Extended Abstract Track

Toward Understanding How the Data Affects Neural Collapse: A Kernel-Based Approach

Editors: List of editors' names

Abstract

Recently, a vast amount of literature has focused on the "Neural Collapse" (NC) phenomenon, which emerges when training neural network (NN) classifiers beyond the zero training error point. The core component of NC is the decrease in the within-class variability of the network's deepest features, dubbed as NC1. The theoretical works that study NC are typically based on simplified unconstrained features models (UFMs) that mask any effect of the data on the extent of collapse. In this paper, we take a step toward addressing this limitation by analyzing NC1 using kernels associated with shallow NNs. By considering the NN Gaussian Process kernel (NNGP), and the complement Neural Tangent Kernel (NTK), we show that the NTK surprisingly does not represent more collapsed features than the NNGP for gaussian data of arbitrary dimensions. We then consider an alternative to NTK: the recently proposed adaptive kernel, which generalizes NNGP to model the feature mapping learned from the training data. Through this *"kernel vs. kernel"* analysis, we present insights into the settings (data dimension, sample size, width) under which the kernel based NC1 aligns with that of shallow NNs.

Keywords: Neural Collapse, Feature Learning, Kernel Methods

1. Introduction

Deep Neural Network classifiers are often trained beyond the zero training error point (Hoffer et al., 2017; Ma et al., 2018). In this regime, a phenomenon dubbed "Neural Collapse" (NC) emerges (Papyan et al., 2020). Recently, a vast amount of literature has been dedicated to exploring NC (as surveyed in Kothapalli (2023)), studying the effect of imbalanced data (Fang et al., 2021; Thrampoulidis et al., 2022), depthwise evolution (Tirer and Bruna, 2022; Rangamani et al., 2023; Sukenik et al., 2023; He and Su, 2023), fine-grained structures (Tirer et al., 2023; Yang et al., 2023; Kothapalli et al., 2023), and implications (Zhu et al., 2021; Galanti et al., 2022; Yang et al., 2022). The most important aspect of NC is the collapse of within-class variability of features (NC1), as one may not gain valuable insights into the structure of the feature class means without NC1 (Tirer et al., 2023; Yang et al., 2023; Kothapalli et al., 2023; Galanti et al., 2022; He and Su, 2023; Xu and Liu, 2023). Notably, most of the works that theoretically analyze the NC behavior are based on variants of the unconstrained features model (UFM) (Mixon et al., 2020), which treats the deepest features of the training samples as free optimization variables. A key limitation of such analyses is that they cannot predict the effect of the data on NC1.

In this paper, we provide a kernel-based analysis of NC1 for Gaussian data, which does not suffer from the limitations of UFM-based analysis. Since kernels provide fixed feature mapping, we propose a *"kernel vs. kernel"* analysis — that is, gaining insights by comparing the properties across NN-related kernels. Our main contributions are as follows:

- We establish expressions for the NC1 metric that depends on the features only through the given arbitrary kernel function.
- We specialize our kernel-based NC1 to kernels associated with shallow NNs. We analyze it for NNGP (Neal, 1995; Lee et al., 2018; Matthews et al., 2018) and NTK (Jacot et al., 2018) and show that, perhaps surprisingly, the NTK does not represent more collapsed features than the NNGP for gaussian data. We also analyze the recently proposed adaptive kernel (Seroussi et al., 2023), which generalizes NNGP to model the feature mapping learned from the training data.
- Finally, we present insights into the settings (data dimension, sample size) under which kernel based NC1 aligns with NC1 of shallow NNs.

2. Problem Setup

Data. Consider dataset $\mathbf{X} \in \mathbb{R}^{d_0 \times N}$, comprising N data points of dimension d_0 belonging to C classes. Each class has size $n_c, c \in [C]$, where $[C] := \{1, 2, \cdots, C\}$ and $\sum_c n_c = N$. The dataset is represented in an "organized" matrix form as $\mathbf{X} = \begin{bmatrix} \mathbf{x}^{1,1} & \cdots & \mathbf{x}^{C,n_C} \end{bmatrix} \in \mathbb{R}^{d_0 \times N}$, where $\mathbf{x}^{c,i} \in \mathbb{R}^{d_0}$ represents the i^{th} sample of the c^{th} class.

Neural Network. We consider a 2-layer fully connected neural network (2L-FCN) $\psi : \mathbb{R}^{d_0} \to \mathbb{R}$ with hidden layer width d_1, and point-wise activation function $\phi(\cdot) : \mathbb{R} \to \mathbb{R}$. Let $\mathbf{W}^{(1)} \in \mathbb{R}^{d_1 \times d_0}$ and $\mathbf{w}^{(2)} \in \mathbb{R}^{d_1}$ denote the weight parameters of the first and second layers, respectively. At initialization, $W_{ij}^{(1)} \sim \mathcal{N}(0, \sigma_w^2/d_0)$ and $w_i^{(2)} \sim \mathcal{N}(0, \sigma_w^2/d_1)$ are drawn i.i.d. For an input $\mathbf{x} \in \mathbb{R}^{d_0}$, the network outputs:

$$\psi(\mathbf{x}) = \sum_{j=1}^{d_1} w_j^{(2)} \phi\left(z_j(\mathbf{x})\right); \qquad z_j(\mathbf{x}) = \sum_{k=1}^{d_0} W_{jk}^{(1)} x_k. \tag{1}$$

Task. We consider a binary classification task, where the network $\psi(\cdot)$ maps the samples $\mathbf{x}^{c,i}, c \in \{1, 2\}$, $i \in [n_c]$ to their respective target labels $y^{c,i} \in \{-1, 1\}$.

Pre- and Post-activation Kernels. For any two inputs $\mathbf{x}^{c,i}, \mathbf{x}^{c',j} \in \mathbb{R}^{d_0}$, we denote their corresponding pre- and post-activation features as $\mathbf{z}^{c,i}, \mathbf{z}^{c',j} \in \mathbb{R}^{d_1}$ and $\phi(\mathbf{z}^{c,i}), \phi(\mathbf{z}^{c',j}) \in \mathbb{R}^{d_1}$, respectively. Here $\mathbf{z}^{c,i} = \mathbf{W}^{(1)}\mathbf{x}^{c,i}$. The pre and post-activation kernels are given by:

$$K^{(1)}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = \mathbf{z}^{c,i\top}\mathbf{z}^{c',j}; \qquad Q^{(1)}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = \phi(\mathbf{z}^{c,i})^\top \phi(\mathbf{z}^{c',j}). \tag{2}$$

3. Main Results

3.1. Within-Class Variability Metric (NC1) for Kernels

Let $\mathbf{H}$ be a matrix of arbitrary features associated with samples of the C classes. Consider the within-class covariance $\mathbf{\Sigma}_W(\mathbf{H})$ and between-class covariance $\mathbf{\Sigma}_B(\mathbf{H})$ matrices as follows:

$$\mathbf{\Sigma}_W(\mathbf{H}) = \frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{n_c} \left(\mathbf{h}^{c,i} - \overline{\mathbf{h}}^c\right) \left(\mathbf{h}^{c,i} - \overline{\mathbf{h}}^c\right)^\top; \;\; \mathbf{\Sigma}_B(\mathbf{H}) = \frac{1}{C} \sum_{c=1}^{C} \left(\overline{\mathbf{h}}^c - \overline{\mathbf{h}}^G\right) \left(\overline{\mathbf{h}}^c - \overline{\mathbf{h}}^G\right)^\top$$

where $\overline{\mathbf{h}}^c = \frac{1}{n_c}\sum_{i=1}^{n_c}\mathbf{h}^{c,i}, \forall c \in [C]$ and $\overline{\mathbf{h}}^G = \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{n_c}\mathbf{h}^{c,i}$ represent the class mean vectors and the global mean vector, respectively. Based on these formulations, we define the variability metric $\mathcal{NC}_1(\mathbf{H})$, introduced in (Tirer et al., 2023) and used also in (Kothapalli et al., 2023; Wang et al., 2023; Yaras et al., 2023), as $\mathcal{NC}_1(\mathbf{H}) := \frac{\mathrm{tr}(\mathbf{\Sigma}_W(\mathbf{H}))}{\mathrm{tr}(\mathbf{\Sigma}_B(\mathbf{H}))}$. In the following theorem, we formulate the traces $\mathrm{tr}(\mathbf{\Sigma}_W(\mathbf{H}))$ and $\mathrm{tr}(\mathbf{\Sigma}_B(\mathbf{H}))$ using an arbitrary kernel function $Q : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$ that expresses inner product of samples in feature space.

Theorem 1 *For any two data points $\mathbf{x}^{c,i}, \mathbf{x}^{c',j}$, let the inner-product of their associated features $\mathbf{h}^{c,i}, \mathbf{h}^{c',j}$ be given by a kernel $Q : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$ as $Q(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = \mathbf{h}^{c,i\top}\mathbf{h}^{c',j}$. Consider $S(c) = \sum_{i=1}^{n_c} Q(\mathbf{x}^{c,i}, \mathbf{x}^{c,i})$, and $S(c,c') = \sum_{i=1}^{n_c}\sum_{j=1}^{n_{c'}} Q(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})$. The traces of covariance matrices $\mathrm{tr}(\mathbf{\Sigma}_W(\mathbf{H}))$ and $\mathrm{tr}(\mathbf{\Sigma}_B(\mathbf{H}))$ can now be formulated as:*

$$\mathrm{tr}(\mathbf{\Sigma}_W(\mathbf{H})) = \sum_{c=1}^{C}\frac{S(c)}{N} - \sum_{c=1}^{C}\frac{S(c,c)}{Cn_c^2}; \quad \mathrm{tr}(\mathbf{\Sigma}_B(\mathbf{H})) = \sum_{c=1}^{C}\frac{S(c,c)}{Cn_c^2} - \sum_{c=1}^{C}\sum_{c'=1}^{C}\frac{S(c,c')}{N^2}. \tag{3}$$

3.2. Activation Variability in the Lazy Learning Regime

Consider the samples of a 1-dimensional Gaussian dataset (i.e., $d_0 = 1$) with $C = 2$ classes as $\{x^{1,i}\} \sim \mathcal{N}(\mu_1, \sigma_1^2), \forall i \in [n_1]$ and $\{x^{2,j}\} \sim \mathcal{N}(\mu_2, \sigma_2^2), \forall j \in [n_2]$.

- **Assumption 1:** For $\mu_1 < 0, \mu_2 > 0$, let $\sigma_1, \sigma_2 > 0$ be small enough such that $|\mu_1| \gg \sigma_1$, $|\mu_2| \gg \sigma_2$ and $\forall i \in [n_1], j \in [n_2], x^{1,i}x^{2,j} < 0$ almost surely.
- **Assumption 2:** The dataset $\mathbf{X} \in \mathbb{R}^{N\times 1}$ consists of large enough samples $n_1, n_2 \gg 1$.

Theorem 2 *Under Assumptions 1-2, let $\phi(\cdot)$ be the ReLU activation. Denote by $\mathbf{H}_{GP}$, $\mathbf{H}_{NTK}$ the features associated with NNGP $Q_{GP}^{(1)}$ and NTK $\Theta_{NTK}^{(2)}$, respectively. Then:*

$$\mathbb{E}\left[\mathcal{NC}_1(\mathbf{H}_{GP})\right] = \mathbb{E}\left[\mathcal{NC}_1(\mathbf{H}_{NTK})\right] = \frac{\sum_{c=1}^{2}\frac{n_c\mu_c^2 + n_c\sigma_c^2}{N} - \frac{\mu_c^2}{2}}{\left(\sum_{c=1}^{2}\frac{\mu_c^2}{2} - \frac{n_c^2\mu_c^2}{N^2}\right) - \frac{2}{N^2}\prod_{c=1}^{2} n_c\mu_c} + \Delta_{h.o.t}, \tag{4}$$

where $\Delta_{h.o.t}$ is a term that vanishes as $\{n_c\}$ increase.

Theorem 2 shows that: *the NTK does not represent more collapsed features than NNGP, despite being associated with NN gradient-based optimization.* Experiments with high dimensional data that empirically justify this result are presented in Appendix.

3.3. Activation Variability in the Feature Learning Regime

A transition from the infinite to finite width regime can introduce various corrections to the pre-and post-activations of a L-layer FCN (Seroussi et al., 2023): (1) The mean and covariance of the pre-activations deviate from that of a random FCN and, (2) the collective effect of activations from the $(l+1)^{th}$ and $(l-1)^{th}$ layers determine the covariance of activations in the l^{th} layer. Based on these observations, Seroussi et al. (2023) propose a Variational Gaussian Approximation (VGA) approach to propose a system of equations, dubbed Equation of State (EoS), for the pre and post-activation kernels $K^{(l)}(\cdot,\cdot), Q^{(l)}(\cdot,\cdot), l \in [L]$ respectively.

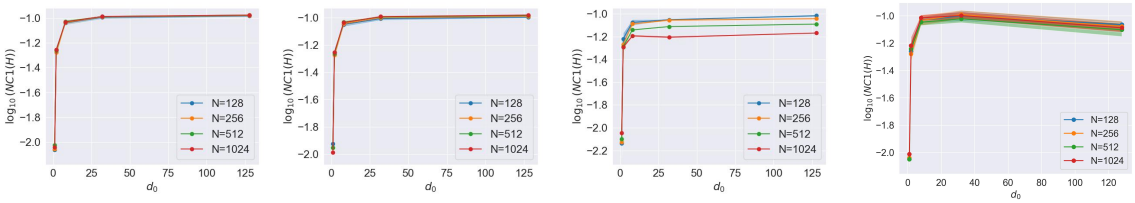
Definition 3 *The "Equations of State" (EoS) for pre and post-activation kernels of a 2L-FCN with* `Erf` *activation are given by:*

$$\bar{\mathbf{f}} = \mathbf{Q}^{(1)}[\sigma^2\mathbf{I} + \mathbf{Q}^{(1)}]^{-1}\mathbf{y}; \quad [\mathbf{Q}^{(1)}]_{ij} = \sigma_a^2 \frac{2}{\pi} \arcsin\left(2K_{ij}^{(1)} \cdot \left(\sqrt{1+2K_{ii}^{(1)}}\sqrt{1+2K_{jj}^{(1)}}\right)^{-1}\right)$$

$$[\mathbf{C}^{-1}]_{ij} = \frac{d_0}{\sigma_w^2}\delta_{ij} + \frac{1}{d_1}\mathrm{tr}\left\{\mathbf{A}^{(1)}\partial_{C_{ij}}\mathbf{Q}^{(1)}\right\}; \quad \mathbf{A}^{(1)} = -(\mathbf{y}-\bar{\mathbf{f}})(\mathbf{y}-\bar{\mathbf{f}})^\top\sigma^{-4} + [\mathbf{Q}^{(1)} + \sigma^2\mathbf{I}]^{-1}$$

Here, $\mathbf{C} \in \mathbb{R}^{d_0 \times d_0}$ *models the statistical covariance of a row of* $\mathbf{W}^{(1)}$*, initialized with* $(\sigma_w^2/d_0)\mathbf{I}$*,* $\mathbf{K}^{(1)} = \mathbf{X}^\top\mathbf{C}\mathbf{X} \in \mathbb{R}^{N\times N}$*,* $\sigma > 0$ *is the regularization parameter, and* $\bar{\mathbf{f}} \in \mathbb{R}^N$ *corresponds to the prediction of the 2-layer FCN (governed by the EoS). Additionally,* $\mathbf{K}^{(1)}, \mathbf{Q}^{(1)} \in \mathbb{R}^{N\times N}$ *are the kernel matrices associated with kernel functions* $K^{(1)}(\cdot,\cdot), Q^{(1)}(\cdot,\cdot)$*.*

3.4. Approximating NC1 of 2L-FCN with Kernels



(a) Q_{GP-Erf} (b) $\Theta_{NTK-Erf}$ (c) EoS (d) 2L-FCN

Figure 1: $\mathcal{NC}_1(\mathbf{H})$ of NNGP $Q^{(1)}_{GP-Erf}$, NTK $\Theta^{(2)}_{NTK-Erf}$, EoS and 2L-FCN on $\mathcal{D}_1(N, d_0)$.

