* (WN) [24]. We limit our focus to the MSE loss. See Appendix E for a discussion of which assumptions hold during training and a deeper discussion. In particular, we show that the spike in the gradient for the NTK and MF scaling are qualitatively different. In particular, in the MF the scaling the residue aligns with the targets y and for NTK aligns with the residual r. See Figure 7. Next the spike direction for MF is stable during training, while the spike direction for NTK is not stable (Figure 8).

Effect of Regularization We analyze three regularization techniques: ℓ_2 weight decay, isotropic input noise, and Jacobian penalization. We show that ℓ_2 can suppress both spikes. On the other hand, isotropic gaussian noise only suppresses the data spike while *promoting* the residue spike. Contrastingly, the Jacobian penalty only suppress the residue spike and *promotes* the data spike. The complete theoretical discussion can be found in Appendix D along with validation on real data.

4. Conclusion

This work shows that in two-layer neural networks, the hidden-layer gradient is approximately rank-two, driven by data-residual (S_1) and data-spike (S_2) components connected by an interpolant (S_{12}). We show that activation function choice, scaling, and regularization can result in qualitatively different gradients. In particular, we have the following rule of thumb for the number of spikes.

Gradient-spike rule-of-thumb: Which spike dominates at initialization?

$\boxed{S_1 \text{ (residue spike)}} \leftrightarrow 2\nu < \min\{\frac{1}{2}, \beta - \alpha, 1 - \alpha\}$ or Large isotropic input noise

$\boxed{S_{12} + S_2 + S_3 \text{ (data spike)}} \leftrightarrow \begin{cases} (i)\ 2\nu > \min\{1, \beta\}, \text{ or } (ii) \text{ Strong Jacobian penalty,} \\ \text{or } (iii) \text{ ReLU and } 2\nu > 1 - \alpha \end{cases}$

If none of the above holds, both spikes remain, and the gradient is typically rank-two.

The coexistence and interplay of the two spike components offer a nuanced understanding of the gradient. We believe that the residue-aligned part propels the network towards fitting the current errors for the specific task, while the data-aligned part reflects the network's adaptation to or influence by the inherent structure and biases present in the input data distribution. This dual influence provides a potential mechanism for reconciling how networks can be both task-specific and data-adaptive. This is an interesting avenue for future work.

References

[1] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/arora19a.html.

[2] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=akddwRG6EGi.

[3] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: A spiked random matrix perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HlIAoCHDWW.

[4] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[5] Chris M Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

[6] Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52(6):2879–2912, 2024.

[7] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

[8] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

[9] Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue Lu, Lenka Zdeborova, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9662–9695. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cui24d.html.

[10] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

[11] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349):1–65, 2024. URL http://jmlr.org/papers/v25/23-1543.html.

[12] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

[13] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

[14] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8580–8589, Red Hook, NY, USA, 2018. Curran Associates Inc.

[15] Chinmaya Kausik, Kashvi Srivastava, and Rishi Sonthalia. Double descent and overfitting under noisy inputs and distribution shift for linear denoisers. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=HxfqTdLIRF.

[16] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0d1a9651497a38d8b1c3871c84528bd4-Paper.pdf.

[17] Xinyue Li and Rishi Sonthalia. Least squares regression can exhibit under-parameterized double descent. *Advances in Neural Information Processing Systems*, 2024.

[18] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[19] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018. doi: 10.1073/pnas.1806579115. URL https://www.pnas.org/doi/abs/10.1073/pnas.1806579115.

[20] Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks, 2024. URL https://arxiv.org/abs/2310.07891.

[21] Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. Gradient-based feature learning under structured data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Fe8PxP2F2p.

[22] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2798–2806. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/pennington17a.html.

[23] Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/196f5641aa9dc87067da4ff90fd81e7b-Paper.pdf.

[24] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/ed265bc903a5a097f61d3ec064d96d2e-Paper.pdf.

[25] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[26] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2019.06.003. URL https://www.sciencedirect.com/science/article/pii/S0304414918306197.

[27] Rishi Sonthalia and Raj Rao Nadakuditi. Training data size induced double descent for denoising feedforward neural networks and the role of training noise. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=FdMWtpVT1I.

[28] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press, 2012.

[29] Yutong Wang, Rishi Sonthalia, and Wei Hu. Near-interpolators: Rapid norm growth and the trade-off between interpolation and generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 4483–4491. PMLR, 2024.

[30] Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4891–4957. PMLR, 2024.

[31] Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How sharpness-aware minimization minimizes sharpness? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=5spDgWmpY6x.

Contents

Table 1: Notation

Symbol	Meaning	Where first defined / used
d	Input dimension	Assumption 1
n	Number of samples	Assumption 1
m	Hidden-layer width	Assumption 1
$\psi_1 = n/d,\ \psi_2 = m/d$	Proportional-scaling ratios	Assumption 1
$\widehat{\Sigma}$	Bulk covariance matrix	Assumption 2
$\Sigma = \widehat{\Sigma} + \zeta^2 qq^\top$	Full covariance (bulk + spike)	Assumption 2
$\lambda_k = k^{-\alpha}$	Bulk eigen-spectrum	Assumption 2
$\alpha \geq 0$	Spectral-decay exponent	Assumption 2
$\zeta = n^\nu,\ \nu \geq 0$	Spike magnitude	Assumption 2
$q \in \mathbb{S}^{d-1}$	Spike direction	Assumption 2
$z \in \mathbb{R}^n$	Latent coordinates of the spike	Equation (2)
$X = X_B + X_S$	Data split bulk + spike	Equation (2)
X_B	Bulk part ($\mathcal{N}(0, \widehat{\Sigma})$ rows)	Equation (2)
$X_S = \zeta z q^\top$	Rank-1 spike part	Equation (2)
$W \in \mathbb{R}^{m \times d}$	Inner-layer weight matrix	Assumption 3
$a \in \{\pm 1\}^m$	Outer weights (fixed)	Assumption 3
γ_m	Width scale (NTK $= 1/\sqrt{m}$, MF $= 1/m$)	Assumption 3:
$\sigma, \sigma', \sigma''$	Activation and derivatives	Assumption 4
$\mu = \mathbb{E}_x[\sigma'(Wx)]$	Mean derivative vector	Assumption 4
$\sigma'_\perp = \sigma' - \mu$	Centered derivative	Assumption 4
r	Residue vector	Equation (1)
β	Alignment exponent ($\frac{1}{\sqrt{n}\,\|r\|_2} z^\top r$)	Assumption 6
S_1	Residue-aligned rank-1 term	Section 3
S_2	Data-spike-aligned rank-1 term	Section 3
S_{12}	Interpolant rank-1 term	Section 3
$G = \nabla_W \mathcal{L}$	Full gradient wrt W	Prop. 12
E	Error term $G - S_1 - S_{12} - S_2$	Thm. 1
E_L	Error term $G - S_{12} - S_2$ (large-spike version)	Thm. 3
E_2	Bulk error from Jacobian-penalty gradient	Prop. 6
S_3	Data-aligned rank-1 term induced by Jacobian penalty	Prop. 6
$\lambda,\ L_{\text{reg}}$	Reg. strength and Jacobian penalty	Appendix D
τ^2	Variance of isotropic Gaussian noise	Appendix D
$\circ,\ \otimes$	Hadamard / outer products	

Appendix A. Assumption Discussion

A.1. Data and Network Assumptions

Assumption 2 models a bulk component via $\hat{\Sigma}$ and a spike component via q (magnitude ζ) and allows general forms of ill-conditioning with $\lambda_d(\hat{\Sigma}) \to 0$ if $\alpha > 0$, and $\lambda_1(\Sigma) \to \infty$ if $\nu > 0$. This generalizes typical data distribution assumptions like isotropic Gaussian ($\Sigma = I_d$) or uniform on a sphere [2, 11, 20, 22, 30], anisotropic data with a bounded condition number [12, 13], divergent largest eigenvalue and bounded smallest eigenvalue [3, 15, 17, 21, 27], or bounded largest eigenvalue and decaying smallest eigenvalue [4, 6, 29].

he assumption on a is standard. The assumption on W (unit-norm rows) relaxes typical literature requirements (e.g., isotropic Gaussian or uniformly spherical w_j). This allows modeling anisotropic weights, possibly dependent on X, to analyze updates throughout training, not just at initialization. The scaling parameter γ_m defines two common regimes: NTK ($\gamma_m \sim 1/\sqrt{m}$) [1, 14, 16, 18], associated with lazy training where inner weights vary little [8, 18], and MF ($\gamma_m \sim 1/m$), associated with feature learning [7, 19, 23, 26]. These scalings yield different initial output variances ($\mathrm{Var}(f(x)) = \Theta(1)$ in NTK vs. $o(1)$ in MF), impacting dynamics.

A.2. Activation Function Properties

We verify the smoothness and lipschitzness conditions for several common activation functions.

A.2.1. SIGMOID FUNCTION

Let $\sigma(u) = (1 + e^{-u})^{-1}$.
Smoothness: The Sigmoid function is infinitely differentiable (C^∞) for all $u \in \mathbb{R}$.

$$\sigma'(u) = \sigma(u)(1 - \sigma(u))$$
$$\sigma''(u) = \sigma'(u)(1 - 2\sigma(u)) = \sigma(u)(1 - \sigma(u))(1 - 2\sigma(u))$$

Both $\sigma'(u)$ and $\sigma''(u)$ exist for all $u \in \mathbb{R}$.

Lipschitzness: Since Sigmoid is bounded and all derivatives of the sigmoid can be written as a polynomial of sigmoid, we see that the derivatives are bounded and hence lipschitz.

Non-Vanishing Derivative Here we show that if the weight vector w_j is drawn uniformly from the unit sphere $\mathbb{S}^{d-1}$, then the expected derivative $\mu_j = \mathbb{E}_x[\sigma'(w_j^T x)]$ is $\Omega(1)$ when $\nu < 1/2$.

The derivative $\sigma'(u) = \sigma(u)(1 - \sigma(u))$ is bounded. We can see that the argument $u_j = w_j^T x$ is Gaussian $N(0, \sigma_{u_j}^2)$, with variance $\sigma_{u_j}^2 = w_j^T \hat{\Sigma} w_j + n^{2\nu}(w_j^T q)^2$. Then the behavior of μ_j is such that if $\sigma_{u_j}^2 = O(1)$, then $\mu_j = \Omega(1)$. Specifically, if $\sigma_{u_j}^2 \to 0$, then $\mu_j \to \sigma'(0) = 0.25$. If $\sigma_{u_j}^2 \to \infty$, then $\mu_j \to 0$.
Spike Contribution $V_S = n^{2\nu}(w_j^T q)^2$: For a fixed $q \in \mathbb{S}^{d-1}$ and random $w_j \in \mathbb{S}^{d-1}$, the term $(w_j^T q)^2$ concentrates around its mean $\mathbb{E}[(w_j^T q)^2] = 1/d$. With high probability for large d, $(w_j^T q)^2 = \Theta(1/d)$. Then in proportional regime, we have that, $V_S = n^{2\nu} \cdot \Theta(1/n) = \Theta(n^{2\nu-1})$. Since $\nu < 1/2$, $2\nu - 1 < 0$, so $V_S = o(1)$ as $n \to \infty$.
Bulk Contribution $V_B = w_j^T \hat{\Sigma} w_j$: For random $w_j \in \mathbb{S}^{d-1}$, $w_j^T \hat{\Sigma} w_j$ concentrates around $\mathbb{E}[w_j^T \hat{\Sigma} w_j] = \frac{1}{d}\mathrm{Tr}(\hat{\Sigma})$. The eigenvalues $\lambda_k(\hat{\Sigma}) \sim k^{-\alpha}$.

- If $\alpha = 0$: $\mathrm{Tr}(\hat{\Sigma}) = \Theta(d)$, so $V_B = \Theta(1)$.

- If $0 < \alpha < 1$: $\mathrm{Tr}(\hat{\Sigma}) = \Theta(d^{1-\alpha})$, so $V_B = \Theta(d^{-\alpha}) = \Theta(n^{-\alpha}) = o(1)$.
- If $\alpha = 1$: $\mathrm{Tr}(\hat{\Sigma}) = \Theta(\log d)$, so $V_B = \Theta((\log d)/d) = \Theta((\log n)/n) = o(1)$.
- If $\alpha > 1$: $\mathrm{Tr}(\hat{\Sigma}) = \Theta(1)$, so $V_B = \Theta(1/d) = \Theta(1/n) = o(1)$.

Thus, V_B is either $\Theta(1)$ (for $\alpha = 0$) or $o(1)$ (for $\alpha > 0$).

A.2.2. HYPERBOLIC TANGENT (TANH) FUNCTION

Let $\sigma(u) = \tanh(u)$.
Smoothness: The Tanh function is C^∞ for all $u \in \mathbb{R}$.

$$\sigma'(u) = 1 - \tanh^2(u) = \text{sech}^2(u)$$
$$\sigma''(u) = -2\tanh(u)\text{sech}^2(u)$$

Both $\sigma'(u)$ and $\sigma''(u)$ exist for all $u \in \mathbb{R}$.

Lipschitzness:

- For $\sigma(u)$: $\max|\sigma'(u)| = \sigma'(0) = 1$. Thus, $\sigma(u)$ is 1-Lipschitz.
- For $\sigma'(u)$: $\max|\sigma''(u)|$ occurs at $u = \text{arctanh}(\pm 1/\sqrt{3})$, giving $|\sigma''(u)| = \frac{4}{3\sqrt{3}} \approx 0.7698$. Thus, $\sigma'(u)$ is Lipschitz with $L \approx 0.77$ (or $L = 1$ as a looser bound).

$L = 2$ serves as a common upper bound.

Non-vanishing Derivative: Let $\sigma(u) = \tanh(u)$. Its derivative is $\sigma'(u) = \text{sech}^2(u)$. This derivative is always positive, $0 < \sigma'(u) \leq 1$, with a maximum of $\sigma'(0) = 1$, and $\sigma'(u) \to 0$ as $|u| \to \infty$. The analysis of the expected derivative $\mu_j = \mathbb{E}_x[\sigma'(w_j^T x)]$ parallels that of the Sigmoid function.

A.2.3. RECTIFIED LINEAR UNIT (RELU) FUNCTION

Let $\sigma(u) = \max(0, u)$.
Smoothness: Here we see that the derivatives for $u \neq 0$ are as follows

$$\sigma'(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1 & \text{if } u > 0 \end{cases}, \qquad \sigma''(u) = 0 \quad \text{for } u \neq 0$$

Lipschitzness:

- For $\sigma(u)$: $|\sigma'(u)| \leq 1$ a.e. Thus, $\sigma(u)$ is 1-Lipschitz.
- For $\sigma'(u)$: $\sigma'(u)$ is a step function. It is bounded, but not Lipschitz over $\mathbb{R}$ due to the discontinuity at $u = 0$. However, its values are 0 or 1.

Non-vanishing Derivative: Since Wx is symmetric, we get that the mean is 0.5.

A.2.4. EXPONENTIAL LINEAR UNIT (ELU) FUNCTION

Let $\sigma(u) = \begin{cases} u & \text{if } u > 0 \\ e^u - 1 & \text{if } u \leq 0 \end{cases}$.
Smoothness: The derivatives are as follows.

$$\sigma'(u) = \begin{cases} 1 & \text{if } u > 0 \\ e^u & \text{if } u \leq 0 \end{cases} \qquad \sigma''(u) = \begin{cases} 0 & \text{if } u > 0 \\ e^u & \text{if } u < 0 \end{cases}$$

Here we have that σ' is continuous, and σ'' is defined everywhere except for 0.