Setup. We train a 2L-FCN with $d_1 = 500, \sigma_w = 1$, `Erf` activation, vanilla GD with learning rate 10^{-3}, weight-decay 10^{-6} for 1000 epochs on dataset $\mathcal{D}_1(N, d_0), \forall i, j \in [N/2]$.:

$$\left\{(\mathbf{x}^{1,i} \sim \mathcal{N}(-2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{1,i} = -1))\right\} \cup \left\{(\mathbf{x}^{2,j} \sim \mathcal{N}(2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{1,i} = 1))\right\}.$$

Results. The trends in $\mathcal{NC}_1(\mathbf{H})$ for NNGP (Figure 1(a)) and NTK (Figure 1(b)) fail to capture the finite width effects of training 2L-FCNs for larger d_0 but provide zero-order reasoning for NN behavior when d_0 is small (i.e intuitively for less complex data distributions). The trends of $\mathcal{NC}_1(\mathbf{H})$ for EoS (Figure 1(c)) (which is solved using an annealing approach), vary depending on the scale of N, d_0. For $d_0 = \{1, 2\}$, the $\mathcal{NC}_1(\mathbf{H})$ of EoS resembles the 2L-FCN case (Figure 1(d)) across a range of sample sizes N. Furthermore, for larger d_0, we observe a proportional scaling behaviour where an increase in N is required to match the 2L-FCN behaviour (while showcasing reduced NC1 compared to NNGP).

4. Conclusion

In this paper, we presented a *"kernel vs. kernel"* approach to analyze the data properties for which the NC1 behavior of an actual FCN can be understood. We believe that future work on analyzing the EoS mathematically and extending it to multiple layers can provide further insights into the reduction of NC1 in deep neural networks.

References

Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2022.

Hangfeng He and Weijie J Su. A law of data separation in deep learning. *Proceedings of the National Academy of Sciences*, 120(36):e2221704120, 2023.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

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

Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research*, 2023.

Vignesh Kothapalli, Tom Tirer, and Joan Bruna. A neural collapse perspective on feature evolution in graph neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

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. *Advances in neural information processing systems*, 32, 2019.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

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

Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 1995.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Akshay Rangamani, Marius Lindegaard, Tomer Galanti, and Tomaso A Poggio. Feature learning in deep classifiers through intermediate neural collapse. In *International Conference on Machine Learning*, pages 28729–28745. PMLR, 2023.

Howard Seltman. Approximations for mean and variance of a ratio. *unpublished note*, 2012.

Inbar Seroussi, Gadi Naveh, and Zohar Ringel. Separation of scales and a thermodynamic description of feature learning in some cnns. *Nature Communications*, 14(1):908, 2023.

Peter Sukenik, Marco Mondelli, and Christoph H Lampert. Deep neural collapse is provably optimal for the deep unconstrained features model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35:27225–27238, 2022.

Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In *International Conference on Machine Learning*, pages 21478–21505. PMLR, 2022.

Tom Tirer, Joan Bruna, and Raja Giryes. Kernel-based smoothness analysis of residual networks. In *Mathematical and Scientific Machine Learning*, pages 921–954. PMLR, 2022.

Tom Tirer, Haoxiang Huang, and Jonathan Niles-Weed. Perturbation analysis of neural collapse. In *International Conference on Machine Learning*, pages 34301–34329. PMLR, 2023.

Roman Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

Peng Wang, Xiao Li, Can Yaras, Zhihui Zhu, Laura Balzano, Wei Hu, and Qing Qu. Understanding deep representation learning via layerwise feature compression and discrimination. *arXiv preprint arXiv:2311.02960*, 2023.

Jing Xu and Haoxiong Liu. Quantifying the variability collapse of neural networks. In *International Conference on Machine Learning*, pages 38535–38550. PMLR, 2023.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? *Advances in Neural Information Processing Systems*, 35:37991–38002, 2022.

Yongyi Yang, Jacob Steinhardt, and Wei Hu. Are neurons actually collapsed? on the fine-grained structure in neural representations. In *International Conference on Machine Learning*, pages 39453–39487. PMLR, 2023.

Can Yaras, Peng Wang, Wei Hu, Zhihui Zhu, Laura Balzano, and Qing Qu. The law of parsimony in gradient descent for learning deep linear networks. *arXiv preprint arXiv:2306.01154*, 2023.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

Appendix A. Proof of Theorem 1

To obtain the NC1 formulation corresponding to an arbitrary feature matrix $\mathbf{H}$, we first define the total covariance $\widetilde{\mathbf{\Sigma}}_T(\mathbf{H})$ and non-centered between-class covariance $\widetilde{\mathbf{\Sigma}}_B(\mathbf{H})$ matrices as follows:

$$\widetilde{\mathbf{\Sigma}}_T(\mathbf{H}) = \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{n_c}\mathbf{h}^{c,i}\mathbf{h}^{c,i\top} \tag{5}$$

$$\widetilde{\mathbf{\Sigma}}_B(\mathbf{H}) = \frac{1}{C}\sum_{c=1}^{C}\overline{\mathbf{h}}^{c}\overline{\mathbf{h}}^{c\top}. \tag{6}$$

A simple relationship between $\widetilde{\mathbf{\Sigma}}_T(\mathbf{H}), \widetilde{\mathbf{\Sigma}}_B(\mathbf{H}), \mathbf{\Sigma}_W(\mathbf{H})$ is as follows:

$$\begin{aligned} \widetilde{\mathbf{\Sigma}}_T(\mathbf{H}) &= \mathbf{\Sigma}_W(\mathbf{H}) + \widetilde{\mathbf{\Sigma}}_B(\mathbf{H}) \\ \implies \operatorname{tr}\left(\mathbf{\Sigma}_W(\mathbf{H})\right) &= \operatorname{tr}\left(\widetilde{\mathbf{\Sigma}}_T(\mathbf{H})\right) - \operatorname{tr}\left(\widetilde{\mathbf{\Sigma}}_B(\mathbf{H})\right). \end{aligned} \tag{7}$$

Similarly, by considering $\mathbf{\Sigma}_G(\mathbf{H}) = \overline{\mathbf{h}}^{G}\overline{\mathbf{h}}^{G\top}$, we get:

$$\begin{aligned} \mathbf{\Sigma}_B(\mathbf{H}) &= \widetilde{\mathbf{\Sigma}}_B(\mathbf{H}) - \mathbf{\Sigma}_G(\mathbf{H}) \\ \implies \operatorname{tr}\left(\mathbf{\Sigma}_B(\mathbf{H})\right) &= \operatorname{tr}\left(\widetilde{\mathbf{\Sigma}}_B(\mathbf{H})\right) - \operatorname{tr}\left(\mathbf{\Sigma}_G(\mathbf{H})\right). \end{aligned} \tag{8}$$

• **Formulating** $\operatorname{tr}\left(\widetilde{\mathbf{\Sigma}}_T(\mathbf{H})\right)$**:** Expanding $\widetilde{\mathbf{\Sigma}}_T(\mathbf{H})$ into individual outer-products of vectors and leveraging the trace properties leads to the following:

$$\begin{aligned} \operatorname{tr}\left(\widetilde{\mathbf{\Sigma}}_T(\mathbf{H})\right) &= \operatorname{tr}\left(\frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{n_c}\mathbf{h}^{c,i}\mathbf{h}^{c,i\top}\right) = \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{n_c}\operatorname{tr}\left(\mathbf{h}^{c,i}\mathbf{h}^{c,i\top}\right) \\ &= \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{n_c}\operatorname{tr}\left(\mathbf{h}^{c,i\top}\mathbf{h}^{c,i}\right) \\ &= \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{n_c}Q(\mathbf{x}^{c,i},\mathbf{x}^{c,i}) \end{aligned}$$

• **Formulating** $\operatorname{tr}\left(\widetilde{\mathbf{\Sigma}}_B(\mathbf{H})\right)$**:** Similar to the above analysis, we can reformulate the trace of non-centered between-class covariance matrix $\widetilde{\mathbf{\Sigma}}_B(\mathbf{H})$ as:

$$\begin{aligned}
\operatorname{tr}(\widetilde{\mathbf{\Sigma}}_B) &= \operatorname{tr}\left(\frac{1}{C}\sum_{c=1}^{C}\overline{\mathbf{h}}^{c}\overline{\mathbf{h}}^{c\top}\right) = \frac{1}{C}\sum_{c=1}^{C}\operatorname{tr}\left(\overline{\mathbf{h}}^{c}\overline{\mathbf{h}}^{c\top}\right) = \frac{1}{C}\sum_{c=1}^{C}\operatorname{tr}\left(\overline{\mathbf{h}}^{c\top}\overline{\mathbf{h}}^{c}\right) \\
&= \frac{1}{C}\sum_{c=1}^{C}\operatorname{tr}\left(\left[\frac{1}{n_c}\sum_{i=1}^{n_c}\mathbf{h}^{c,i}\right]^{\top}\left[\frac{1}{n_c}\sum_{i=1}^{n_c}\mathbf{h}^{c,i}\right]\right) \\
&= \frac{1}{C}\sum_{c=1}^{C}\frac{1}{n_c^2}\operatorname{tr}\left(\sum_{i=1}^{n_c}\sum_{j=1}^{n_c}\mathbf{h}^{c,i\top}\mathbf{h}^{c,j}\right) = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{n_c^2}\sum_{i=1}^{n_c}\sum_{j=1}^{n_c}\operatorname{tr}\left(\mathbf{h}^{c,i\top}\mathbf{h}^{c,j}\right) \\
&= \frac{1}{C}\sum_{c=1}^{C}\frac{1}{n_c^2}\sum_{i=1}^{n_c}\sum_{j=1}^{n_c}Q(\mathbf{x}^{c,i},\mathbf{x}^{c,j})
\end{aligned}$$

• **Formulating** $\operatorname{tr}\left(\mathbf{\Sigma}_G(\mathbf{H})\right)$**:** Reformulation of $\operatorname{tr}\left(\mathbf{\Sigma}_G(\mathbf{H})\right)$ can be approached along the same lines:

$$\begin{aligned}
\operatorname{tr}\left(\mathbf{\Sigma}_G(\mathbf{H})\right) &= \operatorname{tr}\left(\overline{\mathbf{h}}^{G}\overline{\mathbf{h}}^{G\top}\right) = \operatorname{tr}\left(\overline{\mathbf{h}}^{G\top}\overline{\mathbf{h}}^{G}\right) \\
&= \operatorname{tr}\left(\left[\frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{n_c}\mathbf{h}^{c,i}\right]^{\top}\left[\frac{1}{N}\sum_{c=1}^{C}\sum_{j=1}^{n_c}\mathbf{h}^{c,j}\right]\right) \\
&= \frac{1}{N^2}\operatorname{tr}\left(\sum_{c=1}^{C}\sum_{i=1}^{n_c}\sum_{c'=1}^{C}\sum_{j=1}^{n_{c'}}\mathbf{h}^{c,i\top}\mathbf{h}^{c',j}\right) = \frac{1}{N^2}\sum_{c=1}^{C}\sum_{i=1}^{n_c}\sum_{c'=1}^{C}\sum_{j=1}^{n_{c'}}\operatorname{tr}\left(\mathbf{h}^{c,i\top}\mathbf{h}^{c',j}\right) \\
&= \frac{1}{N^2}\sum_{c=1}^{C}\sum_{c'=1}^{C}\sum_{i=1}^{n_c}\sum_{j=1}^{n_{c'}}Q(\mathbf{x}^{c,i},\mathbf{x}^{c',j})
\end{aligned}$$

By using these intermediate results, we can formulate $\operatorname{tr}\left(\mathbf{\Sigma}_W(\mathbf{H})\right), \operatorname{tr}\left(\mathbf{\Sigma}_B(\mathbf{H})\right)$ as:

$$\begin{aligned}
\operatorname{tr}(\mathbf{\Sigma}_W(\mathbf{H})) &= \operatorname{tr}(\widetilde{\mathbf{\Sigma}}_T(\mathbf{H})) - \operatorname{tr}(\widetilde{\mathbf{\Sigma}}_B(\mathbf{H})) \\
&= \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{n_c}Q(\mathbf{x}^{c,i},\mathbf{x}^{c,i}) - \frac{1}{C}\sum_{c=1}^{C}\frac{1}{n_c^2}\sum_{i=1}^{n_c}\sum_{j=1}^{n_c}Q(\mathbf{x}^{c,i},\mathbf{x}^{c,j}) \\
\operatorname{tr}(\mathbf{\Sigma}_B(\mathbf{H})) &= \operatorname{tr}(\widetilde{\mathbf{\Sigma}}_B(\mathbf{H})) - \operatorname{tr}(\mathbf{\Sigma}_G(\mathbf{H}))) \\
&= \frac{1}{C}\sum_{c=1}^{C}\frac{1}{n_c^2}\sum_{i=1}^{n_c}\sum_{j=1}^{n_c}Q(\mathbf{x}^{c,i},\mathbf{x}^{c,j}) - \frac{1}{N^2}\sum_{c=1}^{C}\sum_{c'=1}^{C}\sum_{i=1}^{n_c}\sum_{j=1}^{n_{c'}}Q(\mathbf{x}^{c,i},\mathbf{x}^{c',j}).
\end{aligned}$$

Hence, proving the theorem.

Appendix B. Limiting NNGP and NTK for `ReLU` and `Erf`

Consider the GP limit characterization of the pre-activation kernel $K^{(1)}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})$ as follows:

$$K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = \frac{\sigma_w^2}{d_0}\mathbf{x}^{c,i\top}\mathbf{x}^{c',j}. \tag{9}$$

ReLU Activation. Observe that $K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})$ is independent of the activation function. Now, the closed form representation of the post-activation NNGP kernel $Q^{(1)}_{GP}(\cdot,\cdot)$ for the `ReLU` activation is given by:

$$\begin{aligned} Q^{(1)}_{GP-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) &= \frac{\tau(x^{c,i}, x^{c',j})}{2\pi}\sqrt{K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c,i})K^{(1)}_{GP}(\mathbf{x}^{c',j}, \mathbf{x}^{c',j})}, \\ \tau(x^{c,i}, x^{c',j}) &= \sin\theta^{c',j}_{c,i} + \left(\pi - \theta^{c',j}_{c,i}\right)\cos\theta^{c',j}_{c,i} \\ \theta^{c',j}_{c,i} &= \arccos\left(\frac{K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})}{\sqrt{K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c,i})K^{(1)}_{GP}(\mathbf{x}^{c',j}, \mathbf{x}^{c',j})}}\right). \end{aligned} \tag{10}$$

Next, we define the `ReLU` based derivative kernel $\dot{Q}^{(1)}_{GP-ReLU}(\cdot,\cdot)$ as follows:

$$\dot{Q}^{(1)}_{GP-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = \frac{1}{2\pi}\left(\pi - \theta\right) \tag{11}$$

Erf Activation. The kernel $Q^{(1)}_{GP-Erf}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})$ for the `Erf` activation is given by:

$$Q^{(1)}_{GP-Erf}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = \frac{2}{\pi}\arcsin\left(\frac{2K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})}{\sqrt{1 + 2K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c,i})}\sqrt{1 + 2K^{(1)}_{GP}(\mathbf{x}^{c',j}, \mathbf{x}^{c',j})}}\right). \tag{12}$$

The `Erf` based derivative kernel $\dot{Q}^{(1)}_{GP-Erf}(\cdot,\cdot)$ is formulated as follows:

$$\dot{Q}^{(1)}_{GP-Erf}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = \frac{4}{\pi}\det\left(\begin{bmatrix} 1 + 2K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c,i}) & 2K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) \\ 2K^{(1)}_{GP}(\mathbf{x}^{c',j}, \mathbf{x}^{c,i}) & 1 + 2K^{(1)}_{GP}(\mathbf{x}^{c',j}, \mathbf{x}^{c',j}) \end{bmatrix}\right)^{-1/2}. \tag{13}$$

Finally, the NTK can be formulated (independent of the activation function) as follows:

$$\Theta^{(2)}_{NTK}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = K^{(2)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) + K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})\dot{Q}^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}). \tag{14}$$

Here, $K^{(2)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})$ can be defined using the recursive formulation:

$$K^{(2)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = \sigma_w^2 Q^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}). \tag{15}$$

Depending on the choice of the activation, one can plug in the variant specific NNGP kernels to obtain the specialized NTK formulations.

Appendix C. General Results for NC1 with Kernels

In this section, we present some general results to calculate the expected value of $\mathbb{E}\left[\mathcal{NC}_1(\mathbf{H})\right]$ for any given kernel function $Q(\cdot,\cdot)$ that is associated with the features $\mathbf{H}$. To begin with, we consider a generic formulation of the three cases for $\mathbb{E}\left[Q(x^{c,i}, x^{c',j})\right]$:

$$\mathbb{E}\left[Q(x^{c,i}, x^{c',j})\right] = \begin{cases} V^{(1)}(c) & \text{if } c = c', i = j \\ V^{(2)}(c) & \text{if } c = c', i \neq j \\ V^{(3)}(c, c') & \text{if } c \neq c' \end{cases}. \tag{16}$$

Lemma 4 *Given the cases for the expected values of a kernel function $Q(\cdot,\cdot)$ as per (16), the $\mathbb{E}\left[\mathrm{tr}(\mathbf{\Sigma}_W(\mathbf{H}))\right]$ is given by:*

$$\mathbb{E}\left[\mathrm{tr}(\mathbf{\Sigma}_W(\mathbf{H}))\right] = \sum_{c=1}^{2} \frac{n_c}{N} V^{(1)}(c) - \frac{1}{2n_c^2}\left(n_c(n_c-1)V^{(2)}(c) + n_c V^{(1)}(c)\right) \tag{17}$$

Proof *By leveraging Theorem 4.1, we can compute the expected value of $\mathrm{tr}(\mathbf{\Sigma}_W(\mathbf{H}))$ as follows:*

$$\begin{aligned} \mathbb{E}\left[\mathrm{tr}(\mathbf{\Sigma}_W(\mathbf{H}))\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{n_c} Q(x^{c,i}, x^{c,i})\right] - \mathbb{E}\left[\frac{1}{C}\sum_{c=1}^{C}\frac{1}{n_c^2}\sum_{i=1}^{n_c}\sum_{j=1}^{n_c} Q(x^{c,i}, x^{c,j})\right] \\ &= \frac{1}{N}\sum_{c=1}^{2}\sum_{i=1}^{n_c}\mathbb{E}\left[Q(x^{c,i}, x^{c,i})\right] - \frac{1}{2}\sum_{c=1}^{2}\frac{1}{n_c^2}\sum_{i=1}^{n_c}\sum_{j=1}^{n_c}\mathbb{E}\left[Q(x^{c,i}, x^{c,j})\right] \\ &= \frac{1}{N}\sum_{c=1}^{2}\sum_{i=1}^{n_c} V^{(1)}(c) - \frac{1}{2}\sum_{c=1}^{2}\frac{1}{n_c^2}\left(n_c(n_c-1)V^{(2)}(c) + n_c V^{(1)}(c)\right) \\ &= \sum_{c=1}^{2}\frac{n_c}{N}V^{(1)}(c) - \frac{1}{2n_c^2}\left(n_c(n_c-1)V^{(2)}(c) + n_c V^{(1)}(c)\right). \end{aligned} \tag{18}$$

Hence proving the lemma. ■

Lemma 5 *Given the cases for the expected values of a kernel function $Q(\cdot,\cdot)$ as per (16), the $\mathbb{E}\left[\mathrm{tr}(\mathbf{\Sigma}_B(\mathbf{H}))\right]$ is given by:*

$$\mathbb{E}\left[\mathrm{tr}(\mathbf{\Sigma}_B(\mathbf{H}))\right] = \left[\sum_{c=1}^{2}\left(\frac{1}{2n_c^2} - \frac{1}{N^2}\right)\left(n_c(n_c-1)V^{(2)}(c) + n_c V^{(1)}(c)\right)\right] - \frac{2n_1 n_2}{N^2}V^{(3)}(1,2) \tag{19}$$

Proof *The expected value of* $\text{tr}(\mathbf{\Sigma}_B(\mathbf{H}))$ *can be computed using Theorem 4.1 as:*

$$\begin{aligned}
\mathbb{E}\left[\text{tr}(\mathbf{\Sigma}_B(\mathbf{H}))\right] &= \mathbb{E}\left[\frac{1}{C}\sum_{c=1}^{C}\frac{1}{n_c^2}\sum_{i=1}^{n_c}\sum_{j=1}^{n_c}\mathbf{Q}(x^{c,i},x^{c,j})\right] - \mathbb{E}\left[\frac{1}{N^2}\sum_{c=1}^{C}\sum_{c'=1}^{C}\sum_{i=1}^{n_c}\sum_{j=1}^{n_{c'}}\mathbf{Q}(x^{c,i},x^{c',j})\right] \\
&= \left[\frac{1}{2}\sum_{c=1}^{2}\frac{1}{n_c^2}\left(n_c(n_c-1)V^{(2)}(c)+n_cV^{(1)}(c)\right)\right] \\
&\quad - \frac{1}{N^2}\left[\sum_{c=1}^{2}\left(n_c(n_c-1)V^{(2)}(c)+n_cV^{(1)}(c)\right)\right] \\
&\quad - \frac{1}{N^2}\left[2\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}V^{(3)}(c=1,c'=2)\right] \\
&= \left[\sum_{c=1}^{2}\left(\frac{1}{2n_c^2}-\frac{1}{N^2}\right)\left(n_c(n_c-1)V^{(2)}(c)+n_cV^{(1)}(c)\right)\right] - \frac{2n_1n_2}{N^2}V^{(3)}(1,2)
\end{aligned} \tag{20}$$

Hence proving the lemma. ■

Lemma 6 *Given the cases for the expected values of a kernel function* $Q(\cdot,\cdot)$ *as per* (16)*, the* $\mathbb{E}\left[\mathcal{NC}_1(\mathbf{H})\right]$ *is given by:*

$$\mathbb{E}\left[\mathcal{NC}_1(\mathbf{H})\right] = \frac{\sum_{c=1}^{2}\frac{n_cV^{(1)}(c)}{N} - \frac{\left(n_c(n_c-1)V^{(2)}(c)+n_cV^{(1)}(c)\right)}{2n_c^2}}{\left[\sum_{c=1}^{2}\left(\frac{1}{2n_c^2}-\frac{1}{N^2}\right)\left(n_c(n_c-1)V^{(2)}(c)+n_cV^{(1)}(c)\right)\right] - \frac{2n_1n_2V^{(3)}(1,2)}{N^2}} + \Delta_{h.o.t} \tag{21}$$

Proof *Note that the expectation of the ratios can be given as:*

$$\mathbb{E}\left[\mathcal{NC}_1(\mathbf{H})\right] = \frac{\mathbb{E}\left[\text{tr}(\Sigma_W(\mathbf{H}))\right]}{\mathbb{E}\left[\text{tr}(\Sigma_B(\mathbf{H}))\right]} + \Delta_{h.o.t} \tag{22}$$

$$= \frac{\sum_{c=1}^{2}\frac{n_cV^{(1)}(c)}{N} - \frac{\left(n_c(n_c-1)V^{(2)}(c)+n_cV^{(1)}(c)\right)}{2n_c^2}}{\left[\sum_{c=1}^{2}\left(\frac{1}{2n_c^2}-\frac{1}{N^2}\right)\left(n_c(n_c-1)V^{(2)}(c)+n_cV^{(1)}(c)\right)\right] - \frac{2n_1n_2V^{(3)}(1,2)}{N^2}} + \Delta_{h.o.t} \tag{23}$$

Here, $\Delta_{h.o.t}$ *corresponds to higher order terms given by Seltman (2012):*

$$\Delta_{h.o.t} = \frac{Var(\text{tr}(\mathbf{\Sigma}_B(\mathbf{H})))\mathbb{E}\left[\text{tr}(\Sigma_W(\mathbf{H}))\right]}{\mathbb{E}\left[\text{tr}(\Sigma_B(\mathbf{H}))\right]^3} - \frac{Cov(\text{tr}(\mathbf{\Sigma}_W(\mathbf{H})),\text{tr}(\mathbf{\Sigma}_B(\mathbf{H})))}{\mathbb{E}\left[\text{tr}(\Sigma_B(\mathbf{H}))\right]^2}, \tag{24}$$

where, based on the well-studied concentration of sample covariance matrices around the statistical covariance Vershynin (2012), $\Delta_{h.o.t}$ *tend to* 0 *for large* n_c *values.* ■

Lemma 7 *For a random variable $x^{c,i} \sim \mathcal{N}(\mu_c, \sigma_c^2)$ which represents the i^{th} sample of class c, the expected value $\mathbb{E}\left[\frac{1}{(x^{c,i})^2}\right]$ is given by:*

$$T(c) = \mathbb{E}\left[\frac{1}{(x^{c,i})^2}\right] = \frac{1}{(\mu_c^2 + \sigma_c^2)} + \frac{2\sigma_c^4 + 4\sigma_c^2\mu_c^2}{(\mu_c^2 + \sigma_c^2)^3} \tag{25}$$

Proof *Based on the standard result on the expectation of ratios Seltman (2012), we get:*

$$\mathbb{E}\left[\frac{1}{(x^{c,i})^2}\right] = \frac{1}{\mathbb{E}\left[(x^{c,i})^2\right]} + \frac{Var((x^{c,i})^2)}{\mathbb{E}\left[(x^{c,i})^2\right]^3} \tag{26}$$

$$= \frac{1}{(\mu_c^2 + \sigma_c^2)} + \frac{\mathbb{E}[(x^{c,i})^4] - (\mu_c^2 + \sigma_c^2)^2}{(\mu_c^2 + \sigma_c^2)^3} \tag{27}$$

Based on the results from the moment-generating function, we know that:

$$\mathbb{E}[(x^{c,i})^4] = 3\sigma_c^4 + 6\sigma_c^2\mu_c^2 + \mu_c^4, \tag{28}$$

which gives us:

$$\mathbb{E}\left[\frac{1}{(x^{c,i})^2}\right] = \frac{1}{(\mu_c^2 + \sigma_c^2)} + \frac{3\sigma_c^4 + 6\sigma_c^2\mu_c^2 + \mu_c^4 - (\mu_c^2 + \sigma_c^2)^2}{(\mu_c^2 + \sigma_c^2)^3} \tag{29}$$

$$= \frac{1}{(\mu_c^2 + \sigma_c^2)} + \frac{2\sigma_c^4 + 4\sigma_c^2\mu_c^2}{(\mu_c^2 + \sigma_c^2)^3}. \tag{30}$$

Hence proving the lemma. ■

Appendix D. Proof of Theorem 2

D.1. NC1 of limiting NNGP with `ReLU` activation

In the limit $d_1 \to \infty$, we leverage the kernels in the GP limit. Observe that for any two data points $x^{c,i}, x^{c',j} \in \mathbb{R}$, the value of $\theta_{c,i}^{c',j}$ can be given as:

$$\theta_{c,i}^{c',j} = \arccos\left(\frac{K_{GP}^{(1)}(x^{c,i}, x^{c',j})}{\sqrt{K_{GP}^{(1)}(x^{c,i}, x^{c,i}) K_{GP}^{(1)}(x^{c',j}, x^{c',j})}}\right)$$

$$= \arccos\left(\frac{\frac{\sigma_w^2}{d_0} x^{c,i} x^{c',j}}{\sqrt{\left(\frac{\sigma_w^2}{d_0} x^{c,i} x^{c,i}\right)\left(\frac{\sigma_w^2}{d_0} x^{c',j} x^{c',j}\right)}}\right).$$