Lipschitzness:

- For $\sigma(u)$: For $u > 0$, $\sigma'(u) = 1$. For $u \leq 0$, $\sigma'(u) = e^u \in (0, 1]$. Thus $|\sigma'(u)| \leq 1$. So $\sigma(u)$ is 1-Lipschitz.
- For $\sigma'(u)$: For $u > 0$, $\sigma''(u) = 0$. For $u < 0$, $\sigma''(u) = e^u \in (0, 1)$. On $[-1, 1]$, the function is continuous. Hence lipschitz. Thus, we have global lipschitzness.

Non-vanishing Derivative: The derivative dominates the ReLU case. Hence μ_j is at least 0.5.

A.2.5. SWISH FUNCTION

Let $\sigma(u) = u \cdot \text{sigmoid}(u) = u(1 + e^{-u})^{-1}$.
Smoothness: This follows from smoothness of Sigmoid.

Lipschitzness: Let $S(u) = \text{sigmoid}(u) = (1 + e^{-u})^{-1}$. Then $\sigma(u) = uS(u)$.

- For $\sigma(u)$: The first derivative is:

$$\sigma'(u) = S(u) + uS'(u) = S(u) + uS(u)(1 - S(u))$$

  This is a continuous function that decays to zero. Hence is bounded.

- For $\sigma'(u)$: The second derivative of $\sigma(u)$ is:

$$\begin{aligned} \sigma''(u) &= \frac{d}{du}(S(u) + uS'(u)) = S'(u) + (S'(u) + uS''(u)) \\ &= 2S'(u) + uS''(u) \end{aligned}$$

  This is a continuous function that decays to zero. Hence is bounded.

- For $\sigma''(u)$: The third derivative of $\sigma(u)$ is:

$$\begin{aligned} \sigma'''(u) &= \frac{d}{du}(2S'(u) + uS''(u)) = 2S''(u) + (S''(u) + uS'''(u)) \\ &= 3S''(u) + uS'''(u) \end{aligned}$$

  This is a continuous function that decays to zero. Hence is bounded.

Therefore, $\sigma(u)$, $\sigma'(u)$, and $\sigma''(u)$ are all Lipschitz for Swish with $\beta = 1$.

Non-vanishing Derivative: The expected derivative μ_j is:

$$\begin{aligned}\mu_j = \mathbb{E}[\sigma'(u_j)] &= \mathbb{E}[S(u_j) + u_j S'(u_j)] \\ &= \mathbb{E}[S(u_j)] + \mathbb{E}[u_j S'(u_j)]\end{aligned}$$

We evaluate each term:

For $\mathbb{E}[S(u_j)]$: The function $g(u) = S(u) - 1/2$ is an odd function. Since $u_j \sim N(0, \sigma_{u_j}^2)$ has a probability density function symmetric about 0, the expectation of any odd function of u_j is 0. Thus, $\mathbb{E}[S(u_j) - 1/2] = 0$, which implies $\mathbb{E}[S(u_j)] = 1/2$.

For $\mathbb{E}[u_j S'(u_j)]$: The derivative of sigmoid, $S'(u) = S(u)(1 - S(u))$, is an even function: $S'(-u) = S(-u)(1 - S(-u)) = (1 - S(u))S(u) = S'(u)$. The product $h(u) = uS'(u)$ is an odd function, being the product of an odd function (u) and an even function ($S'(u)$). Since $u_j \sim N(0, \sigma_{u_j}^2)$ has a symmetric PDF about 0, $\mathbb{E}[u_j S'(u_j)] = 0$.

Combining these results:

$$\mu_j = 1/2 + 0 = 1/2$$

The value $1/2$ is a positive constant, independent of other parameters such as d, n, m, ν, α, or the specifics of Σ (provided it is positive definite) and w_j (provided $w_j \in \mathcal{S}^{d-1}$).

A.2.6. SOFTPLUS FUNCTION

Let $\sigma(u) = \log(1 + e^u)$.

Smoothness: The Softplus function is C^∞ for all $u \in \mathbb{R}$.

$$\begin{aligned}\sigma'(u) &= \frac{e^u}{1 + e^u} = \text{sigmoid}(u) \\ \sigma''(u) &= \frac{e^u}{(1 + e^u)^2} = \text{sigmoid}(u)(1 - \text{sigmoid}(u))\end{aligned}$$

Both $\sigma'(u)$ and $\sigma''(u)$ exist for all $u \in \mathbb{R}$.
Lipschitzness: The lipschitzness follows from the boundedness and lipschitzness of sigmoid.

Non-vanishing Derivative: Following the argument presented for the Swish activation function, the mean is 0.5.

A.3. Loss Function Derivatives

Let us see what this is for some common loss functions.

- For the Mean Squared Error (MSE) loss,

$$L(f(X)) = \frac{1}{2}\|f(X) - y\|^2 = \frac{1}{2}\sum_{i=1}^{n}(f(x_i) - y_i)^2 \quad \text{and} \quad L'(f(x)) = f(x) - y.$$

- For the Binary Cross Entropy (BCE) loss, we assume the network produces logits $z = f(X) \in \mathbb{R}^n$ with associated class-one probabilities $p = \text{sigmoid}(z) = \frac{1}{1+e^{-z}} \in \mathbb{R}^n$ computed component wise. Then, for given output data $y \in \{0,1\}^n$,

$$L(f(X)) = -\sum_{i=1}^{n} \Big[y_i \ln(p_i) + (1-y_i)\ln(1-p_i) \Big], \quad L'(f(X)) = p - y = \text{sigmoid}(f(X)) - y.$$

- For the Hinge loss for binary classification with output data $y \in \{-1,1\}^n$, $f(X) \in \mathbb{R}^n$ and

$$L(f(X)) = \sum_{i=1}^{n} \max\left(0,\, 1 - y_i\, f(x_i)\right).$$

  Then $L'(f(X))$ is the vector whose ith entry is given by the subgradient

$$\frac{\partial L}{\partial f(x_i)} = \begin{cases} 0, & \text{if } y_i\, f(x_i) \geq 1, \\ -y_i, & \text{if } y_i\, f(x_i) < 1. \end{cases}$$

A.4. Residue Concentration

1. Suppose the training labels satisfy $y_i = f_*(\mathbf{x}_i) + \xi_i$, where f_* is Lipschitz and ξ_i are i.i.d. subgaussian random variables. Then, for independent W and X, lipschitz activation functions and for either the MSE or Binary Cross Entropy (BCE) loss the residues are subgaussian variables and satisfy this assumption.

2. For binary classification with the hinge loss, then since $a_i \sim \text{Unif}(\pm 1)$ we have with probability $1 - o(1)$ that at least a constant fraction of the data points satisfy $1 - y_i f(x_i) \geq 0$, and therefore $r_i = \pm 1$. As a result the assumption holds at initialization.

A.5. β Alignment

Here we consider `Sigmoid`, `ReLU`, `Tanh`, `ELU`, `Softplus`, and `Swish` activation functions. For each activation function, we consider three different loss functions - MSE, BCE, and Hinge. Then for for each activation and loss function combination, we consider $(\nu, \alpha) \in \{1/8, 3/8, 5/8\} \times \{0, 1/2\}$. This gives us 96 scenarios. We do each each scenario for the Mean Field and NTK scalings. For each scenario we let $\psi_1 = 0.75$ and $\psi_2 = 1.25$. We consider $n \in \{750, 1500, 2250, 3000, 3750\}$. We use triple index targets

$$f(x) = \text{sigmoid}(\beta_1^T x) + \tanh(\beta_2^T x) + \text{relu}(\beta_3^T x)$$

for three unit vectors $\beta_1, \beta_2, \beta_3$. For each value we do 50 trials to get the mean inner product $|\frac{1}{\sqrt{n}\|r\|} z^T r|$. Then we then estimate beta using linear regression.

Figure 3, presents the estimates βs. Here we see that β has a mode around 1. Recall if z_1, z_2 are independent uniformly unit norm vectors. Then $z_1^T z_2 \sim d^{-1}$. Figure 3, however, that many βs are bigger than 1. This suggest z, r are rapidly becoming orthogonal. Note that negative βs are cases, where the alignment improves, so z, r are becoming parallel. Eventually, the inner product will saturate at 1 and β should be close to zero. The reason we get negative βs is due to the limited range of n used for the experiments.

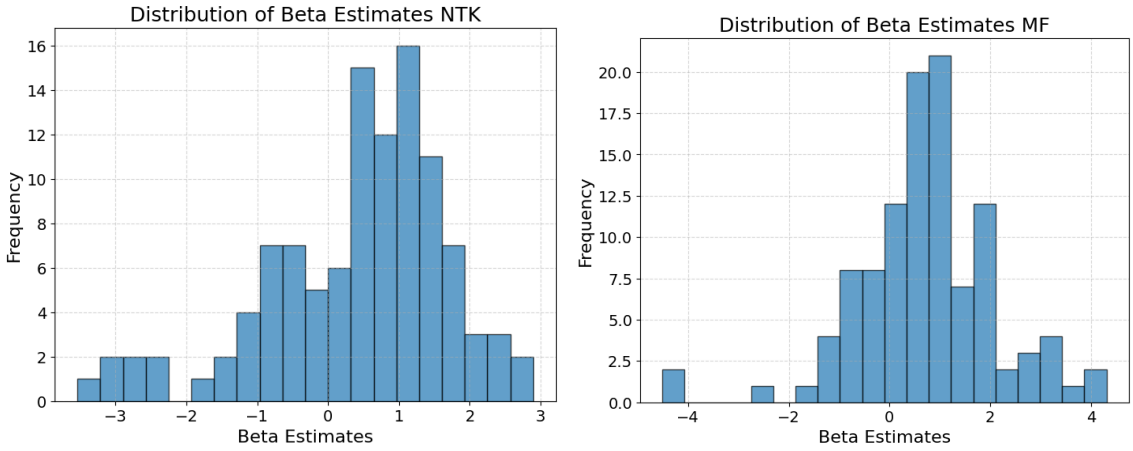


Figure 3: Estimated β values

Appendix B. Empirical Details

All code for the experiments can be found at Link.

The following details are common for all experiments.

Hardware: All experiments were run on Google Colab using an A100.

Data X: We sampled q uniformly randomly from the unit sphere and we used a diagonal $\hat{\Sigma}$.

μ Estimation: We estimate μ using 10000 samples.

Targets: The triple index model we used is as follows.

$$f(x) = \text{sigmod}(\beta_1^T x) + \tanh(\beta_2^T x) + \text{relu}(\beta_3^T x)$$

For three unit vectors $\beta_1, \beta_2, \beta_3$.

When using MSE loss, we let

$$y = f(x) + \varepsilon$$

for standard gaussian noise ε.

When using BCE loss, we use

$$y = f(x).$$

Note that these y are not necessarily in $[0, 1]$. However, the BCE loss is still well defined.

When using Hinge loss,

$$y = \text{sign}(f(x) - 0.5).$$

Note this dataset can be imbalanced.

Alignment determination: To plot the red and blue lines in Figures 1,2,3,7,8, we use the following procedure. We let $B = S_1 + S_{12} + S_2$ (+ S_3 for the gradient penalty). Then we compute its leading left singular vectors for B. We then check if with q and $X_B^T r$. Thus, how we get the associated singular value and we plot the corresponding lines.

B.1. Figure 1

For non-isotropic W, we generate W_S by sampling the rows i.i.d. from the unit sphere. We then introduce anisotropy, by adding $n^{-1/4}\mathbf{1}q^T$ to W_S and then renormalizing to unit norm. This results in the weight concentrating around q.

B.2. Figure 4

For Figure **(b)**, we generate W_S by sampling the rows i.i.d. from the unit sphere. We then introduce anisotropy, by adding $n^{1/2}\mathbf{1}q^T$ to W_S and then renormalizing to unit norm. This results in the weight concentrating around q.

For Figure **(c)**, we generate W_S by sampling the rows i.i.d. from the unit sphere. Then we project onto the ortho-complement of q and renormalize the rows.

For Figure **(d)**, we generate W_S by sampling the rows i.i.d. from the unit sphere. We then let $W = W_S X^T X$ and renormalize the rows. This results in a W that is highly dependent on X.

B.3. Figure 7

Here we use $\zeta = 0, \alpha = 0$. Hence applies for prior work from [2, 20].

We let $n \in \{100, 200, 300, 400, 500, 600, 700, 800\}$ and use $d = n/2$ and $m = n/3$.

B.4. Later in Training Experiments

Here both network are initialized with the same weight matrix for both the inner and outer layers.

We use a step size of $\eta = \gamma_m^{-1}$. Additionally, after each iteration, we re-normalize the rows of W to have unit norm.

For Figure 8**(c)**, the mean principal angle in the following quantity. Given orthonormal basis $u_1, \ldots, u_k$ and $v_1, \ldots, v_k$ for two subspaces, we form the matrix A via

$$A_{ij} = u_i^T v_j$$

We the compute $\cos(\sigma_i(A))$. These are the principal angles between the subspaces. We then report the mean of angels.

B.5. Real Data Experiments

MNIST Dataset: We load the standard MNIST dataset, divide by 256 to have all entries in $[0, 1]$. We use 1000 centered and flattened MNIST images to form $X \in \mathbb{R}^{1000\times 784}$. We estimate $\nu \approx 0.784 > 1/2$. The data is highly ill-conditioned, suggesting a large effective α.

CIFAR Dataset: We use $n = 1000$ CIFAR-10 training images, processed through a pretrained ResNet-18 (on ImageNet) to extract 512-dimensional penultimate-layer activations, forming $X \in \mathbb{R}^{1000\times 512}$. We estimate $\nu \approx 0.3572 < 1/2$ and $\alpha \approx 0.6$.

Specifically, the code for the transformations are as follows.

```
resnet18(weights=ResNet18_Weights.DEFAULT)

transform = transforms.Compose([
    transforms.Resize(224),
    transforms.ToTensor(),
    transforms.Normalize(mean=[0.485, 0.456, 0.406],  # ResNet defaults
                         std=[0.229, 0.224, 0.225])
])
```

Appendix C. Large Spike ($\nu \geq 0.5$): Non $\mathcal{C}^2$ Activations and Dependence between W and X

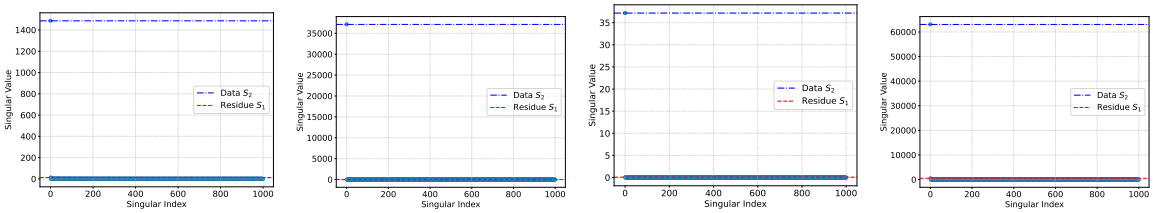
Theorem 3 (Large data-spike gradient approximation) *Suppose Assumptions 1, 2, 3, 4, 5, and 6 are satisfied, and define $E_L = G - S_{12} - S_2$. Then, with probability $1 - o(1)$ for $\nu \geq \frac{1}{2}$ we have*

$$\frac{\|E_L\|_2}{\sqrt{m}\gamma_m \|r\|_\infty} = O(1), \frac{\|S_{12}\|_2}{\|E_L\|_2} = \Omega\left(\frac{n^{\nu-\frac{\beta}{2}}}{\log n}\right), \frac{\|S_2\|_2}{\|E_L\|_2} = \Omega\left(\frac{n^\nu}{\log n}\frac{\|(z\circ r)^T\sigma'_\perp(XW^T)\|_2}{\|\sigma'_\perp(XW^T)\|_2}\right). \tag{6}$$

Note this is a generalization of [3], which required alignment between the targets y and the spike q. Theorem 3 shows that if $\nu > \frac{\beta}{2}$, or if

$$n^\nu = \omega\left(\log n \frac{\|\sigma'_\perp(XW^T)\|_2}{\|(z\circ r)^T\sigma'_\perp(XW^T)\|_2}\right), \tag{7}$$

holds, then the gradient is approximately rank one. In contrast to the $\nu < \frac{1}{4}$ case, this rank-one gradient aligns closely with the data spike plus interpolant $S_{12} + S_2$ rather than the residue S_1. This is empirically verified in Figure 4 for non-$\mathcal{C}^2$ activations ReLU, as well as dependent and independent W and X.