The value of $\theta_{c,i}^{c',j}$ can be simplified to:

$$\theta_{c,i}^{c',j} = \begin{cases} 0 & \text{if } c = c' \\ \pi & \text{if } c \neq c' \end{cases}, \tag{31}$$

which follows from $\frac{x^{c,i}x^{c',j}}{\sqrt{x^{c,i}x^{c,i}}\sqrt{x^{c',j}x^{c',j}}} = \text{sign}(x^{c,i})\,\text{sign}(x^{c',j})$ and $x^{1,i} < 0, x^{2,j} > 0$ almost surely. Thus:

$$Q^{(1)}_{GP-ReLU}(x^{c,i}, x^{c',j}) = \frac{1}{2\pi}\sqrt{\sigma_w^4 (x^{c,i})^2 (x^{c',j})^2}\left(\sin\theta^{c',j}_{c,i} + \left(\pi - \theta^{c',j}_{c,i}\right)\cos\theta^{c',j}_{c,i}\right) \tag{32}$$

$$\implies Q^{(1)}_{GP-ReLU}(x^{c,i}, x^{c,j}) = \begin{cases} \frac{\sigma_w^2}{2}\left|x^{c,i}\right|\left|x^{c',j}\right| & \text{if } c = c' \\ 0 & \text{if } c \neq c' \end{cases} \tag{33}$$

For the $c = c'$ case, the value of the kernel boils down to the product of norms of independent random variables drawn from the same distribution. Since we assume $x^{c,i}x^{c',j} > 0$ if $c = c'$, the equation 33 can be rewritten as:

$$Q^{(1)}_{GP-ReLU}(x^{c,i}, x^{c,j}) = \begin{cases} \frac{\sigma_w^2}{2}x^{c,i}x^{c',j} & \text{if } c = c' \\ 0 & \text{if } c \neq c' \end{cases} \tag{34}$$

Additionally, since $x^{c,i}$ are random variables, the expected value of the kernel can be formulated as:

$$\mathbb{E}\left[Q^{(1)}_{GP-ReLU}(x^{c,i}, x^{c',j})\right] = \begin{cases} \frac{\sigma_w^2}{2}\left(\sigma_c^2 + \mu_c^2\right) & \text{if } c = c', i = j \\ \frac{\sigma_w^2}{2}\mu_c^2 & \text{if } c = c', i \neq j \\ 0 & \text{if } c \neq c' \end{cases} \tag{35}$$

Thus, based on our generic formulation of cases in (16) in Appendix C, we get:

$$V^{(1)}(c) = \frac{\sigma_w^2}{2}\left(\sigma_c^2 + \mu_c^2\right); \quad V^{(2)}(c) = \frac{\sigma_w^2}{2}\mu_c^2; \quad V^{(3)}(c, c') = 0. \tag{36}$$

As $N \gg 1$ and $n_c \gg 1, \forall c \in \{1, 2\}$, Lemma 6 gives us:

$$\begin{aligned} \mathbb{E}[\mathcal{NC}_1(\mathbf{H}_{GP})] &= \frac{\sum_{c=1}^{2}\frac{n_c V^{(1)}(c)}{N} - \frac{V^{(2)}(c)}{2}}{\left[\sum_{c=1}^{2}\left(\frac{1}{2n_c^2} - \frac{1}{N^2}\right)\left(n_c^2 V^{(2)}(c)\right)\right] - \frac{2n_1 n_2}{N^2}V^{(3)}(1,2)} + \Delta_{h.o.t} \\ \implies \mathbb{E}[\mathcal{NC}_1(\mathbf{H}_{GP})] &= \frac{\sum_{c=1}^{2}\frac{n_c\mu_c^2 + n_c\sigma_c^2}{N} - \frac{\mu_c^2}{2}}{\left(\sum_{c=1}^{2}\frac{\mu_c^2}{2} - \frac{n_c^2\mu_c^2}{N^2}\right)} + \Delta_{h.o.t}. \end{aligned} \tag{37}$$

D.2. NC1 of limiting NTK with `ReLU` activation

The recursive relationship between the NTK and NNGP Lee et al. (2019); Tirer et al. (2022) can be given as follows:

$$\Theta^{(2)}_{NTK-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = K^{(2)}_{GP-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) + K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})\dot{Q}^{(1)}_{GP-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) \tag{38}$$

Here, $K^{(2)}_{GP-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})$ can be defined using the following recursive formulation:

$$K^{(2)}_{GP-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = \sigma_w^2 Q^{(1)}_{GP-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}). \tag{39}$$

Based on (11), the derivative $\dot{Q}^{(1)}_{GP-ReLU}$ can be given as follows:

$$\dot{Q}^{(1)}_{GP-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = \frac{1}{2\pi}\left(\pi - \theta^{c',j}_{c,i}\right)$$
$$\theta^{c',j}_{c,i} = \arccos\left(\frac{K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})}{\sqrt{K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c,i})K^{(1)}_{GP}(\mathbf{x}^{c',j}, \mathbf{x}^{c',j})}}\right). \tag{40}$$

We build on the results from the NNGP analysis (with $Q^{(1)}_{GP-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})$) for computing the variability collapse with the limiting NTK. First, note that:

$$\theta^{c',j}_{c,i} = \begin{cases} 0 & \text{if } c = c' \\ \pi & \text{if } c \neq c' \end{cases}. \tag{41}$$

Thus, we get:

$$\Theta^{(2)}_{NTK-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = \sigma_w^2 Q^{(1)}_{GP-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) + K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})\dot{Q}^{(1)}_{GP-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}). \tag{42}$$

From (34), we know that:

$$Q^{(1)}_{GP-ReLU}(x^{c,i}, x^{c,j}) = \begin{cases} \frac{\sigma_w^2}{2} x^{c,i}x^{c',j} & \text{if } c = c' \\ 0 & \text{if } c \neq c' \end{cases} \tag{43}$$

$$\implies \Theta^{(2)}_{NTK-ReLU}(x^{c,i}, x^{c',j}) = \begin{cases} \frac{\sigma_w^4}{2} x^{c,i}x^{c,j} + \frac{\sigma_w^2}{2} x^{c,i}x^{c,j} & \text{if } c = c' \\ 0 & \text{if } c \neq c' \end{cases}, \tag{44}$$

$$= \begin{cases} \left(\frac{\sigma_w^4}{2} + \frac{\sigma_w^2}{2}\right) x^{c,i}x^{c,j} & \text{if } c = c' \\ 0 & \text{if } c \neq c' \end{cases}. \tag{45}$$

Notice that $\Theta^{(2)}_{NTK-ReLU}(x^{c,i}, x^{c',j})$ is a scaled version of $Q^{(1)}_{GP-ReLU}(x^{c,i}, x^{c',j})$ (as per (34)). Thus, we end up with the same result as (37) :

$$\mathbb{E}\left[\mathcal{NC}_1(\mathbf{H}_{NTK})\right] = \frac{\sum_{c=1}^{2} \frac{n_c\mu_c^2 + n_c\sigma_c^2}{N} - \frac{\mu_c^2}{2}}{\left(\sum_{c=1}^{2} \frac{\mu_c^2}{2} - \frac{n_c^2\mu_c^2}{N^2}\right)} + \Delta_{h.o.t}. \tag{46}$$

Appendix E. Results for NC1 with `Erf` activation

E.1. NC1 of Limiting NNGP with `Erf` activation

Under the Assumptions described in Section 3.2 with $d_0 = 1, d_1 \to \infty$, observe that:

$$Q^{(1)}_{GP-Erf}(x^{c,i}, x^{c',j}) = \frac{2}{\pi} \arcsin\left(\frac{2K^{(1)}_{GP}(x^{c,i}, x^{c',j})}{\sqrt{1 + 2K^{(1)}_{GP}(x^{c,i}, x^{c,i})}\sqrt{1 + 2K^{(1)}_{GP}(x^{c',j}, x^{c',j})}} \right) \tag{47}$$

$$= \frac{2}{\pi} \arcsin\left(\frac{2\sigma_w^2 x^{c,i} x^{c',j}}{\sqrt{1 + 2\sigma_w^2 (x^{c,i})^2}\sqrt{1 + 2\sigma_w^2 (x^{c',j})^2}} \right), \tag{48}$$

$$= \frac{2}{\pi} \arcsin\left(\frac{\operatorname{sign}(x^{c,i}) \operatorname{sign}(x^{c',j})}{\sqrt{1 + \frac{1}{2\sigma_w^2 (x^{c,i})^2}}\sqrt{1 + \frac{1}{2\sigma_w^2 (x^{c',j})^2}}} \right), \tag{49}$$

where the last equality comes from:

$$\frac{x^{c,i} x^{c',j}}{|x^{c,i}|\sqrt{1 + \frac{1}{2\sigma_w^2 (x^{c,i})^2}} \cdot |x^{c',j}|\sqrt{1 + \frac{1}{2\sigma_w^2 (x^{c',j})^2}}} = \frac{\operatorname{sign}(x^{c,i}) \operatorname{sign}(x^{c',j})}{\sqrt{1 + \frac{1}{2\sigma_w^2 (x^{c,i})^2}}\sqrt{1 + \frac{1}{2\sigma_w^2 (x^{c',j})^2}}}. \tag{50}$$

For notational simplicity, consider:

$$\rho(x^{c,i}, x^{c',j}) = \sqrt{1 + \frac{1}{2\sigma_w^2 (x^{c,i})^2}}\sqrt{1 + \frac{1}{2\sigma_w^2 (x^{c',j})^2}}, \tag{51}$$

and represent $Q^{(1)}_{GP-Erf}(x^{c,i}, x^{c',j})$ as:

$$Q^{(1)}_{GP-Erf}(x^{c,i}, x^{c',j}) = \frac{2}{\pi} \arcsin\left(\frac{\operatorname{sign}(x^{c,i}) \operatorname{sign}(x^{c',j})}{\rho(x^{c,i}, x^{c',j})} \right). \tag{52}$$

Based on Assumption 1, we know that $x^{1,i} < 0, x^{2,j} > 0$ almost surely. This leads to:

$$Q^{(1)}_{GP-Erf}(x^{c,i}, x^{c',j}) = \begin{cases} \frac{2}{\pi} \arcsin\left(\frac{1}{\rho(x^{c,i}, x^{c,j})} \right) & \text{if } c = c' \\ -\frac{2}{\pi} \arcsin\left(\frac{1}{\rho(x^{c,i}, x^{c',j})} \right) & \text{if } c \neq c' \end{cases}. \tag{53}$$

E.1.1. Calculating $\mathbb{E}\left[Q^{(1)}_{GP-Erf}(x^{c,i}, x^{c',j})\right]$

For $|u| \leq 1$, we consider the expansion of $\arcsin(u) = u + \frac{u^3}{6} + \cdots$ to obtain:

$$\mathbb{E}\left[\arcsin\left(\frac{1}{\rho(x^{c,i}, x^{c',j})} \right) \right] = \mathbb{E}\left[\frac{1}{\rho(x^{c,i}, x^{c',j})} \right] + \mathbb{E}\left[\frac{1}{6\rho(x^{c,i}, x^{c',j})^3} \right] + \cdots. \tag{54}$$

To this end, based on Assumption 1 of large enough $|\mu_1|, |\mu_2|$, we approximate the expectation with only the first term and denote $\xi_{h.o.t}$ to capture the effects of the higher

order terms. Notice that since $\rho(x^{c,i}, x^{c',j}) > 1$ for finite $(x^{c,i}, x^{c',j})$, the effects of $\xi_{h.o.t}$ are finite but decay rapidly compared to the first term. To this end, we get:

$$\mathbb{E}\left[\arcsin\left(\frac{1}{\rho(x^{c,i}, x^{c',j})}\right)\right] = \mathbb{E}\left[\frac{1}{\rho(x^{c,i}, x^{c',j})}\right] + \xi_{h.o.t} \tag{55}$$

Calculating the expectation $\mathbb{E}\left[\frac{1}{\rho(x^{c,i}, x^{c',j})}\right]$ can now be split based on c, c'.

• Case $c = c', i = j$:

$$\rho(x^{c,i}, x^{c,i}) = 1 + \frac{1}{2\sigma_w^2 (x^{c,i})^2} \tag{56}$$

$$\implies \mathbb{E}\left[\rho(x^{c,i}, x^{c,i})\right] = 1 + \frac{1}{2\sigma_w^2}\mathbb{E}\left[\frac{1}{(x^{c,i})^2}\right] \tag{57}$$

$$= 1 + \frac{T(c)}{2\sigma_w^2}. \tag{58}$$

The last equality is based on Lemma 7 which gives the expanded version of $T(c)$.

Finally, the value of $\mathbb{E}\left[\frac{1}{\rho(x^{c,i}, x^{c,i})}\right]$ can be given as:

$$\mathbb{E}\left[\frac{1}{\rho(x^{c,i}, x^{c,i})}\right] = \frac{1}{\mathbb{E}\left[\rho(x^{c,i}, x^{c,i})\right]} + \frac{Var(\rho(x^{c,i}, x^{c,i}))}{\mathbb{E}\left[\rho(x^{c,i}, x^{c,i})\right]^3} \tag{59}$$

$$= \frac{1}{1 + \frac{T(c)}{2\sigma_w^2}} + \delta_{h.o.t}(\rho(x^{c,i}, x^{c,i})) \tag{60}$$

Notice that even in this simple case, the expressions are non-trivial to fully expand. Nonetheless, along with Assumption 1, we consider large enough $|\mu_1|, |\mu_2|$ such that:

$$\frac{T(c)}{2\sigma_w^2} = \frac{1}{2\sigma_w^2}\left[\frac{1}{(\mu_c^2 + \sigma_c^2)} + \frac{2\sigma_c^4 + 4\sigma_c^2\mu_c^2}{(\mu_c^2 + \sigma_c^2)^3}\right] < 1. \tag{61}$$

Thus, based on the expansion of $(1+u)^{-1} = 1 - u + u^2 - u^3 + \cdots$, we obtain the following cleaner approximation of:

$$\mathbb{E}\left[\frac{1}{\rho(x^{c,i}, x^{c,i})}\right] = 1 - \frac{T(c)}{2\sigma_w^2} + \Delta^{(1)}_{h.o.t}(c). \tag{62}$$

Here $\Delta^{(1)}_{h.o.t}(c)$ captures all the higher order terms corresponding to $\left(\frac{T(c)}{2\sigma_w^2}\right)^2 - \left(\frac{T(c)}{2\sigma_w^2}\right)^3 + \cdots$ and $\delta_{h.o.t}(\rho(x^{c,i}, x^{c,i}))$ as denoted above.

• Case $c = c', i \neq j$:

In the case of $c = c', i \neq j$, the expectations on the square roots do not have a particular closed form. To this, end we leverage Assumption 1 to obtain the following approximation:

$$\rho(x^{c,i}, x^{c,j}) = \sqrt{1 + \frac{1}{2\sigma_w^2 (x^{c,i})^2}} \sqrt{1 + \frac{1}{2\sigma_w^2 (x^{c,j})^2}} \tag{63}$$

$$= \left(1 + \frac{1}{4\sigma_w^2 (x^{c,i})^2} + h.o.t\right) \left(1 + \frac{1}{4\sigma_w^2 (x^{c,j})^2} + h.o.t\right) \tag{64}$$

$$\implies \mathbb{E}\left[\rho(x^{c,i}, x^{c,j})\right] = \mathbb{E}\left[1 + \frac{1}{4\sigma_w^2 (x^{c,i})^2} + h.o.t\right] \mathbb{E}\left[1 + \frac{1}{4\sigma_w^2 (x^{c,j})^2} + h.o.t\right] \tag{65}$$

Observe that the inner terms in the expectations are scaled versions of the above case. To this end, we approximate $\mathbb{E}\left[\frac{1}{\rho(x^{c,i}, x^{c,j})}\right]$ as:

$$\mathbb{E}\left[\frac{1}{\rho(x^{c,i}, x^{c,j})}\right] \approx \frac{1}{\left(1 + \frac{T(c)}{4\sigma_w^2}\right)^2} + \delta_{h.o.t}(\rho(x^{c,i}, x^{c,j})) \tag{66}$$

$$= \frac{1}{1 + \frac{T(c)}{2\sigma_w^2} + \frac{T(c)^2}{16\sigma_w^4}} + \delta_{h.o.t}(\rho(x^{c,i}, x^{c,j})) \tag{67}$$

Similar to the assumption that led to (62), we get:

$$\mathbb{E}\left[\frac{1}{\rho(x^{c,i}, x^{c,j})}\right] \approx 1 - \frac{T(c)}{2\sigma_w^2} - \frac{T(c)^2}{16\sigma_w^4} + \Delta^{(2)}_{h.o.t}(c). \tag{68}$$

- **Case** $c \neq c'$

A similar analysis as above applies in this case:

$$\rho(x^{c,i}, x^{c',j}) = \sqrt{1 + \frac{1}{2\sigma_w^2 (x^{c,i})^2}} \sqrt{1 + \frac{1}{2\sigma_w^2 (x^{c',j})^2}} \tag{69}$$

$$= \left(1 + \frac{1}{4\sigma_w^2 (x^{c,i})^2} + h.o.t\right) \left(1 + \frac{1}{4\sigma_w^2 (x^{c',j})^2} + h.o.t\right) \tag{70}$$

$$\implies \mathbb{E}\left[\rho(x^{c,i}, x^{c',j})\right] = \mathbb{E}\left[1 + \frac{1}{4\sigma_w^2 (x^{c,i})^2} + h.o.t\right] \mathbb{E}\left[1 + \frac{1}{4\sigma_w^2 (x^{c',j})^2} + h.o.t\right] \tag{71}$$

Observe that the inner terms in the expectations are similar to the above case. To this end, we approximate $\mathbb{E}\left[\frac{1}{\rho(x^{c,i}, x^{c',j})}\right]$ as:

$$\mathbb{E}\left[\frac{1}{\rho(x^{c,i}, x^{c',j})}\right] \approx \frac{1}{\left(1 + \frac{T(c)}{4\sigma_w^2}\right)\left(1 + \frac{T(c')}{4\sigma_w^2}\right)} + \delta_{h.o.t}(\rho(x^{c,i}, x^{c',j})) \tag{72}$$

$$= \frac{1}{1 + \frac{T(c)+T(c')}{4\sigma_w^2} + \frac{T(c)T(c')}{16\sigma_w^4}} + \delta_{h.o.t}(\rho(x^{c,i}, x^{c',j})) \tag{73}$$

Similar to the assumption that led to (62), we get:

$$\mathbb{E}\left[\frac{1}{\rho(x^{c,i},x^{c',j})}\right] \approx 1 - \frac{T(c)+T(c')}{4\sigma_w^2} - \frac{T(c)T(c')}{16\sigma_w^4} + \Delta^{(3)}_{h.o.t}(c,c'). \tag{74}$$

Finally, based on (62), (68), (74) we obtain the following result for $\mathbb{E}[Q^{(1)}_{GP-Erf}(x^{c,i},x^{c',j})]$ as :

$$\begin{aligned}
&\mathbb{E}\left[Q^{(1)}_{GP-Erf}(x^{c,i},x^{c',j})\right] \\
&\qquad \approx \begin{cases} 1 - \frac{T(c)}{2\sigma_w^2} + \Delta^{(1)}_{h.o.t}(c) & \text{if } c = c', i = j \\ 1 - \frac{T(c)}{2\sigma_w^2} - \frac{T(c)^2}{16\sigma_w^4} + \Delta^{(2)}_{h.o.t}(c) & \text{if } c = c', i \neq j \\ 1 - \frac{T(c)+T(c')}{4\sigma_w^2} - \frac{T(c)T(c')}{16\sigma_w^4} + \Delta^{(3)}_{h.o.t}(c,c') & \text{if } c \neq c' \end{cases}
\end{aligned} \tag{75}$$

Here $\Delta^{(1)}_{h.o.t}(c), \Delta^{(2)}_{h.o.t}(c), \Delta^{(3)}_{h.o.t}(c,c')$ are the collective higher order terms that tend to 0 as $|\mu_c|$ increases relative to smaller values of σ_c. These cases can now be plugged into our generic formulation of expected values of a kernel function (i.e $V^{(1)}(c), V^{(2)}(c), V^{(3)}(c,c')$) as per (16) in Appendix C. Thus, based on Lemma 6 for sufficiently large $\{n_c\}$ we get :

$$\mathbb{E}[\mathcal{NC}_1(\mathbf{H})] = \frac{\sum_{c=1}^{2} \frac{n_c V^{(1)}(c)}{N} - \frac{V^{(2)}(c)}{2}}{\left[\sum_{c=1}^{2}\left(\frac{1}{2n_c^2} - \frac{1}{N^2}\right)\left(n_c^2 V^{(2)}(c)\right)\right] - \frac{2n_1n_2}{N^2}V^{(3)}(1,2)} + \Delta_{h.o.t} \tag{76}$$

• **Numerator in the balanced class setting.**

To better understand the result, let's consider the balanced class scenario with $n_1 = n_2 = N/2$, for which the numerator simplifies to:

$$\sum_{c=1}^{2} \frac{n_c V^{(1)}(c)}{N} - \frac{V^{(2)}(c)}{2} = \sum_{c=1}^{2} \frac{V^{(1)}(c) - V^{(2)}(c)}{2} \tag{77}$$

$$= \sum_{c=1}^{2} \frac{\frac{T(c)^2}{16\sigma_w^4} + \Delta^{(1)}_{h.o.t}(c) - \Delta^{(2)}_{h.o.t}(c)}{2}. \tag{78}$$

If we were to ignore the effects of the higher order terms, then observe that the numerator primarily depends on $T(c)^2$, which can be given based on Lemma 7 as:

$$T(c)^2 = \left[\frac{1}{(\mu_c^2+\sigma_c^2)} + \frac{2\sigma_c^4 + 4\sigma_c^2\mu_c^2}{(\mu_c^2+\sigma_c^2)^3}\right]^2 \tag{79}$$

Thus, showcasing the dependence on μ_c, σ_c in determining the extent of collapse. For sufficiently large $|\mu_c| \gg \sigma_c$, we can approximate this value to:

$$T(c)^2 \approx \left[\frac{1}{\mu_c^2} + \frac{4\sigma_c^2}{\mu_c^4}\right]^2 = \frac{1}{\mu_c^4}\left[1 + \frac{4\sigma_c^2}{\mu_c^2}\right]^2 = \frac{1}{\mu_c^4}\left[1 + \frac{8\sigma_c^2}{\mu_c^2} + \frac{16\sigma_c^4}{\mu_c^4}\right] \tag{80}$$

• **Denominator in the balanced class setting.**

Similar to the numerator analysis, observe that when $n_1 = n_2 = N/2$, the denominator can be given as:

$$\left[\sum_{c=1}^{2}\left(\frac{1}{2n_c^2}-\frac{1}{N^2}\right)\left(n_c^2 V^{(2)}(c)\right)\right]-\frac{2n_1n_2}{N^2}V^{(3)}(1,2) \tag{81}$$

$$=\frac{V^{(2)}(1)+V^{(2)}(2)-2V^{(3)}(1,2)}{4} \tag{82}$$

$$=\frac{-\frac{T(1)}{2\sigma_w^2}-\frac{T(1)^2}{16\sigma_w^4}+\Delta^{(2)}_{h.o.t}(1)-\frac{T(2)}{2\sigma_w^2}-\frac{T(2)^2}{16\sigma_w^4}+\Delta^{(2)}_{h.o.t}(2)}{4} \tag{83}$$

$$+\frac{2\frac{T(1)+T(2)}{4\sigma_w^2}+2\frac{T(1)T(2)}{16\sigma_w^4}-2\Delta^{(3)}_{h.o.t}(1,2)}{4} \tag{84}$$

$$=\frac{-\frac{T(1)^2}{16\sigma_w^4}-\frac{T(2)^2}{16\sigma_w^4}+2\frac{T(1)T(2)}{16\sigma_w^4}+\Delta^{(2)}_{h.o.t}(1)+\Delta^{(2)}_{h.o.t}(2)-2\Delta^{(3)}_{h.o.t}(1,2)}{4} \tag{85}$$

$$=\frac{-\left(\frac{T(1)-T(2)}{4\sigma_w^2}\right)^2+\Delta^{(2)}_{h.o.t}(1)+\Delta^{(2)}_{h.o.t}(2)-2\Delta^{(3)}_{h.o.t}(1,2)}{4}. \tag{86}$$

Observe that the term $T(1)-T(2)$ represents:

$$T(1)-T(2)=\left[\frac{1}{(\mu_1^2+\sigma_1^2)}+\frac{2\sigma_1^4+4\sigma_1^2\mu_1^2}{(\mu_1^2+\sigma_1^2)^3}\right]-\left[\frac{1}{(\mu_2^2+\sigma_2^2)}+\frac{2\sigma_2^4+4\sigma_2^2\mu_2^2}{(\mu_2^2+\sigma_2^2)^3}\right] \tag{87}$$

and for sufficiently large $|\mu_c| \gg \sigma_c$, essentially represents:

$$T(1)-T(2)\approx\frac{1}{\mu_1^2}+\frac{4\sigma_1^2}{\mu_1^4}-\frac{1}{\mu_2^2}-\frac{4\sigma_2^2}{\mu_2^4}. \tag{88}$$

E.2. NC1 of Limiting NTK with `Erf` activation

Recall that the recursive relationship between the NTK and NNGP can be given as follows:

$$\Theta^{(2)}_{NTK-Erf}(x^{c,i},x^{c',j})=K^{(2)}_{GP-Erf}(x^{c,i},x^{c',j})+K^{(1)}_{GP}(x^{c,i},x^{c',j})\dot{Q}^{(1)}_{GP-Erf}(x^{c,i},x^{c',j}), \tag{89}$$

where:

$$K^{(2)}_{GP-Erf}(x^{c,i},x^{c',j})=\sigma_w^2 Q^{(1)}_{GP-Erf}(x^{c,i},x^{c',j}) \tag{90}$$

$$Q^{(1)}_{GP-Erf}(x^{c,i},x^{c',j})=\frac{2}{\pi}\arcsin\left(\frac{2K^{(1)}_{GP}(x^{c,i},x^{c',j})}{\sqrt{1+2K^{(1)}_{GP}(x^{c,i},x^{c,i})}\sqrt{1+2K^{(1)}_{GP}(x^{c',j},x^{c',j})}}\right) \tag{91}$$

$$\dot{Q}^{(1)}_{GP-Erf}(x^{c,i},x^{c',j})=\frac{4}{\pi}\left[\left(1+2K^{(1)}_{GP}(x^{c,i},x^{c,i})\right)\left(1+2K^{(1)}_{GP}(x^{c',j},x^{c',j})\right)-\left(2K^{(1)}_{GP}(x^{c,i},x^{c',j})\right)^2\right]^{-1/2} \tag{92}$$

Considering $d_0 = 1$ (as per the setting and assumptions), we get:

$$K^{(1)}_{GP}(\mathbf{x}^{c,i},\mathbf{x}^{c',j})=\sigma_w^2 x^{c,i}x^{c',j} \tag{93}$$

$$Q^{(1)}_{GP-Erf}(x^{c,i}, x^{c',j}) = \frac{2}{\pi}\arcsin\left(\frac{2\sigma_w^2 x^{c,i}x^{c',j}}{\sqrt{1+2\sigma_w^2(x^{c,i})^2}\sqrt{1+2\sigma_w^2(x^{c',j})^2}}\right). \tag{94}$$

$$\begin{aligned}\dot{Q}^{(1)}_{GP-Erf}(x^{c,i}, x^{c',j}) &= \frac{4}{\pi}\left(\left(1+2\sigma_w^2 x^{c,i}x^{c,i}\right)\left(1+2\sigma_w^2 x^{c',j}x^{c',j}\right) - \left(2\sigma_w^2 x^{c,i}x^{c',j}\right)^2\right)^{-1/2} \\ &= \frac{4}{\pi\sqrt{1+2\sigma_w^2\cdot(x^{c,i})^2 + 2\sigma_w^2\cdot(x^{c',j})^2}}.\end{aligned} \tag{95}$$

This gives us:

$$K^{(1)}_{GP}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j})\dot{Q}^{(1)}_{GP-Erf}(x^{c,i}, x^{c',j}) = \frac{4\sigma_w^2 x^{c,i}x^{c',j}}{\pi\sqrt{1+2\sigma_w^2\cdot(x^{c,i})^2 + 2\sigma_w^2\cdot(x^{c',j})^2}} \tag{96}$$

$$= \frac{4\sigma_w^2 x^{c,i}x^{c',j}}{\pi\sigma_w|x^{c,i}||x^{c',j}|\sqrt{\frac{1}{\sigma_w^2(x^{c,i})^2(x^{c',j})^2} + \frac{2}{(x^{c',j})^2} + \frac{2}{(x^{c,i})^2}}} \tag{97}$$

$$= \frac{4\sigma_w \operatorname{sign}(x^{c,i})\operatorname{sign}(x^{c',j})}{\pi\sqrt{\frac{1}{\sigma_w^2(x^{c,i})^2(x^{c',j})^2} + \frac{2}{(x^{c',j})^2} + \frac{2}{(x^{c,i})^2}}} \tag{98}$$

For notational simplicity, consider:

$$\kappa(x^{c,i}, x^{c',j}) = \sqrt{\frac{1}{\sigma_w^2(x^{c,i})^2(x^{c',j})^2} + \frac{2}{(x^{c',j})^2} + \frac{2}{(x^{c,i})^2}} \tag{99}$$

which simplifies the kernel formulation to:

$$\Theta^{(2)}_{NTK-Erf}(x^{c,i}, x^{c',j}) = \sigma_w^2 Q^{(1)}_{GP-Erf}(x^{c,i}, x^{c',j}) + \frac{4\sigma_w \operatorname{sign}(x^{c,i})\operatorname{sign}(x^{c',j})}{\pi\kappa(x^{c,i}, x^{c',j})} \tag{100}$$

E.2.1. Calculating $\mathbb{E}\left[\Theta^{(2)}_{NTK-Erf}(x^{c,i}, x^{c',j})\right]$

Similar to the NNGP analysis, we break down the calculation of $\mathbb{E}\left[\kappa(x^{c,i}, x^{c',j})\right]$ into three cases.