(a) ReLU, Isotropic W **(b)** Swish, $W_S + \sqrt{n}q\mathbf{1}^T$ **(c)** Sigmoid, $W_{S\perp q}$ **(d)** ELU, $W_S X^T X$

Figure 4: Singular value distributions of the gradient G under various activation functions and weight matrix initializations and structures, with a large data spike $\nu = 3/4$. W_S denotes the random matrix with rows drawn mutually i.i.d. uniformly from the unit sphere. The rows of $W_{S\perp q}$ are uniform on the sphere and orthogonal to q. All weight matrices are subsequently normalized to have unit norm rows. Fixed parameters: bulk decay exponent $\alpha = 0$, $n = 750$, $d = 1000$, $m = 1250$, NTK-like scaling ($\gamma_m = 1/\sqrt{m}$), MSE loss, and triple-index model targets.

Appendix D. Regularization

For what follows let $G^{(0)}$ denote the un-regularized gradient matrix derived in Proposition 12.

ℓ_2 weight decay. Adding the term $\frac{\lambda}{2}\|W\|_F^2$ to the loss function modifies the gradient to $G^{(\lambda)} = G^{(0)} + \lambda W$. Theorem 4 implies that if $\lambda\|W\|_2 = o(\sqrt{m}\gamma_m)$ it cannot suppress S_1 or S_2, however, if $\lambda\|W\|_2 = \omega(\sqrt{m}\gamma_n n^\nu)$ then it suppresses both spikes.

Proposition 4 *Given Assumptions 1, 2, 3, 4, and 6. If $\|r\|_2 = O(\sqrt{n})$, then with probability $1-o(1)$ we have that $\|S_1\|_2 \le O(\sqrt{m}\gamma_m)$, $\|S_{12}\|_2 \le O(\sqrt{m}\gamma_m n^{\nu-\frac{\beta}{2}})$, and $\|S_2\|_2 \le O(\sqrt{m}\gamma_m n^\nu)$.*

Isotropic Gaussian input noise. This regularization technique involves adding independent isotropic Gaussian noise $\xi_i \sim \mathcal{N}(0, \tau^2 I)$ to each input x_i without changing the corresponding labels y_i. [5] showed that training with input noise is equivalent under certain conditions to adding a Tikhonov regularizer to the loss, often related to $\sum_{i=1}^n \|\nabla_x f(x_i)\|_2^2$. More recent work [31] connects adding isotropic noise to the data to controlling the trace of the Hessian of the loss function.

Let us define $x_i' = x_i + \xi_i$. This changes the input data distribution, effectively modifying the bulk covariance from $\hat{\Sigma}$ to $\hat{\Sigma}' = \hat{\Sigma} + \tau^2 I$. Consequently, derived quantities such as the residue vector r', the alignment parameter β', the gradient components S_1', S_2', S_{12}', the error term E', and the effective bulk spectral decay α' are denoted with primes.

Proposition 5 (Isotropic Gaussian noise) *Assume the setup of Assumptions 1, 2, 3 with independent X and W. Assume σ satisfies Assumption 4 for the noisy data X'. Additionally, suppose the modified residues satisfy $r_i' = \Theta(1)$ with probability $1 - o(1)$, and Assumption 6 holds for r' with scaling parameter β'. If $\tau^2 = n^\rho$ and $\|\sigma_\perp'(X'W^T)\|_2 = o(n)$, then with high probability:*

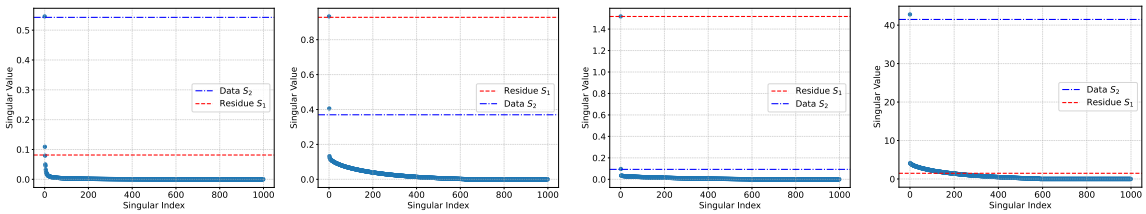
$$\frac{\|S_1'\|_2}{\|E'\|_2} \ge \omega(1), \quad \frac{\|S_2'\|_2}{\|E'\|_2} \le O(n^{\nu-\frac{\rho}{2}}), \quad \frac{\|S_{12}'\|_2}{\|E'\|_2} \le o(n^{\nu-\frac{\rho}{2}-\frac{\beta'}{2}}).$$

Proposition 5 analyzes the effect of input noise. It indicates that the residue spike S_1' remains prominent relative to the error term E'. Conversely, if the noise is sufficiently strong, the data spike components S_2' and S_{12}' become suppressed relative to E'. Intuitively, adding noise with variance $\tau^2 = n^\rho$ increases the variance of the bulk data component. This boosts the overall scale of terms involving $(X_B')^T$. Simultaneously, the added noise tends to make the pre-activations $W^T X'$ more isotropic, which can reduce the operator norm $\|\sigma_\perp'(X'W^T)\|_2$ relative to its Frobenius norm, potentially limiting the growth rate of $\|E'\|_2, \|S_2'\|$. This predicted relative enhancement of S_1' and suppression of S_2' is verified empirically. As discussed in Section 3 (cf. Proposition 2), ReLU can hinder residue spike S_1. However, Figure 5 shows that with small amount of input noise $\tau^2 = 0.25$, an initially suppressed S_1' re-emerges, while S_2' is diminished relative to S_1' and the bulk.

Jacobian penalization. Another form of regularization penalizes the sensitivity of the network output to changes in the inner weights. We consider the Jacobian penalty $L_{reg} = \lambda \frac{1}{2n} \sum_{i=1}^n \|\partial_W f(x_i)\|_2^2$. To analyze this effect of L_{reg} on the gradient, we derive the gradient of L_{reg} with respect to W.

Proposition 6 (Gradient penalty) *Let $\mathrm{Diag}(\|x_i\|^2)$ be the $n \times n$ diagonal matrix, whose entries are $\|x_i\|^2$. If σ is twice differentiable, then*

$$\nabla_W L_{reg} = \frac{1}{n}\lambda\gamma_m^2 \left(\sigma'(WX^T) \odot \sigma''(WX^T)\right) \mathrm{Diag}(\|x_i\|^2) X.$$



(a) $\tau^2 = 0$. ReLU suppresses the residue spike. **(b)** $\tau^2 = 0.25$. The residue spike re-appears. **(c)** $\lambda = 0$. Residue spike is dominant. **(d)** $\lambda = 100$. Data spike is dominant.

Figure 5: Effect of regularization. Panels **(a)**, **(b)** are for isotropic Gaussian noise. Parameters: $n = 750, d = 1000, m = 1250, \nu = 1/8, \alpha = 8/9$ (for original data), triple-index targets, ReLU activation, MSE loss. Panels **(c)**, **(d)** are for Jacobian norm penalization. As λ increases, the size of S_1 does change, size of the bulk E_2 grows, and the size of the data spike S_3 grows. Parameters: NTK, $\nu = 3/8$, $\alpha = 0$, Sigmoid and MSE, and triple-index model targets.

The gradient of the regularizer factorizes into a *data-aligned* rank-one spike S_3 and error E_2:

$$S_3 = \frac{1}{n}\gamma_m^2 X_S^T \Psi, \quad E_2 = \frac{1}{n}\gamma_m^2 X_B^T \Psi, \quad \Psi = \mathrm{Diag}(\|x_i\|^2)\left(\sigma'(XW^T) \odot \sigma''(XW^T)\right).$$

Proposition 7 *Given Assumptions 1, 2, 3, 4, and 6. If* $\|r\|_2 = \Theta(\sqrt{n})$, $\alpha < 1$, *and a constant fraction of the entries of* $\sigma'(XW^T) \odot \sigma''(XW^T)$ *are bounded away from 0, then*

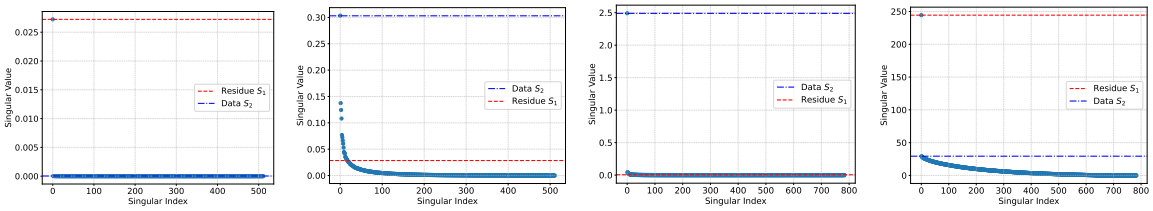
$$\lambda\left(n^{2\nu-\frac{\alpha}{2}-\frac{1}{2}} + n^{\frac{1-3\alpha}{2}}\right) \geq \sqrt{m}\gamma_m \frac{\|\lambda E_2\|_2}{\|S_1\|_2} \geq \lambda\left(n^{2\nu-\frac{\alpha}{2}-1} + n^{-\frac{3}{2}\alpha}\right).$$

If $\nu > \frac{1}{2} + \frac{\alpha}{2}$, then we have that asymptotically the residue spike does not escape the bulk for any $\lambda = \Theta(1)$. If $\nu < \frac{1}{2}$, we see that increasing λ suppresses the residue spike. For the data spike, we have that λS_3 will grow as λ grows. Hence this enhances the data spike. We empirically verify that increasing λ kills the residue spike while promoting the data spike (Figure 5).

Real-Data validation. The identified low-rank spike-plus-bulk gradient structure and the discussed regularization effects are observable in two standard vision datasets - MNIST and CIFAR10. For MNIST, we estimate $\nu \approx 0.784 > 1/2$ and the data is highly ill-conditioned, suggesting a large effective α. Theorem 3 predicts a gradient dominated by data-aligned components (Panel **(c)** of Figure 6). Adding isotropic Gaussian noise with $\sigma^2 = 100$ (Panel **(d)**) suppresses the original data-aligned spike and enhances the residue-aligned spike S_1, consistent with the analysis in Section D. For CIFAR-10 we use a pretrained ResNet-18 (on ImageNet) to extract 512-dimensional embedding. We estimate $\nu \approx 0.3572 < 1/2$ and $\alpha \approx 0.6$. For these parameters Theorem 1 suggests S_1 (residue-aligned) can be prominent. Panel **(a)** of Figure 6 shows a dominant S_1. Applying Jacobian regularization with $\lambda = 10^5$ (Panel **(b)**) suppresses S_1 and promotes a data-aligned spike (akin to S_2), consistent with the behavior analyzed for Jacobian penalization in Section D.

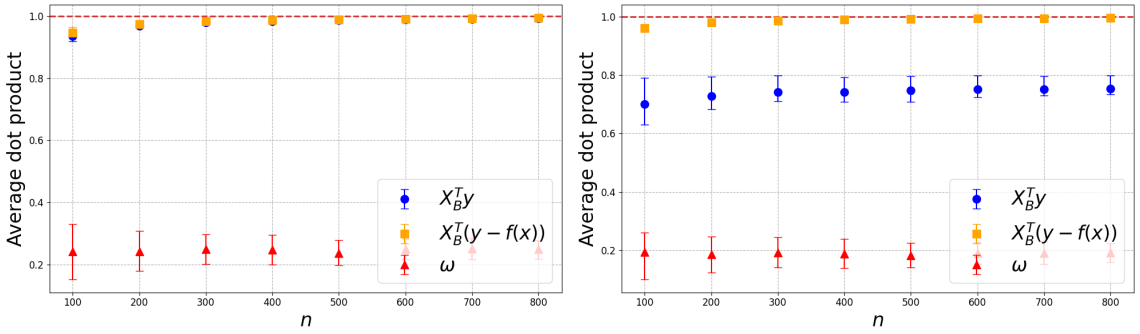
Appendix E. Later in Training

1) Alignment at initialization: residue r versus target vector y. Recall from Theorem 1 that in the small spike regime the gradient is dominated by S_1. Further, for the MF scaling the residue r is approximately equal to the target y, while for the NTK scaling the residue can be quite distinct from y. This implies the alignment of the gradient may differ significantly depending on whether an



(a) CIFAR, $\lambda = 0$ **(b) CIFAR**, $\lambda = 10^5$ **(c) MNIST**, $\sigma^2 = 0$ **(d) MNIST**, $\sigma^2 = 100$

Figure 6: Gradient singular value spectra on real datasets. Each panel displays the singular values of the gradient matrix G under the specified conditions.



(a) MF Scaling **(b)** NTK Scaling

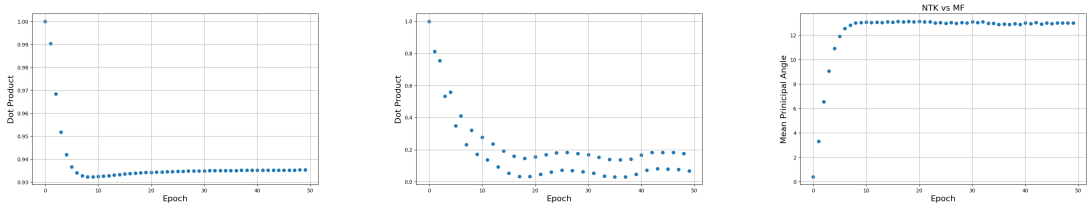
Figure 7: Empirical alignment (normalized inner product) of the top singular vector of the gradient G with $X_B^T y$, $X_B^T r$ and ω for data from a single-index model $y = \text{Sigmoid}(\omega^T x) + \text{noise}$. We use isotropic X, ReLU activation, and MSE loss. We average over 500 samples of a, W, X, y. The error bars are the 25th and 75th percentile.

MF or NTK scaling is used. Suppose $y = \text{sigmoid}(\omega^T x) + \varepsilon$, then Figure 7 presents the normalized inner products between the leading left singular vector of G and three candidate directions $X_B^T y$, $X_B^T r$, and ω. For the MF scaling, we see that the gradient's dominant direction aligns well with $X_B^T r$, $X_B^T y$, consistent with Theorem 1 and [2]. For the NTK scaling, consistent with Theorem 1, the gradient exhibits strong alignment with $X_B^T r$. This differs notably from both $X_B^T y$ and the ω alignment directions predicted in [20] which we believe to be erroneous.

2) Stability of the gradient during early training. Let G_t denote the gradient after t iterations of GD. In Figure 8 we plot the alignment between the leading left singular vector of G_0 and subsequent leading left singular vectors of G_t under both MF and NTK scalings. The following is quite striking: the dominant gradient direction under the MF scaling remains stable throughout training while for the NTK scaling it evolves significantly. This leads to a divergence in the trajectories of the weight matrix even with identical initialization and training data.

Towards explaining this, suppose the conditions of Theorem 1 hold at least approximately up to some iteration $t \leq T$. Then under an MF scaling the gradient is approximated by a rank-one matrix whose left singular vector is nearly constant $X_B^T r_t \approx X_B^T y$. Therefore it remains stable over a number of iterations. If the NTK scaling is used instead, then as S_1 is proportional to $X_B^T r_t \not\approx X_B^T y$ and the gradient depends on the residuals r_t which evolve throughout training.

3) Phase transitions both by epoch and data spike size. In Figure 9 we observe the evolution of the alignment of the gradient versus the data spike and the residue under the MF scaling. Moving

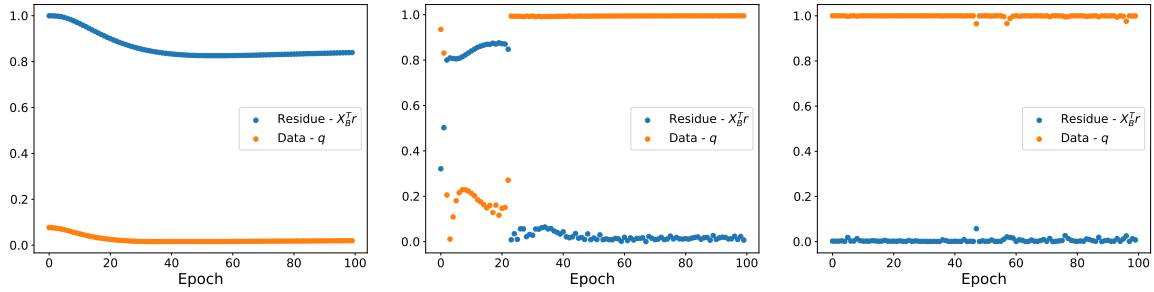


(a) MF scaling (notice the small range)

(b) NTK scaling (oscillatory)

(c) Mean principal angle in degrees between W for the NTK and MF.

Figure 8: Evolution of the gradient direction and weight matrix during training under GD with Weight Normalization (WN) for the MF and NTK scalings. Fixed parameters are $\nu = 0, \alpha = 0$ while using the Sigmoid activation function and the MSE loss. Plots (a) and (b) show the alignment (normalized inner product) between the leading left singular vector of the initial gradient G_0 (epoch 0) and that of G_t (epoch t). Plot (c) shows the mean principal angle between the weight matrices learned under the MF and NTK scalings with identical initialization and training data.



(a) Small ν : $\nu = 0.125$

(b) Medium ν : $\nu = 0.4375$

(c) Large ν : $\nu = 0.75$

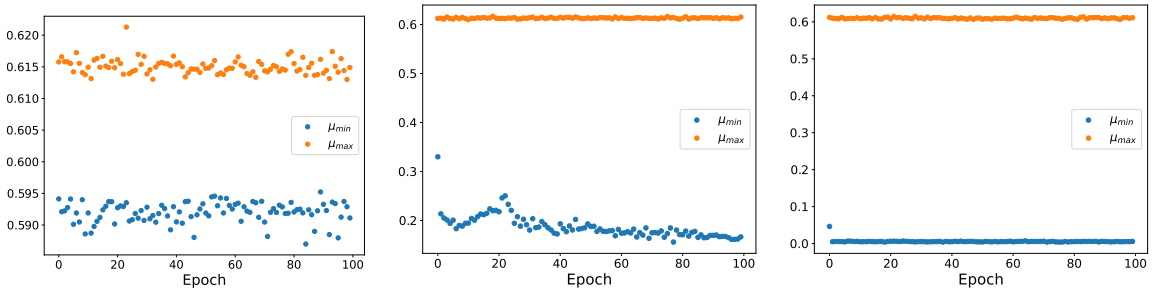
Figure 9: Evolution of the alignment of the leading left singular vector of G_t with data spike q and residue $(X_B^T r_t)$ during training. Fixed parameters: MF scaling, Tanh activation, MSE loss, $\alpha = 0$.

from small to large spike sizes we observe a transition in the gradient alignment from the residue $X_B^T r$ to the data spike q. We remark that this is as predicted by Theorem 1 and Theorem 3 at initialization. Of particular interest is the middle spike size setting, where we witness a phase transition during training of the gradient alignment from residue to data spike. We only pause to highlight this interesting phenomenon here and leave a more thorough analysis to future work.

E.1. Assumption Satisfactions

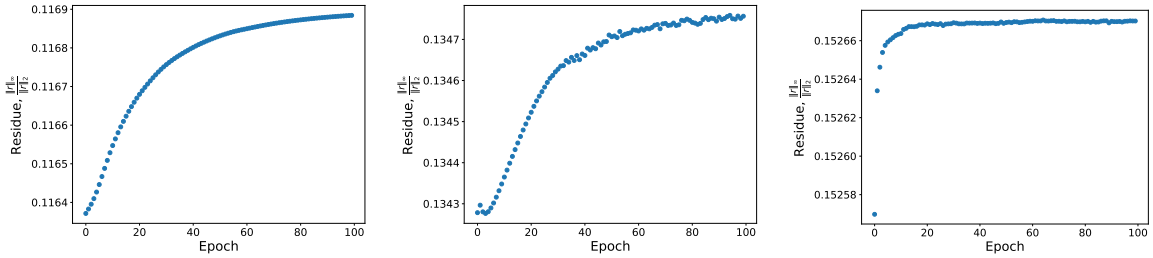
For Theorem 1 and Theorem 3 to apply beyond initialization, during training we require certain assumptions to hold. We begin by considering the common assumptions needed for both theorems.

1. Assumption 1 concerns the proportional regime and hence holds during training.
2. Assumption 2 concerns the data generation process and hence also holds during training.
3. Assumption 3 concerns the network initialization, and scaling, hence the assumptions on a and γ_m continue to hold during training. Moreover, through the use of weight normalization the assumption that the rows of W are on the unit sphere also holds.
4. Assumption 4 concerns the activation function, namely its smoothness and Lipschitzness which of course also hold during training. However, it is not clear that the assumption on the non-vanishing



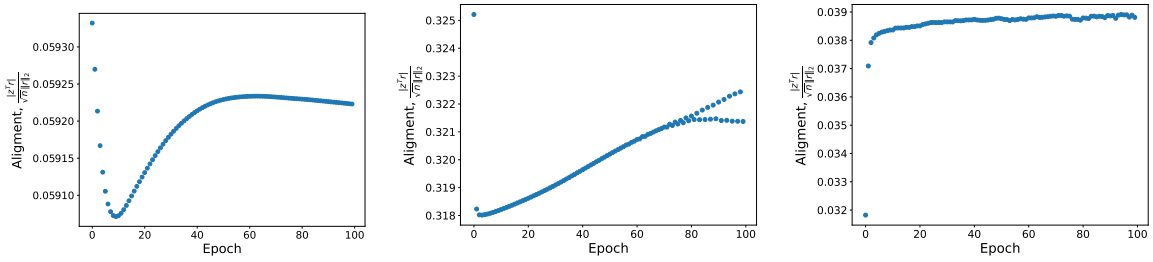
(a) Small ν : $\nu = 0.125$ **(b)** Medium ν : $\nu = 0.4375$ **(c)** Large ν : $\nu = 0.75$

Figure 10: Evolution of $\mu_{\min}$ and μ_{max} during training. Fixed parameters: MF scaling, Tanh activation, MSE loss, $\alpha = 0$.



(a) Small ν : $\nu = 0.125$ **(b)** Medium ν : $\nu = 0.4375$ **(c)** Large ν : $\nu = 0.75$

Figure 11: Evolution of $\|r\|_\infty / \|r\|_2$ during training. Fixed parameters: MF scaling, Tanh activation, MSE loss, $\alpha = 0$.



(a) Small ν : $\nu = 0.125$ **(b)** Medium ν : $\nu = 0.4375$ **(c)** Large ν : $\nu = 0.75$

Figure 12: Evolution of $|z^T r/(\sqrt{n}\|r\|_2)$ during training. Fixed parameters: MF scaling, Tanh activation, MSE loss, $\alpha = 0$.

gradient is satisfied. Despite this, we empirically verify as per Figure 10 that it does hold during training at least for small ν. For moderate $\nu = 7/16$ we observe that μ_{min} appears to decrease, hence later in training this assumption may be violated. For large $\nu = 3/4$,the assumption only appears to hold for the first iteration. We remark that this results in the suppression of S_1 and S_{12} but does not effect S_2 or E. As a result, we suspect that the data spike q remains dominant.

5. Assumption 5. This is the assumption that

$$\frac{\|r\|_\infty}{\|r\|_2} = O\left(\frac{\log n}{\sqrt{n}}\right).$$

Figure 11 shows that while this ratio grows, the change is very small. Hence, we believe that this assumptions holds.

6. Assumption 6. This is about the alignment between z and r. Figure 12 shows that while this ratio grows, the change is very small. Hence, we believe that this assumptions holds.

For the additional assumptions required for Theorem 1, clearly if the activation is $\mathcal{C}^2$ at initialization then it is also $\mathcal{C}^2$ throughout training. Finally, although clearly the independence of W_t and X is violated, due to the near constant gradient direction, (at least for the MF scaling) the correlation between W and X remains small.

Appendix F. Proofs

Notation In the appendix, we shall use $f \lesssim g$ to mean that $f = O(g)$ with probability $1 - o(1)$.

F.1. Regularization Proofs

Proposition 8 *Given Assumptions 1, 2, 3, 4, and 6. If $\|r\|_2 = O(\sqrt{n})$, then with probability $1-o(1)$ we have that $\|S_1\|_2 \leq O(\sqrt{m}\gamma_m)$, $\|S_{12}\|_2 \leq O(\sqrt{m}\gamma_m n^{\nu-\frac{\beta}{2}})$, and $\|S_2\|_2 \leq O(\sqrt{m}\gamma_m n^{\nu})$.*

Proof These bound immediately follow from Theorem 13, Theorem 14, and Theorem 15. ■

Proposition 9 (Isotropic Gaussian noise) *Assume the setup of Assumptions 1, 2, 3 with independent X and W. Assume σ satisfies Assumption 4 for the noisy data X'. Additionally, suppose the modified residues satisfy $r'_i = \Theta(1)$ with probability $1 - o(1)$, and Assumption 6 holds for r' with scaling parameter β'. If $\tau^2 = n^\rho$ and $\|\sigma'_\perp(X'W^T)\|_2 = o(n)$, then with high probability:*

$$\frac{\|S'_1\|_2}{\|E'\|_2} \geq \omega(1), \quad \frac{\|S'_2\|_2}{\|E'\|_2} \leq O(n^{\nu-\frac{\rho}{2}}), \quad \frac{\|S'_{12}\|_2}{\|E'\|_2} \leq o(n^{\nu-\frac{\rho}{2}-\frac{\beta'}{2}}).$$

Proof We prove each bound in turn.

S'_1 **Bound:** Recall that $S'_1 = \frac{\gamma_m}{n}(X'_B)^T r'(a \circ \mu')^T$. Since $d > n$, and $X'_B \in \mathbb{R}^{n\times d}$ is full rank with probability 1, we have that with probability 1, for any vector v

$$\|(X'_B)^T v\|_2 \geq \sigma_{min}(X)\|v\|_2$$

Since the smallest eigenvalue of $\hat{\Sigma}'$ is n^ρ, with probability $1 - o(1)$, we have that

$$\sigma_{min}(X'_B) \geq n^{\frac{1}{2}+\frac{\rho}{2}}.$$

Applying to S_1', we get

$$\|S_1'\|_2 \gtrsim \gamma_m n^{\rho/2} \|a \circ \mu'\|_2 \frac{\|r\|_2}{\sqrt{n}}$$

Then using Assumption 4, the fact that the entries of a are ± 1, and $r_i' = \Theta(1)$, we get

$$\|S_1'\|_2 \gtrsim \gamma_m n^{\rho/2} \sqrt{m}$$

E **Bound:** Next, we have that $E' = \frac{\gamma_m}{n}(X_B')^T((r'a^T) \circ \sigma_\perp'(X'W^T))$. Using the fact that with probability $1 - o(1)$, $r_i' = \Theta(1)$ and $a_i = \pm 1$, we have that with probability $1 - o(1)$

$$\|(r'a^T) \circ \sigma_\perp'(X'W^T)\|_2 = \|\sigma_\perp'(X'W^T)\|_2.$$

Thus, we have that

$$\begin{aligned}\|E'\|_2 &\lesssim \frac{\gamma_m}{n} \|X_B'\|_2 \|\sigma_\perp'(X'W^T)\|_2 \\ &\lesssim \frac{\gamma_m}{n} \sqrt{n} n^{\rho/2} \|\sigma_\perp'(X'W^T)\|_2\end{aligned}$$

Since $d > n$ and X_B' if full rank with probability 1, we have that with probability 1,

$$\begin{aligned}\|(X_B')^T((r'a^T) \circ \sigma_\perp'(X'W^T))\|_2 &\geq \sigma_{min}(X_B') \|((r'a^T) \circ \sigma_\perp'(X'W^T))\|_2 \\ &= \sigma_{min}(X_B') \|\sigma_\perp'(X'W^T))\|_2\end{aligned}$$

Hence we get

$$\|E'\|_2 \gtrsim \frac{\gamma_m}{n} \sqrt{n} n^{\rho/2} \|\sigma_\perp'(X'W^T)\|_2$$

S_2' **Bound:** Recall that

$$S_2' = \frac{\gamma_m}{n} n^\nu q z^T((r'a^T) \circ \sigma_\perp'(X'W^T))$$

Hence we get that

$$\begin{aligned}\|S_2'\|_2 &= \frac{\gamma_m}{n} n^\nu \|q\|_2 \|z^T((r'a^T) \circ \sigma_\perp'(X'W^T))\|_2 \\ &\leq \frac{\gamma_m}{n} n^\nu \|z\|_2 \|((r'a^T) \circ \sigma_\perp'(X'W^T))\|_2 \\ &\lesssim \frac{\gamma_m}{n} n^{\nu + \frac{1}{2}} \|\sigma_\perp'(X'W^T))\|_2\end{aligned}$$

S_{12}' **Bound:** Recall that

$$S_{12}' = \frac{\gamma_m}{n} n^{nu} q z^T r'(a \circ \mu')^T$$

Thus, we have that

$$\begin{aligned}\|S'_{12}\|_2 &= \frac{\gamma_m}{n} n^{nu} \|z^T r(a \circ \mu')\|_2 \\ &= \frac{\gamma_m}{n} n^{nu} \|z^T r\|_2 \|(a \circ \mu')\|_2 \\ &\lesssim \frac{\gamma_m}{n} n^{nu-\beta'/2} \|r'\|_2 \|z\|_2 \|(a \circ \mu')\|_2 \\ &\lesssim \sqrt{m}\gamma_m n^{\nu - \frac{\beta'}{2}}\end{aligned}$$

Relative Bounds: Thus, we have that using $\|\sigma'_\perp(X'W^T)\|_2 = o(n)$

$$\frac{\|S'_1\|_2}{\|E'\|_2} \gtrsim \frac{n}{\|\sigma'_\perp(X'W^T)\|_2} = \omega(1)$$

For the upper bounds we see that

$$\frac{\|S'_2\|_2}{\|E'\|_2} \lesssim n^{\nu - \frac{\rho}{2}}, \quad \frac{\|S'_{12}\|_2}{\|E'\|_2} \lesssim n^{\nu - \frac{\beta'}{2} - \frac{\rho}{2}} \cdot \frac{\|\sigma'_\perp(X'W^T)\|_2}{n} = n^{\nu - \frac{\beta'}{2} - \frac{\rho}{2}} o(1).$$

■

Proposition 10 (Gradient penalty) *Let* $\mathrm{Diag}(\|x_i\|^2)$ *be the* $n \times n$ *diagonal matrix, whose entries are* $\|x_i\|^2$*. If* σ *is twice differentiable, then*

$$\nabla_W L_{reg} = \frac{1}{n}\lambda\gamma_m^2 \left(\sigma'(WX^T) \odot \sigma''(WX^T)\right) \mathrm{Diag}(\|x_i\|^2) X.$$