- **Case** $c = c', i = j$

$$\kappa(x^{c,i}, x^{c,i}) = \sqrt{\frac{1}{\sigma_w^2(x^{c,i})^4} + \frac{4}{(x^{c,i})^2}} = \sqrt{1 - \left(1 - \frac{1}{\sigma_w^2(x^{c,i})^4} - \frac{4}{(x^{c,i})^2}\right)} \tag{101}$$

$$= 1 - \frac{1}{2}\left(1 - \frac{1}{\sigma_w^2(x^{c,i})^4} - \frac{4}{(x^{c,i})^2}\right) + \xi_{h.o.t} \tag{102}$$

$$= \frac{1}{2} + \frac{1}{2\sigma_w^2(x^{c,i})^4} + \frac{2}{(x^{c,i})^2} + \xi_{h.o.t} \tag{103}$$

This gives us:

$$\begin{aligned}\mathbb{E}\left[\kappa(x^{c,i},x^{c,i})\right] &= \frac{1}{2} + \mathbb{E}\left[\frac{1}{2\sigma_w^2(x^{c,i})^4}\right] + \mathbb{E}\left[\frac{2}{(x^{c,i})^2}\right] + \mathbb{E}[\xi_{h.o.t}] \\ &= \frac{1}{2} + \frac{1}{2\sigma_w^2}\left[\frac{1}{\mathbb{E}\left[(x^{c,i})^4\right]} + \frac{Var((x^{c,i})^4)}{\mathbb{E}\left[(x^{c,i})^4\right]^3}\right] + 2\left[\frac{1}{\mathbb{E}\left[(x^{c,i})^2\right]} + \frac{Var((x^{c,i})^2)}{\mathbb{E}\left[(x^{c,i})^2\right]^3}\right] \\ &\quad + \mathbb{E}[\xi_{h.o.t}]\end{aligned} \tag{104}$$

Based on the results from the moment-generating function, we know that:

$$\mathbb{E}[(x^{c,i})^4] = 3\sigma_c^4 + 6\sigma_c^2\mu_c^2 + \mu_c^4, \tag{105}$$

which can be used along with Lemma 7 to obtain:

$$\mathbb{E}\left[\kappa(x^{c,i},x^{c,i})\right] = \frac{1}{2} + 2T(c) + \mathbb{E}[\xi_{h.o.t}] \tag{106}$$

For notational simplicity, we define a helper function as follows:

$$S(\mu_c,\sigma_c) = -\frac{1}{2} + 2T(c) + \mathbb{E}[\xi_{h.o.t}], \tag{107}$$

which gives us:

$$\mathbb{E}\left[\kappa(x^{c,i},x^{c,i})\right] = 1 + S(\mu_c,\sigma_c) \tag{108}$$

Finally, the value of $\mathbb{E}\left[\frac{1}{\kappa(x^{c,i},x^{c,i})}\right]$ can be given as:

$$\mathbb{E}\left[\frac{1}{\kappa(x^{c,i},x^{c,i})}\right] = \frac{1}{\mathbb{E}\left[\kappa(x^{c,i},x^{c,i})\right]} + \frac{Var(\kappa(x^{c,i},x^{c,i}))}{\mathbb{E}\left[\kappa(x^{c,i},x^{c,i})\right]^3} \tag{109}$$

$$= \frac{1}{1+S(\mu_c,\sigma_c)} + \delta_{h.o.t}(\kappa(x^{c,i},x^{c,i})) \tag{110}$$

Notice that even in this simple case, the expressions are non-trivial to fully expand. Nonetheless, along with Assumption 1, we consider large enough $|\mu_1|, |\mu_2|$ such that:

$$S(\mu_c,\sigma_c) < 1. \tag{111}$$

Thus, based on the expansion of $(1+u)^{-1} = 1 - u + u^2 - u^3 + \cdots$, we obtain the following cleaner approximation of:

$$\mathbb{E}\left[\frac{1}{\kappa(x^{c,i},x^{c,i})}\right] = 1 - S(\mu_c,\sigma_c) + \widetilde{\delta}_{h.o.t}(\kappa(x^{c,i},x^{c,i})) \tag{112}$$

- **Case** $c = c', i \neq j$**:**

$$\kappa(x^{c,i}, x^{c,j}) = \sqrt{\frac{1}{\sigma_w^2 (x^{c,i})^2 (x^{c,j})^2} + \frac{2}{(x^{c,j})^2} + \frac{2}{(x^{c,i})^2}} \tag{113}$$

$$= \sqrt{1 - \left(1 - \frac{1}{\sigma_w^2 (x^{c,i})^2 (x^{c,j})^2} - \frac{2}{(x^{c,j})^2} - \frac{2}{(x^{c,i})^2}\right)} \tag{114}$$

$$= 1 - \frac{1}{2}\left(1 - \frac{1}{\sigma_w^2 (x^{c,i})^2 (x^{c,j})^2} - \frac{2}{(x^{c,j})^2} - \frac{2}{(x^{c,i})^2}\right) + \xi'_{h.o.t} \tag{115}$$

$$= \frac{1}{2} + \frac{1}{2\sigma_w^2 (x^{c,i})^2 (x^{c,j})^2} + \frac{1}{(x^{c,j})^2} + \frac{1}{(x^{c,i})^2} + \xi'_{h.o.t} \tag{116}$$

Thus, based on Lemma 7, we get:

$$\mathbb{E}\left[\kappa(x^{c,i}, x^{c,j})\right] = \frac{1}{2} + \mathbb{E}\left[\frac{1}{2\sigma_w^2 (x^{c,i})^2 (x^{c,j})^2}\right] + \mathbb{E}\left[\frac{1}{(x^{c,j})^2}\right] + \mathbb{E}\left[\frac{1}{(x^{c,i})^2}\right] + \mathbb{E}[\xi'_{h.o.t}] \tag{117}$$

$$= \frac{1}{2} + \frac{1}{2\sigma_w^2}\mathbb{E}\left[\frac{1}{(x^{c,i})^2}\right]\mathbb{E}\left[\frac{1}{(x^{c,j})^2}\right] + \mathbb{E}\left[\frac{1}{(x^{c,j})^2}\right] + \mathbb{E}\left[\frac{1}{(x^{c,i})^2}\right] + \mathbb{E}[\xi'_{h.o.t}] \tag{118}$$

$$= \frac{1}{2} + \frac{T(c)^2}{2\sigma_w^2} + 2T(c) + \mathbb{E}[\xi'_{h.o.t}] \tag{119}$$

This leads to:

$$\mathbb{E}\left[\frac{1}{\kappa(x^{c,i}, x^{c,j})}\right] = \mathbb{E}\left[\frac{1}{1 + \left(-\frac{1}{2} + \frac{T(c)^2}{2\sigma_w^2} + 2T(c) + \mathbb{E}[\xi'_{h.o.t}]\right)}\right] \tag{120}$$

$$= 1 - \left(-\frac{1}{2} + \frac{T(c)^2}{2\sigma_w^2} + 2T(c) + \mathbb{E}[\xi'_{h.o.t}]\right) + \delta'_{h.o.t}(\kappa(x^{c,i}, x^{c,j})) \tag{121}$$

$$= \frac{3}{2} - \frac{T(c)^2}{2\sigma_w^2} - 2T(c) + \widetilde{\delta}_{h.o.t}(\kappa(x^{c,i}, x^{c,j})) \tag{122}$$

- **Case $c \neq c'$:**

$$\kappa(x^{c,i}, x^{c',j}) = \sqrt{\frac{1}{\sigma_w^2 (x^{c,i})^2 (x^{c',j})^2} + \frac{2}{(x^{c',j})^2} + \frac{2}{(x^{c,i})^2}} \tag{123}$$

$$= \frac{1}{2} + \frac{1}{2\sigma_w^2 (x^{c,i})^2 (x^{c',j})^2} + \frac{1}{(x^{c',j})^2} + \frac{1}{(x^{c,i})^2} + \xi''_{h.o.t} \tag{124}$$

Thus, based on Lemma 7, we get:

$$\mathbb{E}\left[\kappa(x^{c,i}, x^{c',j})\right] = \frac{1}{2} + \mathbb{E}\left[\frac{1}{2\sigma_w^2 (x^{c,i})^2 (x^{c',j})^2}\right] + \mathbb{E}\left[\frac{1}{(x^{c',j})^2}\right] + \mathbb{E}\left[\frac{1}{(x^{c,i})^2}\right] + \mathbb{E}[\xi''_{h.o.t}] \tag{125}$$

$$= \frac{1}{2} + \frac{1}{2\sigma_w^2}\mathbb{E}\left[\frac{1}{(x^{c,i})^2}\right]\mathbb{E}\left[\frac{1}{(x^{c',j})^2}\right] + \mathbb{E}\left[\frac{1}{(x^{c',j})^2}\right] + \mathbb{E}\left[\frac{1}{(x^{c,i})^2}\right] + \mathbb{E}[\xi''_{h.o.t}] \tag{126}$$

$$= \frac{1}{2} + \frac{T(c)T(c')}{2\sigma_w^2} + T(c') + T(c) + \mathbb{E}[\xi''_{h.o.t}]. \tag{127}$$

This gives us:

$$\mathbb{E}\left[\frac{1}{\kappa(x^{c,i},x^{c',j})}\right] = \mathbb{E}\left[\frac{1}{1+\left(-\frac{1}{2}+\frac{T(c)T(c')}{2\sigma_w^2}+T(c')+T(c)+\mathbb{E}[\xi''_{h.o.t}]\right)}\right] \tag{128}$$

$$= 1-\left(-\frac{1}{2}+\frac{T(c)T(c')}{2\sigma_w^2}+T(c')+T(c)+\mathbb{E}[\xi''_{h.o.t}]\right)+\delta'_{h.o.t}(\kappa(x^{c,i},x^{c,j})) \tag{129}$$

$$= \frac{3}{2}-\frac{T(c)T(c')}{2\sigma_w^2}-T(c)-T(c')+\widetilde{\delta}_{h.o.t}(\kappa(x^{c,i},x^{c',j})) \tag{130}$$

Finally, the cases for the expected value of the kernel can be given as:

$$\mathbb{E}\left[\Theta^{(2)}_{NTK-Erf}(x^{c,i},x^{c',j})\right] = \begin{cases}\mathbb{E}\left[\sigma_w^2 Q^{(1)}_{GP-Erf}(x^{c,i},x^{c,j})\right]+\mathbb{E}\left[\frac{4\sigma_w}{\pi\kappa(x^{c,i},x^{c,j})}\right] & c=c'\\ \mathbb{E}\left[\sigma_w^2 Q^{(1)}_{GP-Erf}(x^{c,i},x^{c',j})\right]-\mathbb{E}\left[\frac{4\sigma_w}{\pi\kappa(x^{c,i},x^{c',j})}\right] & c\neq c'\end{cases}, \tag{131}$$

From (75), we know that:

$$\mathbb{E}\left[Q^{(1)}_{GP-Erf}(x^{c,i},x^{c',j})\right] \approx \begin{cases}1-\frac{T(c)}{2\sigma_w^2}+\Delta^{(1)}_{h.o.t}(c) & \text{if } c=c', i=j\\ 1-\frac{T(c)}{2\sigma_w^2}-\frac{T(c)^2}{16\sigma_w^4}+\Delta^{(2)}_{h.o.t}(c) & \text{if } c=c', i\neq j\\ 1-\frac{T(c)+T(c')}{4\sigma_w^2}-\frac{T(c)T(c')}{16\sigma_w^4}+\Delta^{(3)}_{h.o.t}(c,c') & \text{if } c\neq c'\end{cases}. \tag{132}$$

To simplify the presentation, we can ignore the higher-order terms and obtain:

$$\mathbb{E}\left[\Theta^{(2)}_{NTK-Erf}(x^{c,i},x^{c',j})\right] \tag{133}$$

$$\approx \begin{cases}\sigma_w^2\left(1-\frac{T(c)}{2\sigma_w^2}\right)+\frac{4\sigma_w}{\pi}\left(\frac{3}{2}-2T(c)\right) & c=c'; i=j\\ \sigma_w^2\left(1-\frac{T(c)}{2\sigma_w^2}-\frac{T(c)^2}{16\sigma_w^4}\right)+\frac{4\sigma_w}{\pi}\left(\frac{3}{2}-\frac{T(c)^2}{2\sigma_w^2}-2T(c)\right), & c=c', i\neq j\\ \sigma_w^2\left(1-\frac{T(c)+T(c')}{4\sigma_w^2}-\frac{T(c)T(c')}{16\sigma_w^4}\right)-\frac{4\sigma_w}{\pi}\left(\frac{3}{2}-\frac{T(c)T(c')}{2\sigma_w^2}-T(c)-T(c')\right), & c\neq c'\end{cases} \tag{134}$$

Observe that the order of the $T(c)$ terms involved here resemble that of the NNGP scenario in (75). Thus, we can make similar conclusions regarding the role of the order of μ_c, σ_c in determining the value of $\mathbb{E}\left[\mathcal{NC}_1(\mathbf{H})\right]$.