Proof Letting $Z = WX^T$ and $f_i = f(x_i)$ then note

$$f_i = a^T \sigma(Wx_i) = a^T h_i, \text{ and } \partial_{h_i} f_i = a.$$

It follows that

$$\partial_{z_i} f_i = \partial_{h_i} f_i \odot \sigma'(z_i) = a \odot \sigma'(z_i).$$

Recall

$$\frac{\partial Z_{rc}}{\partial W_{kj}} = \mathbf{1}_{\{c=k\}} X_{rj},$$

then

$$\frac{\partial f_i}{\partial W_{kj}} = \sum_{c=1}^{m} \frac{\partial f_i}{\partial Z_{ic}} \frac{\partial Z_{ic}}{\partial W_{kj}} = \frac{\partial f_i}{Z_{ik}} X_{ij}$$

and therefore

$$\partial_W f_i = (a \odot \sigma'(Wx_i)) x_i^T$$

Let $g_i = a \odot \sigma'(h_i)$, then

$$\|\partial_W f_i\|_F^2 = \|g_i x_i^T\|_F^2 = \sum_{j,k} g_{ij}^2 x_{ik}^2 = \|g_i\|_2^2 \|x_i\|_2^2.$$

Now

$$\begin{aligned}\frac{\partial}{\partial W_{rc}} \|g_i\|_2^2 &= \frac{\partial}{\partial W_{rc}} \sum_{j=1} a_j^2 \frac{\partial}{\partial W_{rc}} \sigma'(w_j^T x_i)^2 \\ &= \left(2a_r^2 \sigma'(w_j^T x_i)\sigma''(w_j^T x_i)\right) x_{ic}.\end{aligned}$$

The term inside the brackets is independent of c while the term outside the brackets is independent of r. As a result this is an outer product and

$$\partial_W \|g_i\|_2^2 = 2\left(a^{\circ 2} \odot \sigma'(W x_i) \odot \sigma''(W x_i)\right) x_i^T.$$

Note above $a^{\circ 2}$ refers squaring operation being applied elementwise to the vector a. Therefore

$$\begin{aligned}\partial_W R &= \frac{1}{2} \sum_{i=1}^n \partial_W \|\partial_W f_i\|_F^2 && (8)\\ &= \frac{1}{2} \sum_{i=1}^n \|x_i\|_2^2 \partial_W \|g_i\|_2^2 && (9)\\ &= \sum_{i=1}^n \|x_i\|_2^2 \left(a^{\circ 2} \odot \sigma'(W x_i) \odot \sigma''(W x_i)\right) x_i^T. && (10)\\ &= \left(a^{\circ 2} \mathbf{1}^T \circ \sigma'(W X^T) \circ \sigma''(W X^T)\right) \mathrm{Diag}(\|x_i\|^2) X && (11)\\ &= \left(\sigma'(W X^T) \circ \sigma''(W X^T)\right) \mathrm{Diag}(\|x_i\|^2) X && (12)\end{aligned}$$

■

Proposition 11 *Given Assumptions 1, 2, 3, 4, and 6. If $\|r\|_2 = \Theta(\sqrt{n})$, $\alpha < 1$, and a constant fraction of the entries of $\sigma'(XW^T) \odot \sigma''(XW^T)$ are bounded away from 0, then*

$$\lambda\left(n^{2\nu - \frac{\alpha}{2} - \frac{1}{2}} + n^{\frac{1-3\alpha}{2}}\right) \geq \sqrt{m}\gamma_m \frac{\|\lambda E_2\|_2}{\|S_1\|_2} \geq \lambda\left(n^{2\nu - \frac{\alpha}{2} - 1} + n^{-\frac{3}{2}\alpha}\right).$$

Proof We begin by noting that since σ, σ' are lipschitz, we have that σ', σ'' are bounded. Hence

$$\sigma'(XW^T) \odot \sigma''(XW^T)$$

has an operator norm that is at most $O(n)$. Since a constant fraction p of the entries are at least some universal constant c, then in the proportional regime, we have that

$$\|\sigma'(XW^T) \odot \sigma''(XW^T)\|_2 \geq \frac{1}{\sqrt{n}} \|\sigma'(XW^T) \odot \sigma''(XW^T)\|_F \gtrsim \sqrt{mc} = \Omega(\sqrt{n})$$

Recall that

$$E_2 = \frac{1}{n}\gamma_m^2 X_B^T \operatorname{Diag}(\|x_i\|^2)\left(\sigma'(XW^T)\odot\sigma''(XW^T)\right).$$

Then since $d > n$, $X_B^T \operatorname{Diag}(\|x_i\|^2)$ is full rank with probability 1, we have that

$$\frac{1}{n}\gamma_m^2\sigma_{min}\left(X_B^T \operatorname{Diag}(\|x_i\|^2)\right)\|\sigma'(XW^T)\odot\sigma''(XW^T)\|_2 \lesssim \|E_2\|_2$$

and

$$\|E_2\|_2 \lesssim \frac{1}{n}\gamma_m^2\sigma_{max}\left(X_B^T \operatorname{Diag}(\|x_i\|^2)\right)\|\sigma'(XW^T)\odot\sigma''(XW^T)\|_2$$

Due to Assumption 2, with high probability $1-o(1)$, we have that

$$\sigma_{max}(X_B) \lesssim \sqrt{n} \text{ and } \sigma_{min}(X_B) \gtrsim n^{\frac{1-\alpha}{2}}$$

Then since $\|x_i\|^2$ concentrates to $n^{2\nu}+n^{1-\alpha}$ (for $\alpha<1$), we have that

$$\|E_2\|_2 \lesssim \frac{\gamma_m^2}{n}\sqrt{n}(n^{2\nu}+n^{1-\alpha})\|\sigma'(XW^T)\odot\sigma''(XW^T)\|_2$$

and

$$\|E_2\|_2 \gtrsim \frac{\gamma_m^2}{n}n^{\frac{1-\alpha}{2}}(n^{2\nu}+n^{1-\alpha})\|\sigma'(XW^T)\odot\sigma''(XW^T)\|_2$$

Then using the $O(n)$ upper bound on $\|\sigma'(XW^T)\odot\sigma''(XW^T)\|_2$, in the proportional regime, with high probability $1-o(1)$, we get that

$$\|E_2\|_2 \lesssim m\gamma_m^2(n^{2\nu-\frac{1}{2}}+n^{\frac{1}{2}-\alpha})$$

Using our $\Omega(\sqrt{n})$ lower bound on $\|\sigma'(XW^T)\odot\sigma''(XW^T)\|_2$, we get

$$\|E_2\|_2 \gtrsim m\gamma_m^2(n^{2\nu-\frac{\alpha}{2}-1}+n^{-\frac{3\alpha}{2}})$$

On the other hand, if $\|r\|_2=\Theta(\sqrt{n})$ we have that

$$\sqrt{m}\gamma_m n^{-\frac{\alpha}{2}} \lesssim \|S_1\|_2 \lesssim \sqrt{m}\gamma_m$$

For the NTK regime, we have that

$$n^{2\nu-\frac{\alpha}{2}-\frac{1}{2}}+n^{\frac{1-3\alpha}{2}} \gtrsim \sqrt{m}\gamma_m\frac{\|E_2\|_2}{\|S_1\|_2} \gtrsim n^{2\nu-\frac{\alpha}{2}-1}+n^{-\frac{3}{2}\alpha}$$

■

F.2. Spikey Gradient Proof

Proposition 12 (Gradient of the loss) *If Assumption 4 holds and R is differentiable, then*

$$G := \nabla_{W^T} L = X^T \left[(ra^T) \circ \sigma'(XW^T)\right] + \lambda \nabla_{W^T} R(W) \in \mathbb{R}^{d \times m}$$

exists for almost every W in $\mathbb{R}^{m \times d}$.

Proof The first thing we need to do is to compute the gradient. To begin, we compute

$$f(x_i) = \sum_{j=1}^{m} a_j \sigma \left(\sum_{k=1}^{d} w_{jk} (x_i)_k \right)$$

Thus, we see that

$$\begin{aligned}
\frac{\partial}{\partial w_{rs}} L(f(x)) &= \frac{1}{n} \sum_{i=1}^{n} \ell'(f(x_i)) \frac{\partial}{\partial w_{rs}} \left(\sum_{j=1}^{m} a_j \sigma \left(\sum_{k=1}^{d} w_{jk} (x_i)_k \right) \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \ell'(f(x_i)) \sum_{j=1}^{m} a_j \frac{\partial}{\partial w_{rs}} \left(\sigma \left(\sum_{k=1}^{d} w_{jk} (x_i)_k \right) \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \ell'(f(x_i)) \sum_{j=1}^{m} a_j \sigma' \left(w_j^T x_i \right) \frac{\partial}{\partial w_{rs}} \left(\sum_{k=1}^{d} w_{jk} (x_i)_k \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \ell'(f(x_i)) a_r \sigma' \left(w_r^T x_i \right) (x_i)_s \\
&= \frac{1}{n} \sum_{i=1}^{n} (L'(f(X))a)_{ir} \sigma'(XW^T)_{ir} X_{is} \\
&= \frac{1}{n} (X^T [(L'(f(X))a) \circ \sigma'(XW^T)])_{sr}
\end{aligned}$$

■

We begin by decomposing the gradient

$$G = \frac{\gamma_m}{n} X^T \left((ra^T) \circ \sigma'(XW^T) \right).$$

This algebraic decomposition holds for the current state (X, W, r, a), irrespective of any statistical dependence between W and X. Recall the data decomposition

$$X = X_B + X_S = X_B + \zeta z q^T \in \mathbb{R}^{n \times d}$$

where rows of X_B are from $\mathcal{N}(0, \hat{\Sigma})$, $z \sim \mathcal{N}(0, I)$, $\|q\| = 1$ and the activation derivative decomposition $\sigma'(XW^T) = \mathbf{1}_n \mu^T + \sigma'_\perp(XW^T)$, where $\mu = \mathbb{E}_x[\sigma'(Wx)]$ depends on the current W.

Substituting these into the gradient expression yields:

$$\begin{aligned}
G &= \frac{\gamma_m}{n} X^T \left((ra^T) \circ \left[\mathbf{1}_n \mu^T + \sigma'_\perp(XW^T) \right] \right) \\
&= \frac{\gamma_m}{n} X^T \left(r(a \circ \mu)^T + (ra^T) \circ \sigma'_\perp(XW^T) \right) \\
&= \frac{\gamma_m}{n} (X_B^T + X_S^T) \left(r(a \circ \mu)^T \right) + \frac{\gamma_m}{n} (X_B^T + X_S^T) \left((ra^T) \circ \sigma'_\perp(XW^T) \right) \\
&= \underbrace{\frac{\gamma_m}{n} X_B^T r (a \circ \mu)^T}_{S_1} + \underbrace{\frac{\gamma_m}{n} X_S^T r (a \circ \mu)^T}_{S_{12}} \\
&\quad + \underbrace{\frac{\gamma_m}{n} X_S^T ((ra^T) \circ \sigma'_\perp(XW^T))}_{S_2} + \underbrace{\frac{\gamma_m}{n} X_B^T ((ra^T) \circ \sigma'_\perp(XW^T))}_{E}.
\end{aligned}$$

Using $X_S = \zeta z q^T$, we identify the components explicitly:

$$\begin{aligned}
S_1 &= \gamma_m \frac{X_B^T r}{n} (a \circ \mu)^T \\
S_{12} &= \gamma_m \zeta \left(\frac{z^T r}{n} \right) q (a \circ \mu)^T \\
S_2 &= \frac{\gamma_m \zeta}{n} q \left(z^T ((ra^T) \circ \sigma'_\perp(XW^T)) \right) \\
E &= \frac{\gamma_m}{n} X_B^T ((ra^T) \circ \sigma'_\perp(XW^T)).
\end{aligned}$$

Note that S_{12} shares its right singular vector $(a \circ \mu)$ with S_1 (up to scaling) and its left singular vector q with S_2. Understanding the gradient structure requires bounding the norms of these terms, which depends on the properties of the current W, r, μ, and the data statistics.

F.2.1. UPPER AND LOWER BOUNDS

Given our helper results, we now provide bounds for the S_1, S_{12}, S_2, and E appearing in Section 3.

Lemma 13 (S_1 Bound) *Let W be the weight matrix (e.g., at step t) with unit norm rows, and let $S_1 = \gamma_m \frac{X_B^T r}{n} (a \circ \mu)^T$. Suppose X_B is from Assumption 2, a has fixed ± 1 entries (Assumption 3), r is the current residual, and $\mu = \mathbb{E}_x[\sigma'(Wx)]$ satisfies $\mu_k = \Theta(1)$ for all k (Assumption 4). Assume $d > n$. Then with high probability:*

$$\sqrt{m} \gamma_m \mu_{\min} \|r\|_2 n^{-\frac{\alpha+1}{2}} \lesssim \|S_1\|_2 \lesssim \sqrt{m} \gamma_m \mu_{\max} \|r\|_2 n^{-\frac{1}{2}},$$

where $\mu_{\min} = \min_k |\mu_k| = \Omega(1)$ and $\mu_{\max} = \max_k |\mu_k| = O(1)$.

Proof The operator norm is

$$\|S_1\|_2 = \frac{\gamma_m}{n} \|X_B^T r\|_2 \|a \circ \mu\|_2.$$

First, consider $a \circ \mu$, where $a_k = \pm 1$ and $\mu_k = \mathbb{E}_x[\sigma'(w_k^T x)]$. By assumption, $\mu_{\min} = \min_k |\mu_k| = \Omega(1)$ and $\mu_{\max} = \max_k |\mu_k| = O(1)$ (since σ' is bounded). We have:

$$\|a \circ \mu\|_2^2 = \sum_{k=1}^m a_k^2 \mu_k^2 = \sum_{k=1}^m \mu_k^2.$$

Thus, we see that

$$\mu_{\min}\sqrt{m} \leq \|a \circ \mu\|_2 \leq \mu_{\max}\sqrt{m}.$$

By Assumption 2, if $d > n$ we have that with high probability

$$n^{\frac{1-\alpha}{2}}\|r\|_2 \lesssim \|X_B^T r\|_2 \lesssim n^{\frac{1}{2}}\|r\|_2.$$

Substituting the bounds for $\|X_B^T r\|_2$ and $\|a \circ \mu\|_2$ into the expression for $\|S_1\|_2 = \frac{\gamma_m}{n}\|X_B^T r\|_2\|a \circ \mu\|_2$:

$$\text{Lower:} \quad \|S_1\|_2 \gtrsim \frac{\gamma_m}{n}(n^{\frac{1-\alpha}{2}}\|r\|_2)(\sqrt{m}\mu_{\min}) = \gamma_m\sqrt{m}\mu_{\min}\|r\|_2 n^{-\frac{\alpha+1}{2}}$$
$$\text{Upper:} \quad \|S_1\|_2 \lesssim \frac{\gamma_m}{n}(n^{\frac{1}{2}}\|r\|_2)(\sqrt{m}\mu_{\max}) = \gamma_m\sqrt{m}\mu_{\max}\|r\|_2 n^{-\frac{1}{2}}.$$

This completes the proof. ■

Lemma 14 (**S_{12} Bound**) *Let W be the weight matrix (e.g., at step t) with unit norm rows. Let $S_{12} = \gamma_m\zeta(\frac{z^T r}{n})q(a \circ \mu)^T$. Suppose $z, q, \zeta = n^\nu$ are from Assumption 2, a has fixed ± 1 entries (Assumption 3), $\mu = \mathbb{E}_x[\sigma'(Wx)]$ satisfies $\mu_k = \Theta(1)$ (Assumption 4), and the current residual r satisfies $|\frac{z^T r}{n}| = \Theta(\|r\|_2 n^{-\beta/2-1/2})$ (Assumption 6). Assume $d > n$. Then w.h.p.:*

$$\|S_{12}\|_2 = \Theta\left(\sqrt{m}\gamma_m\|r\|_2 n^{\nu-\frac{\beta}{2}-\frac{1}{2}}\right).$$

Proof Since S_{12} is a rank-1 matrix and $\|q\|_2 = 1$, its operator norm is:

$$\|S_{12}\|_2 = \left|\gamma_m\zeta\left(\frac{z^T r}{n}\right)\right| \|q\|_2\|a \circ \mu\|_2 = \gamma_m n^\nu \left|\frac{z^T r}{n}\right| \|a \circ \mu\|_2.$$

By Assumption 6 applied to the current residual r, we have

$$\left|\frac{z^T r}{n}\right| = \Theta\left(\|r\|_2 n^{-\frac{\beta}{2}-\frac{1}{2}}\right).$$

Substituting this scaling, we get

$$\|S_{12}\|_2 = \gamma_m n^\nu \Theta\left(\|r\|_2 n^{-\frac{\beta}{2}-\frac{1}{2}}\right)\|a \circ \mu\|_2 = \Theta\left(\gamma_m n^{\nu-\frac{\beta+1}{2}}\|r\|_2\|a \circ \mu\|_2\right).$$

As established in the proof of Theorem 13, using the assumptions on a and μ (specifically $\mu_k = \Theta(1)$), we have $\|a \circ \mu\|_2 = \Theta(\sqrt{m})$. Combining these gives the final result:

$$\|S_{12}\|_2 = \Theta\left(\gamma_m n^{\nu-\frac{\beta+1}{2}}\|r\|_2\Theta(\sqrt{m})\right) = \Theta\left(\sqrt{m}\gamma_m\|r\|_2 n^{\nu-\frac{\beta}{2}-\frac{1}{2}}\right).$$

■

Lemma 15 (S_2 Bound) *Let W be the weight matrix (e.g., at step t) with unit norm rows. Let $S_2 = \frac{\gamma_m \zeta}{n} q z^T \left[(ra^T) \circ \sigma'_\perp(XW^T)\right]$. Suppose $z, q, \zeta = n^\nu$ are from Assumption 2, a has fixed ± 1 entries (Assumption 3), $\mu = \mathbb{E}_x[\sigma'(Wx)]$ satisfies $\mu_k = \Theta(1)$ (Assumption 4), and the current residual r satisfies $|\frac{z^T r}{n}| = \Theta(\|r\|_2 n^{-\beta/2-1/2})$ (Assumption 6).Then, w.h.p.:*

$$\gamma_m n^{\nu-\frac{\beta}{2}-1}\|r\|_2 \sigma_{\min}(\sigma'_\perp(XW^T)) \lesssim \|S_2\|_2 \lesssim \gamma_m \sqrt{m}\, \|r\|_\infty \, \min(n^\nu, \|W\|_2 n^{2\nu-\frac{1}{2}}).$$

Where $\lesssim$ hides universal constants C_1, C_2.

Proof The operator norm is

$$\|S_2\|_2 = \frac{\gamma_m n^\nu}{n} \|z^T((ra^T) \circ \sigma'_\perp(XW^T))\|_2.$$

Upper Bound: Using Theorem 20 and Assumption 3 that $a_i \sim \mathrm{Unif}(\pm 1)$, we have the upper bound

$$\begin{aligned}\|z^T((ra^T) \circ \sigma'_\perp(XW^T))\|_2 &\lesssim \|z\|_2 \|r\|_\infty \|a\|_\infty \|\sigma'_\perp(XW^T))\|_2 \\ &\lesssim \|z\|_2 \|r\|_\infty \|\sigma'_\perp(XW^T))\|_2.\end{aligned}$$

Then with probability $1 - o(1)$, since $z \sim \mathcal{N}(0, I)$, we have that $\|z\|_2 \lesssim C\sqrt{n}$. Hence we get that

$$\|z^T((ra^T) \circ \sigma'_\perp(XW^T))\|_2 \lesssim C\|r\|_\infty \|\sigma'_\perp(XW^T))\|_2 \sqrt{n}.$$

Then since we have Assumption 4, we can use Theorem 19 to bound the norm $\|\sigma'_\perp(XW^T)\|_2$, which gives us that with probability $1 - o(1)$,

$$\begin{aligned}\|z^T((ra^T) \circ \sigma'_\perp(XW^T))\|_2 &\lesssim C\|r\|_\infty \sqrt{n} \min\left(n, \sqrt{n}\|W\Sigma^{1/2}\|_2\right) \\ &= C\|r\|_\infty \sqrt{n} \min(n, \|W\|_2 n^{\nu+1/2}).\end{aligned}$$

Thus, we get that

$$\begin{aligned}\|S_2\|_2 &\lesssim \frac{\gamma_m}{n} n^\nu C \|r\|_\infty \sqrt{n} \min(n, \|W\|_2 n^{\nu+1/2}) \\ &= C\sqrt{m}\, \gamma_m \, \|r\|_\infty \, \min(n^\nu, \|W\|_2 n^{2\nu-\frac{1}{2}}),\end{aligned}$$

where we used the proportional scaling of n and m, Assumption 1, in the second line.

Lower Bound: For a lower bound, we start by writing

$$(ra^T) \circ \sigma'_\perp(XW^T) = \mathrm{Diag}(r)\, \sigma'_\perp(XW^T)\, \mathrm{Diag}(a).$$

Thus, we have that

$$\begin{aligned}q z^T \left((ra^T) \circ \sigma'_\perp(XW^T)\right) &= q \left(z^T \mathrm{Diag}(r)\right) \sigma'_\perp(XW^T)\, \mathrm{Diag}(a) \\ &= q(z \circ r)^T \sigma'_\perp(XW^T)\, \mathrm{Diag}(a).\end{aligned}$$

Taking the norm and recalling that $\zeta = n^\nu$, we get

$$\|\zeta q z^T \left((ra^T) \circ \sigma'_\perp(XW^T)\right) \|_2 = n^\nu \|q\| \left\|(z \circ r)^T \sigma'_\perp(XW^T) \operatorname{Diag}(a)\right\|.$$

Since the entries of a are ± 1 and q has unit norm, we have that this is the same as

$$n^\nu \|q\| \left\|(z \circ r)^T \sigma'_\perp(XW^T)\right\| = n^\nu \left\|(z \circ r)^T \sigma'_\perp(XW^T)\right\|.$$

By Cauchy-Schwarz, we have using Assumption 6 $|z^T r/\sqrt{n}\|r\|_2| = \Theta(d^{-\beta/2})$ that

$$\|z \circ r\| = \sqrt{\sum_{i=1}^n (z_i r_i)^2} \geq \frac{|\sum_{i=1}^n z_i r_i|}{\sqrt{\sum_{i=1}^n 1}} = \frac{|z^T r| \|r\|_2}{\sqrt{n}\|r\|_2} = \Omega(n^{-\frac{\beta}{2}} \|r\|_2).$$

Thus, we get that for some constant C

$$\|S_2\| \gtrsim C\gamma_m \frac{1}{n} n^{\nu - \frac{\beta}{2}} \|r\|_2 \sigma_{\min}(\sigma'_\perp(XW^T)).$$

■

Lemma 16 (Upper Bound on E) *Assuming Assumption 1], Assumption 3, Assumption 2, and Assumption 4, we have that with probability at least* $1 - o(1)$

$$\|E\|_2 \lesssim C\sqrt{m}\gamma_m \|r\|_\infty \min\left(1, n^{\nu - \frac{1}{2}} \|W\|_2\right).$$

Proof Recall $E = \frac{\gamma_m}{n} X_B^T((ra^T) \circ \sigma'_\perp(XW^T))$. Using Theorem 20, we have that

$$\frac{n}{\gamma_m} \|E\|_2 \lesssim \|X_B\|_2 \|r\|_\infty \|a\|_\infty \|\sigma'_\perp(XW^T)\|_2.$$

Then using Assumption 2, whereby the rows of X_B are iid from $\mathcal{N}(0, \hat{\Sigma})$, we have with probability $1 - o(1)$ that

$$\|X_B\|_2 \lesssim C\sqrt{n},$$

and using Assumption 3, we trivially have that

$$\|a\|_\infty = 1.$$

Thus, we have that

$$\frac{n}{\gamma_m} \|E\|_2 \lesssim C\sqrt{n} \|r\|_\infty \|\sigma'_\perp(XW^T)\|_2.$$

Then using Theorem 19, we have that with probability $1 - o(1)$

$$\|\sigma'_\perp(XW^T)\|_2 \lesssim C \min\left(n, \sqrt{n}\|W\Sigma^{1/2}\|_2\right).$$

Since $\|\Sigma^{1/2}\| = n^\nu$, we get the result in the proportional scaling of Assumption 1. ■

Theorem 1 (Gradient approximation) *Suppose Assumptions 1, 2, 3, 4, 5, 6 are satisfied, X and W are independent, and σ is a $\mathcal{C}^2$ function. Define $E = G - S_1 - S_{12} - S_2$. Then, for all $\nu, \alpha \in \mathbb{R}_{\geq 0}$,*

$$\frac{\|G - S_1 - S_{12}\|_2}{\sqrt{m}\gamma_m \|r\|_\infty} = O\left(\|W\|_2 n^{2\nu - \frac{1}{2}}\right), \quad \frac{\|G - S_1 - S_{12} - S_2\|_2}{\sqrt{m}\gamma_m \|r\|_\infty} = O\left(\|W\|_2 n^{\nu - \frac{1}{2}}\right) \tag{3}$$

with probability $1 - o(1)$ as $d, n, m \to \infty$. Moreover, if $\nu < \frac{1}{2}$ then with the same probability

$$\frac{\|S_1\|_2}{\|E\|_2} = \Omega\left(\frac{n^{\frac{1}{2} - \nu - \frac{\alpha}{2}}}{\log n \|W\|_2}\right), \qquad \frac{\|S_2\|_2}{\|E\|_2} = \Omega\left(\frac{n^\nu}{\log n} \frac{\|(z \circ r)^T \sigma'_\perp(XW^T)\|_2}{\|\sigma'_\perp(XW^T)\|_2}\right), \tag{4}$$

$$\frac{\|S_{12}\|_2}{\|E\|_2} = \Omega\left(\frac{n^{\frac{1}{2} - \frac{\beta}{2}}}{\log n \|W\|_2}\right), \qquad \Omega(n^{\nu - \frac{\beta}{2}}) \leq \frac{\|S_{12}\|_2}{\|S_1\|_2} \leq O(n^{\nu - \frac{\beta}{2} + \frac{\alpha}{2}}). \tag{5}$$

Proof We start with the gradient decomposition derived in Section 3:

$$G = S_1 + S_{12} + S_2 + E$$

where

$$\begin{aligned}
S_1 &= \gamma_m \frac{X_B^T r}{n} (a \circ \mu)^T \\
S_{12} &= \gamma_m \zeta \left(\frac{z^T r}{n}\right) q (a \circ \mu)^T \\
S_2 &= \frac{\gamma_m \zeta}{n} q \left(z^T ((ra^T) \circ \sigma'_\perp(XW^T))\right) \\
E &= \frac{\gamma_m}{n} X_B^T ((ra^T) \circ \sigma'_\perp(XW^T)).
\end{aligned}$$

We assume the conditions of the theorem hold, including the scaling $\sqrt{m}\gamma_m = O(1)$ and the residual concentration $\|r\|_2 / \|r\|_\infty = \Theta(\sqrt{n}/\log n)$ (Assumption 5).

Proof of Upper Bounds:

For the first upper bound, we have $G - S_1 - S_{12} - S_2 = E$. Using the upper bound on $\|E\|_2$ from Theorem 16 and the assumption $\sqrt{m}\gamma_m = O(1)$:

$$\begin{aligned}
\frac{\|G - S_1 - S_{12} - S_2\|_2}{\|r\|_\infty} &= \frac{\|E\|_2}{\|r\|_\infty} \\
&\lesssim \frac{C\sqrt{m}\gamma_m \min\left(1, n^{\nu - \frac{1}{2}} \|W\|_2\right)}{\|r\|_\infty} \\
&= O\left(\min(1, \|W\|_2 n^{\nu - \frac{1}{2}})\right).
\end{aligned}$$

For the second upper bound, we have $G - S_1 - S_{12} = S_2 + E$. Using the triangle inequality and the upper bounds on $\|S_2\|_2$ from Theorem 15 and $\|E\|_2$ from Theorem 16, along with $\sqrt{m}\gamma_m = O(1)$:

$$\begin{aligned}
\frac{\|G - S_1 - S_{12}\|_2}{\|r\|_\infty} &\leq \frac{\|S_2\|_2 + \|E\|_2}{\|r\|_\infty} \\
&\lesssim \frac{\sqrt{m}\gamma_m \|r\|_\infty \min(n^\nu, \|W\|_2 n^{2\nu - \frac{1}{2}}) + \sqrt{m}\gamma_m \|r\|_\infty \min\left(1, n^{\nu - \frac{1}{2}} \|W\|_2\right)}{\|r\|_\infty} \\
&= O\left(\min(n^\nu, \|W\|_2 n^{2\nu - \frac{1}{2}}) + \min(1, \|W\|_2 n^{\nu - \frac{1}{2}})\right).
\end{aligned}$$

Proof of Lower Bounds:

We establish lower bounds for the ratios $\|S_1\|/\|E\|$, $\|S_{12}\|/\|E\|$, and $\|S_2\|/\|E\|$. These rely on the lower bounds for $\|S_1\|, \|S_{12}\|, \|S_2\|$ and the upper bound for $\|E\|$. We use the result $\|r\|_2/\|r\|_\infty = \Theta(\sqrt{n}/\log n)$.