Appendix F. Activation Variability Relative to Data

In this section, we introduce a relative measure of activation variability collapse with respect to the data. First, we begin by defining the within-class and between-class data covariance matrices $\mathbf{\Sigma}_W(\mathbf{X}), \mathbf{\Sigma}_B(\mathbf{X}) \in \mathbb{R}^{d_0\times d_0}$ for the data samples as:

$$\mathbf{\Sigma}_W(\mathbf{X}) = \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{n_c}\left(\mathbf{x}^{c,i}-\overline{\mathbf{x}}^c\right)\left(\mathbf{x}^{c,i}-\overline{\mathbf{x}}^c\right)^\top; \quad \mathbf{\Sigma}_B(\mathbf{X}) = \frac{1}{C}\sum_{c=1}^{C}\left(\overline{\mathbf{x}}^c-\overline{\mathbf{x}}^G\right)\left(\overline{\mathbf{x}}^c-\overline{\mathbf{x}}^G\right)^\top, \tag{135}$$

where $\overline{\mathbf{x}}^c = \frac{1}{n_c}\sum_{i=1}^{n_c}\mathbf{x}^{c,i}, \forall c \in [C]$ and $\overline{\mathbf{x}}^G = \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{n_c}\mathbf{x}^{c,i}$ represent the data class mean vectors and the data global mean vector respectively.

Definition 8 *Set a small $\tau > 0$. The variability collapse relative to the data is given by:*

$$\mathcal{NC}_1(\mathbf{H}|\mathbf{X}) := \frac{\mathcal{NC}_1(\mathbf{H})}{\mathcal{NC}_1(\mathbf{X}) + \tau}, \quad \textit{where } \mathcal{NC}_1(\mathbf{X}) := \frac{\mathrm{tr}(\mathbf{\Sigma}_W(\mathbf{X}))}{\mathrm{tr}(\mathbf{\Sigma}_B(\mathbf{X}))} \tag{136}$$

The constant τ prevents numerical instabilities. Through this approach, we capture the extent of variability collapse of activation features relative to the variability collapse of the data samples itself.

Corollary 9 *Under Assumptions 1-2 (as per Section 5.3 in main text), let $\phi(\cdot)$ be the ReLU activation, and the limiting NNGP kernel be $Q^{(1)}_{GP-ReLU}(\mathbf{x}^{c,i}, \mathbf{x}^{c',j}) = \mathbf{h}^{c,i\top}\mathbf{h}^{c',j}$, then:*

$$\frac{\mathbb{E}\left[\mathcal{NC}_1(\mathbf{H})\right]}{\mathbb{E}\left[\mathcal{NC}_1(\mathbf{X})\right]} \approx 1 - \frac{\frac{2}{N^2}\prod_{c=1}^{2} n_c\mu_c}{\left(\sum_{c=1}^{2}\frac{\mu_c^2}{2} - \frac{n_c^2\mu_c^2}{N^2}\right)} \tag{137}$$

Proof *To keep the derivation similar to those for the kernel formulation in equation 34, we consider a simplified kernel on $\mathbf{X}$ (identity feature map):*

$$K_{data}(x^{c,i}, x^{c',j}) = x^{c,i}x^{c',j}. \tag{138}$$

Additionally, since $\mathbf{x}^{c,i}$ are 1-d random variables, the expected value of the kernel is:

$$\mathbb{E}\left[K_{data}(x^{c,i}, x^{c',j})\right] = \begin{cases} \sigma_c^2 + \mu_c^2 & \text{if } c = c', i = j \\ \mu_c^2 & \text{if } c = c', i \neq j \\ \mu_c\mu_{c'} & \text{if } c \neq c' \end{cases} \tag{139}$$

We use Lemma 6 with cases $V^{(1)}(c) = \sigma_c^2 + \mu_c^2$, $V^{(2)}(c) = \mu_c^2$ and $V^{(3)}(c, c') = \mu_c\mu_{c'}$ to obtain:

$$\mathbb{E}\left[\mathcal{NC}_1(\mathbf{X})\right] = \frac{\mathbb{E}\left[\mathrm{tr}(\Sigma_W(\mathbf{X}))\right]}{\mathbb{E}\left[\mathrm{tr}(\Sigma_B(\mathbf{X}))\right]} = \frac{\sum_{c=1}^{2}\frac{n_c\mu_c^2 + n_c\sigma_c^2}{N} - \frac{n_c^2\mu_c^2 + n_c\sigma_c^2}{2n_c^2}}{\left(\sum_{c=1}^{2}\frac{n_c^2\mu_c^2 + n_c\sigma_c^2}{2n_c^2} - \frac{n_c^2\mu_c^2 + n_c\sigma_c^2}{N^2}\right) - \frac{2}{N^2}\prod_{c=1}^{2} n_c\mu_c} + \Delta^X_{h.o.t} \tag{140}$$

Finally, the ratio $\frac{\mathbb{E}\left[\mathcal{NC}_1(\mathbf{H})\right]}{\mathbb{E}\left[\mathcal{NC}_1(\mathbf{X})\right]}$ for `ReLU` (Theorem 2) with large enough $n_c \gg 1$ is:

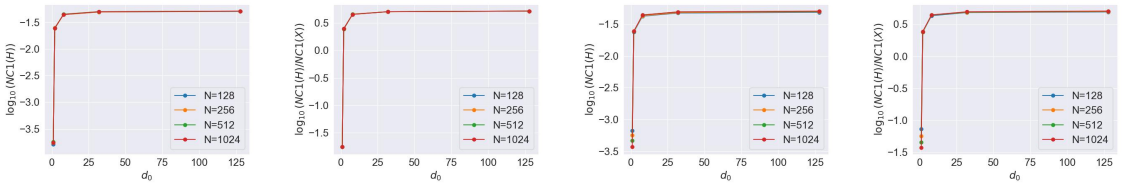
$$\frac{\mathbb{E}\left[\mathcal{NC}_1(\mathbf{H})\right]}{\mathbb{E}\left[\mathcal{NC}_1(\mathbf{X})\right]} = \frac{\sum_{c=1}^{2}\frac{n_c\mu_c^2 + n_c\sigma_c^2}{N} - \frac{\mu_c^2}{2}}{\left(\sum_{c=1}^{2}\frac{\mu_c^2}{2} - \frac{n_c^2\mu_c^2}{N^2}\right)} \cdot \frac{\left(\sum_{c=1}^{2}\frac{\mu_c^2}{2} - \frac{n_c^2\mu_c^2}{N^2}\right) - \frac{2}{N^2}\prod_{c=1}^{2} n_c\mu_c}{\sum_{c=1}^{2}\frac{n_c\mu_c^2 + n_c\sigma_c^2}{N} - \frac{\mu_c^2}{2}} + \Delta'_{h.o.t}. \tag{141}$$

$$= \frac{\left(\sum_{c=1}^{2}\frac{\mu_c^2}{2} - \frac{n_c^2\mu_c^2}{N^2}\right) - \frac{2}{N^2}\prod_{c=1}^{2} n_c\mu_c}{\left(\sum_{c=1}^{2}\frac{\mu_c^2}{2} - \frac{n_c^2\mu_c^2}{N^2}\right)} + \Delta'_{h.o.t} \tag{142}$$

$$= 1 - \frac{\frac{2}{N^2}\prod_{c=1}^{2} n_c\mu_c}{\left(\sum_{c=1}^{2}\frac{\mu_c^2}{2} - \frac{n_c^2\mu_c^2}{N^2}\right)} + \Delta'_{h.o.t} \tag{143}$$

To better understand the result, let us consider the balanced class scenario where $n_1 = n_2 = n = N/2$. This results in a ratio of $\approx 1 - (2\mu_1\mu_2)/(\mu_1^2 + \mu_2^2)$. Furthermore, if $|\mu_1| = |\mu_2|$ (so $\mu_1 = -\mu_2$), then the ratio ≈ 2. Thus, it emphasizes the interplay between class imbalance/balance and the expected class means on the relative variability collapse.

Addressing misleading $\mathcal{NC}_1(\mathbf{H})$ values. Consider the case where $\sigma_1, \sigma_2 \to 0$. Then Theorem 5.1 for $Q_{GP-ReLU}$ indicates that $\mathbb{E}\left[\mathcal{NC}_1(\mathbf{H})\right] \to 0$ (considering smaller fluctuations from $\Delta_{h.o.t}$) in the balanced class setting. Such an observation can be misleading if one were to ignore $\mathcal{NC}_1(\mathbf{X})$. For instance, such an empirical result while training deep neural networks fails to differentiate between settings where the network learned meaningful features and learned to classify complex datasets or was simply able to leverage the already collapsed data vectors. This applies to `Erf` activation as well. We justify this argument with the following experiment. For a sample size N chosen from $\{128, 256, 512, 1024\}$, and input dimension d_0 chosen from $\{1, 2, 8, 32, 128\}$, we sample the vectors $\mathbf{x}^{1,i} \sim \mathcal{N}(-10 * \mathbf{1}_{d_0}, \mathbf{I}_{d_0}), i \in [N/2]$ for class 1 and $\mathbf{x}^{2,j} \sim \mathcal{N}(10 * \mathbf{1}_{d_0}, \mathbf{I}_{d_0}), j \in [N/2]$ for class 2 as our dataset. From Figure 2(a), 2(b), observe that $\mathcal{NC}_1(\mathbf{H}|\mathbf{X})$ values for Q_{GP-Erf} can be orders of magnitude larger than $\mathcal{NC}_1(\mathbf{H})$, and for high-dimensions $\mathcal{NC}_1(\mathbf{H}|\mathbf{X}) > 1$. Essentially, the raw data is 'more' collapsed than the activations in these settings. Similar observations can be made for the NTK $\Theta_{NTK-Erf}$ in Figure 2(c), 2(d).

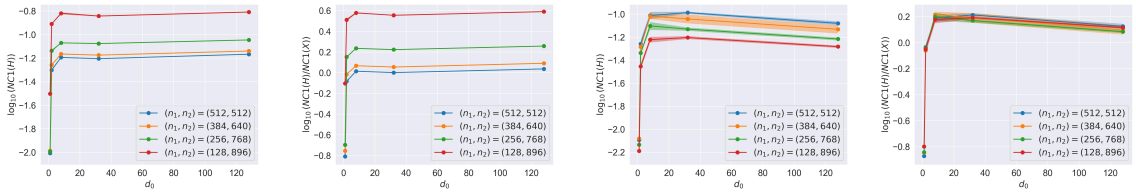


(a) Q_{GP},$\mathcal{NC}_1(\mathbf{H})$ (b) Q_{GP}, $\mathcal{NC}_1(\mathbf{H}|\mathbf{X})$ (c) Θ_{NTK}, $\mathcal{NC}_1(\mathbf{H})$ (d) Θ_{NTK}, $\mathcal{NC}_1(\mathbf{H}|\mathbf{X})$

Figure 2: $\mathcal{NC}_1(\mathbf{H}), \mathcal{NC}_1(\mathbf{H}|\mathbf{X})$ of $Q^{(1)}_{GP-Erf}$ and $\Theta^{(2)}_{NTK-Erf}$. The dimension d_0 on the x-axis is chosen from $\{1, 2, 8, 32, 128\}$. For a particular N, we sample the vectors $\mathbf{x}^{1,i} \sim \mathcal{N}(-10 * \mathbf{1}_{d_0}, \mathbf{I}_{d_0}), y^{1,i} = -1, i \in [N/2]$ for class 1 and $\mathbf{x}^{2,j} \sim \mathcal{N}(10 * \mathbf{1}_{d_0}, \mathbf{I}_{d_0}), y^{2,j} = 1, j \in [N/2]$ for class 2.

Appendix G. Numerical solutions of EoS

We solve the EoS using the Newton-Krylov method with an annealing schedule (as originally proposed by Seroussi et al. (2023)) using the `scipy.optimize.newton_krylov` python API. We initialize $\mathbf{C}$ with the GP limit value of $(\sigma_w^2/d_0)\mathbf{I}_{d_0}$ and choose a large annealing factor (ex: 10^5) as the value for d_1. The result of optimizing with `newton_krylov` is a new $\mathbf{C}$, which in addition to a lower annealing factor is used as an input for the next `newton_krylov` function call. This loop is repeated until the end of an annealing schedule. For instance,



(a) EoS $\mathcal{NC}_1(\mathbf{H})$ (b) EoS $\mathcal{NC}_1(\mathbf{H}|\mathbf{X})$ (c) 2L-FCN $\mathcal{NC}_1(\mathbf{H})$ (d) 2L-FCN $\mathcal{NC}_1(\mathbf{H}|\mathbf{X})$

Figure 3: $\mathcal{NC}_1(\mathbf{H}), \mathcal{NC}_1(\mathbf{H}|\mathbf{X})$ of the adaptive kernel (EoS) with final annealing factor $d_1 = 500$ and 2L-FCN with $d_1 = 500$ and `Erf` activation. The dimension d_0 on the x-axis is chosen from $\{1, 2, 8, 32, 128\}$. For a tuple (n_1, n_2) such that $n_1 + n_2 = N = 1024$, we sample the vectors $\mathbf{x}^{1,i} \sim \mathcal{N}(-2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{1,i} = -1, i \in [n_1]$ for class 1 and $\mathbf{x}^{2,j} \sim \mathcal{N}(2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{2,j} = 1, j \in [n_2]$ for class 2.

to analyze the EoS corresponding to $d_1 = 500$, we choose the following list of step-wise annealing factors:

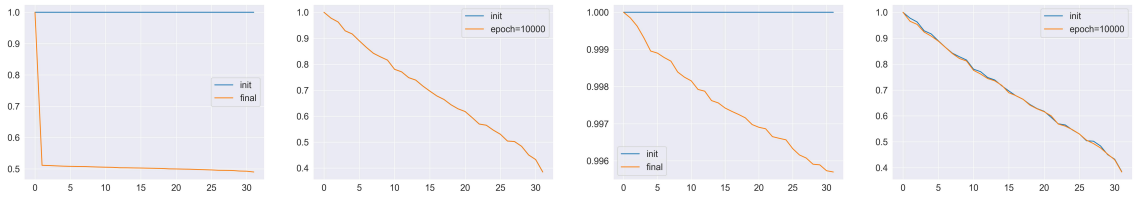
$$\texttt{factors} = [\underbrace{10^5, 9 * 10^4, \cdots, 2 * 10^4}_{\texttt{step}=-10^4}, \underbrace{10^4, 9 * 10^3, \cdots, 2 * 10^3}_{\texttt{step}=-10^3}, \underbrace{10^3, \cdots, 500}_{\texttt{step}=-10^2}]. \tag{144}$$

Similarly, for a choice of $d_1 = 2000$, we select the slice of the above list up to 2000. Selecting the schedule is a manual operation and can be treated as a hyper-parameter. In our experiments, we observed that this schedule is sufficient to obtain insights on the NC1 metrics of $\mathbf{Q}^{(1)}$. Thus, we leave the exploration of various annealing strategies as future work.