Ratio $\|S_1\|/\|E\|$*:* Using Theorem 13 (lower bound) and Theorem 16 (upper bound), we have that

$$\begin{aligned}
\frac{\|S_1\|_2}{\|E\|_2} &\gtrsim \frac{\sqrt{m}\gamma_m \mu_{\min} \|r\|_2 n^{-\frac{\alpha+1}{2}}}{\sqrt{m}\gamma_m \|r\|_\infty \min\left(1, n^{\nu-\frac{1}{2}}\|W\|_2\right)} \\
&\gtrsim \frac{\|r\|_2}{\|r\|_\infty} \frac{n^{-(\alpha+1)/2}}{\min(1, \|W\|_2 n^{\nu-1/2})} \\
&= \frac{\sqrt{n}}{\log n} \frac{n^{-(\alpha+1)/2}}{\min(1, \|W\|_2 n^{\nu-1/2})} \\
&= \frac{n^{-\alpha/2}}{\log n \min(1, \|W\|_2 n^{\nu-1/2})}.
\end{aligned}$$

If $\nu < 1/2$ and we assume $\|W\|_2 n^{\nu-1/2} = O(1)$ is the dominant term in the minimum, the ratio is

$$\Omega\left(\frac{n^{1/2-\nu-\alpha/2}}{\log n \|W\|_2}\right).$$

If $\nu \geq 1/2$ and assume $\|W\|_2 n^{\nu-1/2} \geq \Omega(1)$, the minimum is $O(1)$. The ratio is

$$\Omega\left(\frac{n^{-\alpha/2}}{\log n}\right).$$

Ratio $\|S_{12}\|/\|E\|$*:* Using Theorem 14 (lower bound) and Theorem 16 (upper bound):

$$\begin{aligned}
\frac{\|S_{12}\|_2}{\|E\|_2} &\gtrsim \frac{\sqrt{m}\gamma_m \|r\|_2 n^{\nu-\beta/2-1/2}}{\sqrt{m}\gamma_m \|r\|_\infty \min(1, \|W\|_2 n^{\nu-1/2})} \\
&\gtrsim \frac{\|r\|_2}{\|r\|_\infty} \frac{n^{\nu-\beta/2-1/2}}{\min(1, \|W\|_2 n^{\nu-1/2})} \\
&= \frac{\sqrt{n}}{\log n} \frac{n^{\nu-\beta/2-1/2}}{\min(1, \|W\|_2 n^{\nu-1/2})} \\
&= \frac{n^{\nu-\beta/2}}{\log n \min(1, \|W\|_2 n^{\nu-1/2})}.
\end{aligned}$$

If $\nu < 1/2$ and assume $\|W\|_2 n^{\nu-1/2} = O(1)$ dominates the minimum, the ratio is

$$\Omega\left(\frac{n^{1/2-\beta/2}}{\log n \|W\|_2}\right).$$

If $\nu \geq 1/2$ and assume $\|W\|_2 n^{\nu-1/2} \geq \Omega(1)$, the minimum is $O(1)$. The ratio is

$$\Omega\left(\frac{n^{\nu-\beta/2}}{\log n}\right).$$

Ratio $\|S_2\|/\|E\|$*:* We have that

$$\begin{aligned}\frac{\|S_2\|}{\|E\|} &\gtrsim \frac{\frac{\gamma_m}{n} n^{\nu} \|(z \circ r)^T \sigma'_\perp(XW^T)\|}{\frac{\gamma_m}{n}\|X_B\|_2 \|r\|_\infty \|\sigma'_\perp(XW^T)\|} \\ &\gtrsim n^{\nu-\frac{1}{2}} \frac{\|z \circ r\|}{\|r\|_\infty} \kappa\left(\sigma'_\perp(XW^T)\right) \\ &\gtrsim n^{\nu-\frac{1}{2}-\frac{\beta}{2}} \frac{\|r\|_2}{\|r\|_\infty} \kappa\left(\sigma'_\perp(XW^T)\right) \\ &\gtrsim \frac{n^{\nu-\frac{\beta}{2}}}{\log n} \kappa\left(\sigma'_\perp(XW^T)\right)\end{aligned}$$

Relative Sizes Next, we prove the relative bounds. First, we have that

$$\frac{\|S_{12}\|}{\|S_1\|} = \frac{\|X_S^T r\|\|a \circ \mu\|}{\|X_B^T r\|\|a \circ \mu\|} = \frac{n^{\nu+\frac{1}{2}-\frac{\beta}{2}}\|r\|_2}{\|X_B^T r\|}$$

Then since

$$n^{-\frac{\alpha}{2}+\frac{1}{2}}\|r\|_2 \lesssim \|X_B^T r\|_2 \lesssim \sqrt{n}\|r\|_2,$$

we get that

$$n^{\nu-\frac{\beta}{2}} \lesssim \frac{\|S_{12}\|}{\|S_1\|} \lesssim n^{\nu-\frac{\beta}{2}+\frac{\alpha}{2}}$$

For the second relative bound, we have that

$$\frac{\|S_{12}\|}{\|S_2\|} = \frac{n^{\nu+\frac{1}{2}-\frac{\beta}{2}}\|r\|_2\|a \circ \mu\|}{n^{\nu}\|(z \circ r)^T \sigma'_\perp(XW^T)\|} = \Theta\left(\frac{n^{1-\frac{\beta}{2}}\|r\|_2}{\|(z \circ r)^T \sigma'_\perp(XW^T)\|}\right)$$

For a lower bound, we get that

$$\frac{\|S_{12}\|}{\|S_2\|} \gtrsim C \frac{\|r\|_2}{\|z\|_2\|r\|_2} \frac{n^{1-\frac{\beta}{2}}}{n^{\nu+\frac{1}{2}}} = \frac{1}{n^{\nu+\frac{\beta}{2}}}$$

For an upper bound, we have that

$$\frac{\|S_{12}\|}{\|S_2\|} \lesssim \frac{n^{1-\frac{\beta}{2}}\|r\|_2}{n^{-\frac{\beta}{2}}\|r\|_2 \sigma_{\min}(\sigma'_\perp(XW^T))} = \frac{n}{\sigma_{\min}(\sigma'_\perp(XW^T))}$$

■

Theorem 3 (Large data-spike gradient approximation) *Suppose Assumptions 1, 2, 3, 4, 5, and 6 are satisfied, and define* $E_L = G - S_{12} - S_2$*. Then, with probability* $1 - o(1)$ *for* $\nu \geq \frac{1}{2}$ *we have*

$$\frac{\|E_L\|_2}{\sqrt{m}\gamma_m\|r\|_\infty} = O(1), \frac{\|S_{12}\|_2}{\|E_L\|_2} = \Omega\left(\frac{n^{\nu-\frac{\beta}{2}}}{\log n}\right), \frac{\|S_2\|_2}{\|E_L\|_2} = \Omega\left(\frac{n^{\nu}}{\log n}\frac{\|(z \circ r)^T \sigma'_\perp(XW^T)\|_2}{\|\sigma'_\perp(XW^T)\|_2}\right). \tag{6}$$

Proof This proof is exactly the same as Theorem 1. Except we use the following upper bounds.

Data Spike: The operator norm is

$$\|S_2\|_2 = \frac{\gamma_m n^\nu}{n} \|z^T((ra^T) \circ \sigma'_\perp(XW^T))\|_2.$$

Using Theorem 20 and Assumption 3 that $a_i \sim \text{Unif}(\pm 1)$, we have the upper bound

$$\begin{aligned}\|z^T((ra^T) \circ \sigma'_\perp(XW^T))\|_2 &\lesssim \|z\|_2 \|r\|_\infty \|a\|_\infty \|\sigma'_\perp(XW^T))\|_2 \\ &\lesssim \|z\|_2 \|r\|_\infty \|\sigma'_\perp(XW^T))\|_2.\end{aligned}$$

Then with probability $1 - o(1)$, since $z \sim \mathcal{N}(0, I)$, we have that $\|z\|_2 \lesssim C\sqrt{n}$. Hence we get that

$$\|z^T((ra^T) \circ \sigma'_\perp(XW^T))\|_2 \lesssim C\|r\|_\infty \|\sigma'_\perp(XW^T))\|_2 \sqrt{n}.$$

Then since we have Assumption 4, we can bound the norm $\|\sigma'_\perp(XW^T)\|_2$ by $O(n)$

$$\|z^T((ra^T) \circ \sigma'_\perp(XW^T))\|_2 \lesssim C\|r\|_\infty \sqrt{n} n$$

Thus, we get that

$$\|S_2\|_2 \lesssim C\sqrt{m}\, \gamma_m \, \|r\|_\infty \, n^\nu,$$

where we used the proportional scaling of n and m, Assumption 1, in the second line.

Error Term: Recall $E = \frac{\gamma_m}{n} X_B^T((ra^T) \circ \sigma'_\perp(XW^T))$. Using Theorem 20, we have that

$$\frac{n}{\gamma_m} \|E\|_2 \lesssim \|X_B\|_2 \|r\|_\infty \|a\|_\infty \|\sigma'_\perp(XW^T)\|_2.$$

Then using Assumption 2, whereby the rows of X_B are iid from $\mathcal{N}(0, \hat{\Sigma})$, we have with probability $1 - o(1)$ that

$$\|X_B\|_2 \lesssim C\sqrt{n},$$

and using Assumption 3, we trivially have that

$$\|a\|_\infty = 1.$$

Thus, we have that

$$\frac{n}{\gamma_m} \|E\|_2 \lesssim C\sqrt{n} \|r\|_\infty \|\sigma'_\perp(XW^T)\|_2.$$

Then

$$\|\sigma'_\perp(XW^T)\|_2 \leq O(n).$$

Since $\|\Sigma^{1/2}\| = n^\nu$, we get the result in the proportional scaling of Assumption 1. ■

F.2.2. HELPER RESULTS: SUBGAUSSIANITY AND CONCENTRATION

Lemma 17 *Let $Z \in \mathbb{R}^{n\times d}$ be a matrix with standard normal IID entries. If $n < d$, then as $n/d \to c \in (0,1)$, we have that with probability 1, the eigenvalues of $\frac{1}{d}ZZ^T$ are $\Theta(1)$. Further,*

$$\sigma_{\min}(Z) = \Theta(\sqrt{d} - \sqrt{n}), \quad \sigma_{\max}(Z) = \Theta(\sqrt{d} + \sqrt{n}).$$

Proof As $\frac{1}{d}ZZ^T$ is a Wishart matrix, the limiting empirical spectral distribution almost surely weakly converges to the Marchenko-Pastur distribution supported on $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$. ■

Lemma 18 *Let $X_B \in \mathbb{R}^{n\times d}$ have IID rows from $\mathcal{N}(0, \hat{\Sigma})$, where $\lambda_k(\hat{\Sigma}) \sim k^{-\alpha}$ as per Assumption 2. Then with probability $1 - 2\exp(-cn)$ for positive universal constants c, we have that*

$$\Omega\left(n^{\frac{1-\alpha}{2}}\right) \le \|X_B\|_2 \le O\left(n^{\frac{1}{2}}\right)$$

Proof We can write $X_B = \hat{\Sigma}^{1/2}Z$ where $Z \in \mathbb{R}^{n\times d}$ has IID standard normal entries. Using Theorem 17, we have that in the proportional regime (Assumption 1), $\|Z\|_2 = \Theta(\sqrt{n})$. The result follows using the fact that

$$\sigma_{\min}(\hat{\Sigma}^{1/2})\|Z\|_2 \le \|X_B\|_2 = \|\hat{\Sigma}^{1/2}Z\|_2 \le \sigma_{\max}(\hat{\Sigma}^{1/2})\|Z\|_2,$$

and noting that

$$\sigma_{\min}(\Sigma^{1/2}) = \Theta(n^{-\alpha/2}) \text{ and } \sigma_{\max}(\Sigma^{1/2}) = \Theta(1).$$

■

Lemma 19 *Let W be a given fixed matrix indepedent of X. If Assumption 4 is satisfied and σ is $\mathcal{C}^2$, then we have with probability $1 - C\exp(-cn)$ for positive universal constants c, C, that*

$$\|\sigma'_\perp(XW^T)\|_2 \lesssim C' \min\left(n, \sqrt{n}\|W\Sigma^{1/2}\|_2\right).$$

for some constant $C' > 0$. Here $\Sigma = \hat{\Sigma} + \zeta^2 qq^T$ is the full data covariance from Assumption 2.

Proof Since σ is L-Lipschitz (Assumption 4), its derivative σ' is bounded by L. As $\mu = \mathbb{E}_x[\sigma'(Wx)]$, the centered term $\sigma'_\perp(XW^T) = \sigma'(XW^T) - \mathbf{1}_n\mu^T$ has entries bounded by some M (e.g., $M = 2L$). Thus, using the relation between operator and Frobenius norms:

$$\|\sigma'_\perp(XW^T)\|_2^2 \le \|\sigma'_\perp(XW^T)\|_F^2 \le Mnm.$$

Thus, we have that in the proportional regime

$$\|\sigma'_\perp(XW^T)\|_2 = O(n).$$

On the other hand, $\sigma'_\perp(XW^T)$ represents mean-centered features and is Lipschitz, using Corollary 25, with probability $1 - C\exp(-cn)$, we have that

$$\|\sigma'_\perp(XW^T)\|_2 = O\left(\sqrt{n}\|W\Sigma^{1/2}\|_2\right).$$

The overall bound follows by taking the minimum of the two derived bounds. ■

Lemma 20 *For any vectors u, v and matrix A, we have that*

$$\min_i |u_i| \min_j |v_j| \|A\|_2 \leq \|(uv^T) \circ A\|_2 \leq \|u\|_\infty \|v\|_\infty \|A\|_2.$$

Proof This follows from the observation that

$$(uv^T) \circ A = \text{diag}(u) A \text{diag}(v),$$

where $\text{diag}(u)$ is the diagonal matrix with u in the diagonal. Then using the fact that

$$\sigma_{\min}(B)\|A\|_2 \leq \|AB\|_2 \leq \sigma_{\max}(B)\|A\|_2,$$

where $\sigma_{\min}$ is allowed to be zero and noticing that

$$\sigma_{\max}(\text{diag}(u)) = \|u\|_\infty \quad \text{and} \quad \sigma_{\min}(\text{diag}(u)) = \min_i |u_i|.$$

The bounds follow from applying the matrix norm inequality twice. ■

Lemma 21 (Sub-Gaussianity) *For $x \sim \mathcal{N}(0, \Sigma)$, a fixed vector $w \in \mathbb{R}^d$, and an $\mathcal{L}_f$-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$, the random variable $f(w^T x)$ is sub-gaussian with subgaussian norm at most $C\mathcal{L}_f^2 \|w^T \Sigma^{1/2}\|_2^2$ for some constant C. Furthermore,*

$$\mathbb{E}[|f(w^T x)|] = |f(0)| + O\left(\mathcal{L}_f \|w^T \Sigma^{1/2}\|_2\right) = O(1 + \mathcal{L}_f \|w^T \Sigma^{1/2}\|_2).$$

Proof Using Lipschitzness,

$$\left|f(x^T w) - f(0^T w)\right| \leq \mathcal{L}_f |x^T w - 0| = \mathcal{L}_f |x^T w|.$$

The variable $w^T x \sim \mathcal{N}(0, \sigma_w^2)$ where $\sigma_w^2 = \|w^T \Sigma^{1/2}\|_2^2$. Thus, $w^T x$ is (σ_w^2)-sub-gaussian. For $t \geq 0$,

$$\Pr\left[|f(x^T w) - f(0)| \geq t\right] \leq \Pr\left[\mathcal{L}_f |x^T w| \geq t\right] \leq 2 \exp\left(-\frac{t^2}{2\mathcal{L}_f^2 \|w^T \Sigma^{1/2}\|_2^2}\right).$$

Thus, we see that

$$\Pr[|f(x^T w)| \geq t] \leq 2 \exp\left(-\frac{(t-c)^2}{2\mathcal{L}_f^2 \|w^T \Sigma^{1/2}\|_2^2}\right),$$

where $c = |f(0)|$. For the expectations, taking expectations, we get that

$$\mathbb{E}\left[|f(x^T w) - f(0)|\right] \leq \mathbb{E}\left[\mathcal{L}_f |x^T w|\right] = \mathcal{L}_f \sqrt{\frac{2}{\pi}} \|w^T \Sigma^{1/2}\|_2^2.$$

Using $|f(w^T x)| \leq |f(w^T x) - f(0)| + |f(0)|$ and the triangle inequality for expectations, $\mathbb{E}[|f(w^T x)|] \leq \mathbb{E}[|f(w^T x) - f(0)|] + |f(0)| = |f(0)| + O(\mathcal{L}_f \sigma_w)$, giving the result. ■

Lemma 22 (Covariance Operator Norm Bound) *Let $W \in \mathbb{R}^{m\times d}$ be a fixed matrix whose rows have unit norm and let $x \sim \mathcal{N}(0,\Sigma)$. Suppose that $f:\mathbb{R}\to\mathbb{R}$ is $\mathcal{L}_f$ Lipschitz respectively. Define the population second moment matrix*

$$\Phi = \mathbb{E}_x[f(Wx)f(Wx)^T],$$

where f is applied element-wise to the vector $Wx \in \mathbb{R}^m$. Then

$$\|\Phi\|_2 \leq \|\mathbb{E}_x\left[f(Wx)\right]\|_2^2 + \|W\Sigma^{1/2}\|_2^2\mathcal{L}_f^2$$

for some universal constants C_1, C_2.

Proof We note that Φ is the uncentered covariance matrix. However, to bound the operator norm of Φ we need to consider the centered covariance matrix $\check{\Phi}$

$$\check{\Phi} = \underbrace{\mathbb{E}\left[f(Wx)f(Wx)^T\right]}_{\Phi} - \mathbb{E}\left[f(Wx)\right]\mathbb{E}\left[f(Wx)\right]^T$$

Then we see that

$$\begin{aligned}
\left\|\check{\Phi}\right\|_2 &= \sup_{\|v\|=1} v^T\check{\Phi}v \\
&= \sup_{\|v\|=1} v^T\Phi v - \left(\mathbb{E}\left[v^Tf(Wx)\right]\right)^2 \\
&= \sup_{\|v\|=1} \mathbb{E}\left[\left(v^Tf(Wx)\right)\left(v^Tf(Wx)\right)^T\right] - \left(\mathbb{E}\left[v^Tf(Wx)\right]\right)^2 \\
&= \sup_{\|v\|=1} \mathrm{Var}\left(v^Tf(Wx)\right)
\end{aligned}$$

We want to bound this using the Gaussian Poincare inequality. Which we recall here (Link). Let $g:\mathbb{R}^d\to\mathbb{R}$ be a C^1 function then

$$\mathrm{Var}_{z\sim\mathcal{N}(0,I)}(g(z)) \leq \mathbb{E}_{z\sim\mathcal{N}(0,I)}\left[\|\nabla g(z)\|^2\right]$$

Since $x\sim\mathcal{N}(0,\Sigma)$, we can write it as $x=\Sigma^{1/2}z$. Thus, define the function

$$g(z) := f(Wx) = v^Tf\left(W\Sigma^{1/2}x\right) = \sum_{k=1}^m v_kf\left(w_k^T\Sigma^{1/2}x\right).$$

Let us then define

$$u = \begin{bmatrix} v_1f'\left(w_1^T\Sigma^{1/2}x\right) & \dots & v_mf'\left(w_m^T\Sigma^{1/2}x\right)\end{bmatrix}^T$$

Then we see that

$$\nabla g(z)^T = \sum_{k=1}^m v_kf'\left(w_k^T\Sigma^{1/2}x\right)\left(w_k^T\Sigma^{1/2}\right) = u^TW\Sigma^{1/2}$$

Thus, we see that

$$\mathbb{E}_z\left[\|\nabla_z g(z)\|^2\right] \leq \mathbb{E}_x\left[\|W\Sigma^{1/2}\|_2^2\|u\|^2\right] \leq \|W\Sigma^{1/2}\|_2^2\mathbb{E}_x\left[\|u\|^2\right]$$

Then using Lemma 21 and noting that f' is bounded by $\mathcal{L}_f$, we get that

$$\begin{aligned}\mathbb{E}_x\left[\sum_{k=1}^m u_k^2\right] &= \sum_{k=1}^m v_k^2 \mathbb{E}_x\left[\left(f'(w_k^T x)\right)^2\right] \\ &\leq \sum_{k=1}^m v_k^2 \mathcal{L}_f^2 \qquad\qquad \leq \mathcal{L}_f^2\end{aligned}$$

Thus, we have that

$$\mathbb{E}\left[\|\nabla g(z)\|^2\right] \leq \|W\Sigma^{1/2}\|_2^2 \mathcal{L}_f^2$$

Thus, using the Gaussian Poincare inequality, we see that

$$\left\|\check{\Phi}\right\|_2 \leq \|W\Sigma^{1/2}\|_2^2 \mathcal{L}_f^2$$

Thus, we see that

$$\|\Phi\|_2 \leq \|\check{\Phi} - \Phi\|_2 + \|W\Sigma^{1/2}\|_2^2 \mathcal{L}_f^2$$

Finally, we see that

$$\begin{aligned}\left\|\check{\Phi} - \Phi\right\|_2 &= \left\|\mathbb{E}\left[f(Wx)\right]\mathbb{E}\left[f(Wx)\right]^T\right\|_2 \\ &= \left\|\mathbb{E}\left[f(Wx)\right]\right\|_2^2\end{aligned}$$

Thus,

$$\|\Phi\|_2 \leq \left\|\mathbb{E}\left[f(Wx)\right]\right\|_2^2 + \|W\Sigma^{1/2}\|_2^2 \cdot \mathcal{L}_f^2$$

■

We are going to instantiate a few corollaries for cases that we care about. Specifically, we shall $f = \sigma'_\perp$ as the non-linearity. In this case we have that $\mathbb{E}\left[f(Wx)\right] = 0$.

Corollary 23 *If $\mathbb{E}\left[f(Wx)\right] = 0$, we have that*

$$\|\Phi\|_2 \leq \|W\Sigma^{1/2}\|_2^2 \mathcal{L}_f^2.$$

We shall also need to bound the norm of the expectation. In the case, when σ is bounded, we get that the expectation

Lemma 24 (Feature Norm Bound) *Let $x_i \sim \mathcal{N}(0, \Sigma)$ be IID for $i = 1 \ldots n$, forming rows of X. Let $W \in \mathbb{R}^{m \times d}$ be a fixed matrix whose rows w_j have norm $\|w_j\|_2 = 1$. Let $f : \mathbb{R} \to \mathbb{R}$ be $\mathcal{L}_f$-Lipschitz. Define the population second moment matrix*

$$\Phi = \mathbb{E}_x[f(Wx)f(Wx)^T]$$

(as in Theorem 22). Then with probability $1 - 2e^{-cn}$ for some universal constant $c > 0$,

$$\left\|\frac{1}{\sqrt{n}} f(XW^T)\right\|_2 \leq \left(1 + C'\sqrt{\frac{m}{n}}\right)\sqrt{\|\Phi\|_2}$$

for some universal constant C'.

Proof Since x_i are IID, we have the rows of $f(XW^T) \in \mathbb{R}^{n\times m}$ are IID. Additionally, by Lemma 21 the entries are $\mathcal{L}_f^2 \|w_i^T \Sigma^{1/2}\|_2^2$ sub-gaussian entries. Thus, we have that

$$\check{X} = \frac{1}{\mathcal{L}_f \max_{i=1\dots m} \|w_i^T \Sigma^{1/2}\|_2} f(XW^T)$$

has IID rows whose sub-Gaussian norm is at most a universal constant. Let

$$\check{\Phi} = \frac{1}{n}\mathbb{E}\left[\check{X}^T\check{X}\right] = \frac{1}{\mathcal{L}_f^2 \max_{i=1\dots m} \|w_i^T \Sigma^{1/2}\|_2^2}\Phi$$

Then using Equation 5.26 from [28], there exists universal constant C, c such that

$$\Pr\left[\left\|\frac{1}{n}\check{X}^T\check{X} - \check{\Phi}\right\|_2 \geq \max(\delta, \delta^2)\|\check{\Phi}\|_2\right] < 2e^{-ct^2}, \quad \delta = C\sqrt{\frac{m}{n}} + \frac{t}{\sqrt{n}}$$

Thus, with probability $1 - 2e^{-ct^2}$, we have that

$$\left\|\frac{1}{n}\check{X}^T\check{X} - \check{\Phi}\right\|_2 \leq \max(\delta, \delta^2)\|\check{\Phi}\|_2$$

Using the reverse triangle inequality, we have that

$$\frac{1}{n}\|\check{X}^T\check{X}\|_2 \leq \left\|\frac{1}{n}\check{X}^T\check{X} - \check{\Phi}\right\|_2 + \|\check{\Phi}\|_2$$

Thus, with probability at least $1 - 2e^{-ct^2}$, we have that

$$\frac{1}{n}\|\check{X}^T\check{X}\|_2 \leq \|\check{\Phi}\|_2 + \max(\delta, \delta^2)\|\check{\Phi}\|_2$$

Thus, we get that

$$\frac{1}{\sqrt{n}}\|\check{X}\|_2 \leq \sqrt{\|\check{\Phi}\|_2 + \max(\delta, \delta^2)\|\check{\Phi}\|_2}$$

Multiplying both sides by $\mathcal{L}_f \max_{i=1\dots m} \|w_i^T \Sigma^{1/2}\|_2$, we see that

$$\begin{aligned}
\left\|\frac{1}{\sqrt{n}} f(W_0\tilde{X}^T)\right\|_2 &\leq \mathcal{L}_f \max_{i=1\dots m} \|w_i^T\Sigma^{1/2}\|_2(1 + C'\delta)\sqrt{\|\check{\Phi}\|_2} \\
&\leq (1 + C'\delta)\sqrt{\mathcal{L}_f^2 \max_{i=1\dots m} \|w_i^T\Sigma^{1/2}\|_2^2\|\check{\Phi}\|_2} \\
&\leq \left(1 + C'\delta\right)\sqrt{\|\Phi\|_2}
\end{aligned}$$

Using $t = \sqrt{m}$, we see that with probability $1 - 2e^{-cm}$,

$$\left\|\frac{1}{\sqrt{n}} f(W_0\tilde{X}^T)\right\|_2 \leq \left(1 + C'\sqrt{\frac{m}{n}}\right)\sqrt{\|\Phi\|_2}$$

■

Hence, we can again instantiate some simple corollaries.

Corollary 25 *If* $\mathbb{E}[f(Wx)] = 0$*, we have that*

$$\left\| f(W_0 \tilde{X}^T) \right\|_2 \leq \mathcal{L}_f C \|W\Sigma^{1/2}\|_2 \sqrt{n}$$

Another important case, if f is uniformly bounded. This is the case, when we apply it for σ', σ''. Here we either have the expectation is zero. In which Corollary 25 applies. If the mean in non-zero then we get the following.

Corollary 26 *If* $|\mathbb{E}[f(z)]| = M$*, we have that*

$$\left\| f(W_0 \tilde{X}^T) \right\|_2 \leq C \left[n + \mathcal{L}_f \|W\Sigma^{1/2}\|_2 \sqrt{n} \right].$$

F.3. ReLU Data Alignment

Lemma 27 *Let* $M = uv^T$ *be a non-zero rank 1 matrix, where* $u \in \mathbb{R}^m$ *and* $v \in \mathbb{R}^n$*. Assume that all entries of* u *and* v *are non-zero, i.e.,* $u_i \neq 0$ *for all* $i = 1, \ldots, m$ *and* $v_j \neq 0$ *for all* $j = 1, \ldots, n$*. Let* $\tilde{M}$ *be the matrix with entries* $\tilde{M}_{ij} = \delta_{u_i v_j > 0}$*. Let* $\hat{M} = \tilde{M} - 0.5J$*, where* J *is the* $m \times n$ *matrix of all ones. Then, rank*$(\hat{M}) = 1$*.*

Proof Let $u', u'' \in \{0,1\}^m$ and $v', v'' \in \{0,1\}^n$ be indicator vectors defined as follows:

- $u'_i = \delta_{u_i > 0}$
- $u''_i = \delta_{u_i < 0}$
- $v'_j = \delta_{v_j > 0}$
- $v''_j = \delta_{v_j < 0}$

Since we assume $u_i \neq 0$ and $v_j \neq 0$ for all i, j, every entry in u is either positive or negative, and similarly for v. This means $1_m = u' + u''$ and $1_n = v' + v''$, where 1 denotes a vector of all ones of the appropriate dimension.

The entry $\tilde{M}_{ij} = \delta_{u_i v_j > 0}$ is 1 if and only if ($u_i > 0$ and $v_j > 0$) or ($u_i < 0$ and $v_j < 0$). This can be written as:

$$\tilde{M} = u'(v')^T + u''(v'')^T$$

The all-ones matrix J can be written as $J = 1_m 1_n^T$. Using the property that $1 = u' + u''$ and $1 = v' + v''$:

$$\begin{aligned} J &= (u' + u'')(v' + v'')^T \\ &= u'(v')^T + u'(v'')^T + u''(v')^T + u''(v'')^T \end{aligned}$$

Now we compute $\hat{M} = \tilde{M} - 0.5J$:

$$\begin{aligned} \hat{M} &= (u'(v')^T + u''(v'')^T) - 0.5(u'(v')^T + u'(v'')^T + u''(v')^T + u''(v'')^T) \\ &= 0.5u'(v')^T + 0.5u''(v'')^T - 0.5u'(v'')^T - 0.5u''(v')^T \\ &= 0.5\left[u'(v')^T - u'(v'')^T - u''(v')^T + u''(v'')^T\right] \\ &= 0.5\left[u'((v')^T - (v'')^T) - u''((v')^T - (v'')^T)\right] \\ &= 0.5(u' - u'')((v')^T - (v'')^T) \\ &= 0.5(u' - u'')(v' - v'')^T \end{aligned}$$

Let $\text{sign}(u)$ denote the vector with entries $\text{sign}(u_i)$, where $\text{sign}(x) = 1$ if $x > 0$ and $\text{sign}(x) = -1$ if $x < 0$. Since no u_i is zero, $(u' - u'')_i = \delta_{u_i>0} - \delta_{u_i<0} = \text{sign}(u_i)$. Similarly, $(v' - v'')_j = \text{sign}(v_j)$. Thus, we have shown:

$$\hat{M} = 0.5 \cdot \text{sign}(u) \cdot \text{sign}(v)^T$$

Since $M = uv^T$ is non-zero, both u and v must be non-zero vectors. Because we assumed no zero entries, the vectors $\text{sign}(u)$ (containing only ± 1) and $\text{sign}(v)$ (containing only ± 1) are non-zero vectors. The matrix $\hat{M}$ is expressed as the outer product of two non-zero vectors. Therefore, $\text{rank}(\hat{M}) = 1$. ■

Proposition 28 (ReLU gradient) *If $2\nu > 1 - \alpha$, and the row of W are i.i.d. from the unit sphere, then with probability $1 - o(1)$ we have that $\sigma'_\perp(XW^T) = \frac{1}{2}\,\text{sign}(z_i)\,\text{sign}(Wq)^T$.*

Proof Recall the data decomposition $x_i = \zeta z_i q + x_{b,i}$, where the *spike direction* $q \in \mathbb{R}^d$ is unit-norm, $z_i \sim \mathcal{N}(0, 1)$, the bulk component $x_{b,i}$ has spectrum exponent α, and the spike magnitude scales as $\zeta = n^\nu$. Since each row w_k^T of W is uniform on $\mathbb{S}^{d-1}$, $\|Wq\|_2 \approx \sqrt{m/d}$ with high probability. Using standard concentration for random projections, with probability $1 - o(1)$,

$$\|Wx_{b,i}\|_2^2 \;\le\; C\,\|x_{b,i}\|_2^2 \;=\; C\sum_{j=1}^{d} j^{-\alpha} \;=\; \begin{cases} \Theta\big(d^{1-\alpha}\big) & \alpha < 1, \\ \Theta(\log d) & \alpha = 1, \\ O(1) & \alpha > 1. \end{cases} \tag{13}$$

For the spike term $\|\,W(\zeta z_i q)\|_2 \;=\; |z_i|\,\zeta\,\|Wq\|_2 \;\gtrsim\; n^\nu\sqrt{\frac{m}{d}}\,|z_i| \;\ge\; n^\nu$, since $|z_i| \ge c$ with probability $1 - o(1)$ for some universal $c > 0$. Hence, whenever $2\nu > 1 - \alpha$, the spike contribution $W(\zeta z_i q)$ dominates the bulk, so that $\text{sign}(Wx_i) = \text{sign}\big(W(\zeta z_i q)\big)$. Then Lemma 27 then implies for ReLU that

$$\sigma'_\perp(XW^T) = \tfrac{1}{2}\;\text{sign}(z_i)\;\text{sign}(Wq)^T.$$

■