Comparing the spectrum of weight covariance matrices. Since $\mathbf{C}$ is subject to change while obtaining the stable state of the EoS, we analyze its initial and final (normalized) spectra for two different datasets of dimension $d_0 = 32$ and $N = 1024$. **Dataset 1**: $\mathbf{x}^{1,i} \sim \mathcal{N}(-2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), i \in [N/2]$, $\mathbf{x}^{2,j} \sim \mathcal{N}(2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), j \in [N/2]$. **Dataset 2:** $\mathbf{x}^{c,i} \sim \mathcal{N}(\mathbf{0}_{d_0}, 4 * \mathbf{I}_{d_0}), i \in [N/2], c \in [2]$ The first dataset is our running example, and the second is pure random noise data. Surprisingly, we observed that the EoS solution captures correlations in the data for both datasets, which is reflected in its final spectrum. In particular, the singular values shift from being constant at initialization to exhibiting a decay in their values (Figure 4). Such a shift does not exactly match the case of 2L-FCN because of (1) the difference in the dynamics of GD and Newton-Krylov with annealing, and (2) we start with a GP-based initial value for $\mathbf{C}$ in EoS. A rigorous analysis of the EoS dynamics is an open research direction (as also highlighted by the Seroussi et al. (2023)). Nonetheless, the EoS offers a richer data-dependent setup to analyze the activations and weights, than the UFM.

Appendix H. Additional Experiments

Compute Resources. All the experiments in this paper were executed on a machine with 16 GB of host memory and 8 CPU cores. Experiments with the EoS on datasets of varying dimensions and sample sizes took the longest time (≈ 1 hour) to finish.

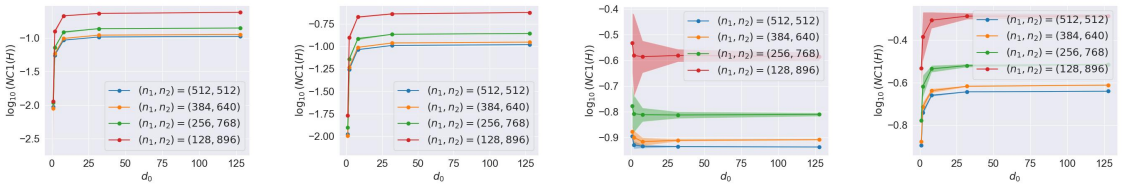


(a) EoS: Dataset 1 (b) 2L-FCN: Dataset 1 (c) EoS: Dataset 2 (d) 2L-FCN: Dataset 2

Figure 4: Normalized singular values sorted in descending order $\lambda_i/\lambda_{max}, \forall i \in [32]$ for $\mathbf{C} \in \mathbb{R}^{d_0 \times d_0}$ in case of EoS and for $\mathbf{W}^{(1)\top}\mathbf{W}^{(1)} \in \mathbb{R}^{d_0 \times d_0}$ in case of 2L-FCN. Here `init` represents the initialized state of an EoS and 2L-FCN in their respective plots. The `final` state of EoS is obtained by solving it using Newton-Krylov with a final factor 500. The 2L-FCN with $d_1 = 500$ is trained for $10,000$ epochs using GD with a learning rate 10^{-3}, weight decay 10^{-6}, $\sigma_w = 1$.

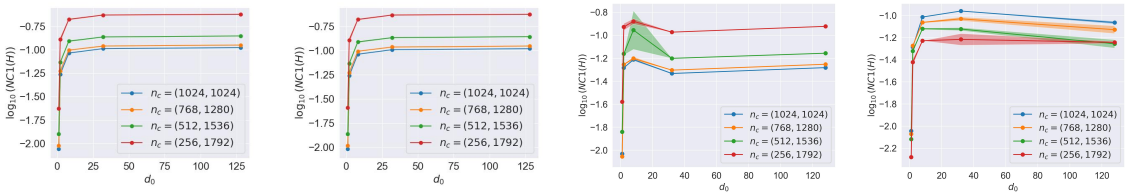
Dataset with $C = 4$. Similar to the formulation of $\mathcal{D}_1(N, d_0)$ for $C = 2$ in the main text, we formulate $\mathcal{D}_2(N, d_0)$ for $C = 4$ and $\forall i, j, k, l \in [N/4]$ as follows:

$$\begin{aligned}
&\left\{(\mathbf{x}^{1,i} \sim \mathcal{N}(-6 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{1,i} = -3))\right\} \\
&\cup \left\{(\mathbf{x}^{2,j} \sim \mathcal{N}(-2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{2,j} = -1)\right\} \\
&\cup \left\{(\mathbf{x}^{3,k} \sim \mathcal{N}(2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{3,k} = 1))\right\} \\
&\cup \left\{(\mathbf{x}^{4,l} \sim \mathcal{N}(6 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{4,l} = 3))\right\}.
\end{aligned} \tag{145}$$



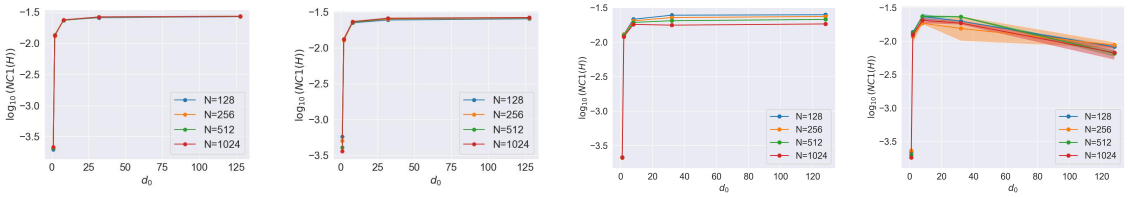
(a) Q_{GP-Erf} (b) $\Theta_{NTK-Erf}$ (c) $Q_{GP-ReLU}$ (d) $\Theta_{NTK-ReLU}$

Figure 5: $\mathcal{NC}_1(\mathbf{H})$ of the post-activation NNGP kernel ($Q_{GP}^{(1)}$) and NTK ($\Theta^{(2)}$) corresponding to `Erf`, `ReLU` activations. The dimension d_0 on the x-axis is chosen from $\{1, 2, 8, 32, 128\}$. For (n_1, n_2) such that $n_1 + n_2 = N = 1024$, we sample the vectors $\mathbf{x}^{1,i} \sim \mathcal{N}(-2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{1,i} = -1, i \in [n_1]$ for class 1 and $\mathbf{x}^{2,j} \sim \mathcal{N}(2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{2,j} = 1, j \in [n_2]$ for class 2 as our dataset.



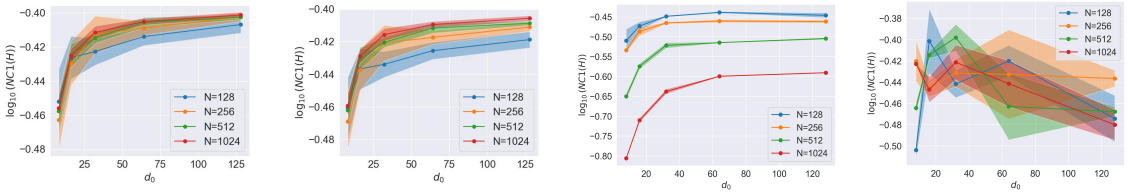
(a) Q_{GP-Erf} (b) $\Theta_{NTK-Erf}$ (c) EoS (d) 2L-FCN

Figure 6: $\mathcal{NC}_1(\mathbf{H})$ of the limiting kernels, adaptive kernel (EoS) with final annealing factor $d_1 = 500$ and 2L-FCN with $d_1 = 500$ and `Erf` activation. d_0 on the x-axis is chosen from $\{1, 2, 8, 32, 128\}$. For a tuple $n_c = (n_1, n_2)$ such that $n_1 + n_2 = N = 2048$, we sample the vectors $\mathbf{x}^{1,i} \sim \mathcal{N}(-2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{1,i} = -1, i \in [n_1]$ for class 1 and $\mathbf{x}^{2,j} \sim \mathcal{N}(2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{2,j} = 1, j \in [n_2]$ for class 2.



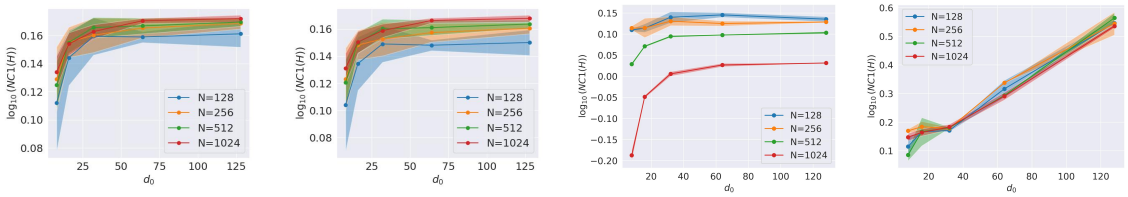
(a) Q_{GP-Erf} (b) $\Theta_{NTK-Erf}$ (c) EoS (d) 2L-FCN

Figure 7: $\mathcal{NC}_1(\mathbf{H})$ of the limiting kernels, adaptive kernel (EoS) with final annealing factor $d_1 = 500$ and 2L-FCN with $d_1 = 500$ and `Erf` activation. The dimension d_0 on the x-axis is chosen from $\{1, 2, 8, 32, 128\}$. For a particular N, we sample the vectors $\mathbf{x}^{1,i} \sim \mathcal{N}(-6 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{1,i} = -1, i \in [N/2]$ for class 1, and $\mathbf{x}^{4,j} \sim \mathcal{N}(6 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{2,j} = 1, j \in [N/2]$ for class 2.



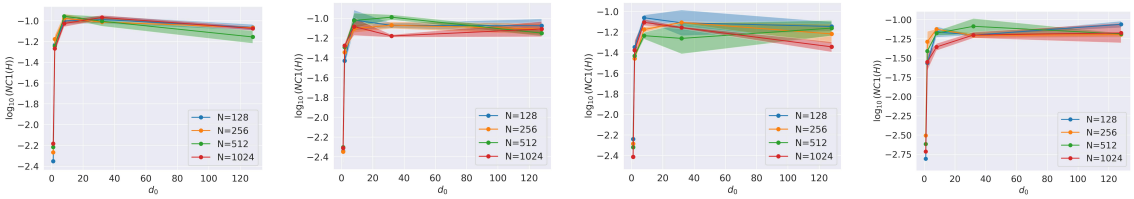
(a) Q_{GP-Erf} (b) $\Theta_{NTK-Erf}$ (c) EoS (d) 2L-FCN

Figure 8: $\mathcal{NC}_1(\mathbf{H})$ of the limiting kernels, adaptive kernel (EoS) with final annealing factor $d_1 = 500$ and 2L-FCN with $d_1 = 500$ and `Erf` activation. The dimension d_0 on the x-axis is chosen from $\{1, 2, 8, 32, 128\}$. For a particular N, we sample the vectors $\mathbf{x}^{1,i} \sim \mathcal{N}(-2 * \mathbf{1}_{d_0}, \mathbf{I}_{d_0}), y^{1,i} = -1, i \in [N/2]$ for class 1 and $\mathbf{x}^{2,j} \sim \mathcal{N}(2 * \mathbf{1}_{d_0}, \mathbf{I}_{d_0}), y^{2,j} = 1, j \in [N/2]$ for class 2.



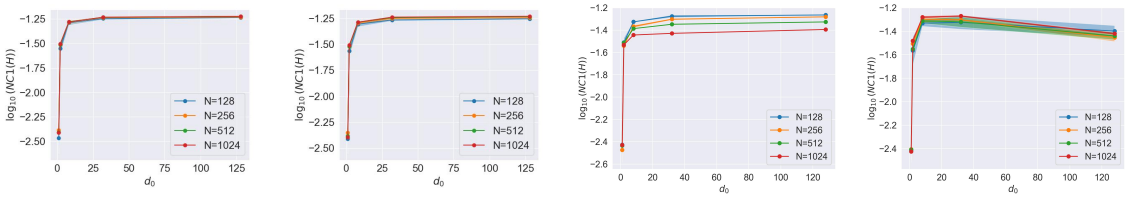
(a) Q_{GP-Erf} (b) $\Theta_{NTK-Erf}$ (c) EoS (d) 2L-FCN

Figure 9: $\mathcal{NC}_1(\mathbf{H})$ of the limiting kernels, adaptive kernel (EoS) with final annealing factor $d_1 = 500$ and 2L-FCN with $d_1 = 500$ and `Erf` activation. The dimension d_0 on the x-axis is chosen from $\{8, 16, 32, 64, 128\}$. For a particular N, we sample the vectors $\mathbf{x}^{1,i} \sim \mathcal{N}(-2 * \mathbf{1}_{d_0}, 4 * \mathbf{I}_{d_0}), y^{1,i} = -1, i \in [N/2]$ for class 1 and $\mathbf{x}^{2,j} \sim \mathcal{N}(2 * \mathbf{1}_{d_0}, 4 * \mathbf{I}_{d_0}), y^{2,j} = 1, j \in [N/2]$ for class 2.



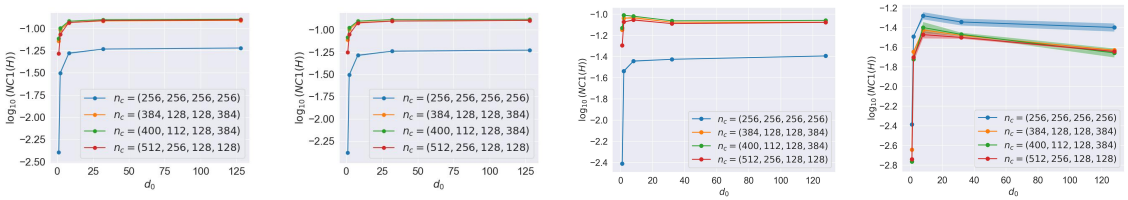
(a) $L = 3$ (b) $L = 4$ (c) $L = 5$ (d) $L = 6$

Figure 10: $\mathcal{NC}_1(\mathbf{H})$ of deeper FCN networks with `Erf` activation and hidden later width 500. The dimension d_0 on the x-axis is chosen from $\{1, 2, 8, 32, 128\}$. For a particular N, we sample the vectors $\mathbf{x}^{1,i} \sim \mathcal{N}(-2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{1,i} = -1, i \in [N/2]$ for class 1 and $\mathbf{x}^{2,j} \sim \mathcal{N}(2 * \mathbf{1}_{d_0}, 0.25 * \mathbf{I}_{d_0}), y^{2,j} = 1, j \in [N/2]$ for class 2.



(a) Q_{GP-Erf} (b) $\Theta_{NTK-Erf}$ (c) EoS (d) 2L-FCN

Figure 11: $\mathcal{NC}_1(\mathbf{H})$ of the limiting kernels, adaptive kernel (EoS) with final annealing factor 500 and 2L-FCN with $d_1 = 500$ and Erf activation on dataset $\mathcal{D}_2(N, d_0)$ (145).



(a) Q_{GP-Erf} (b) $\Theta_{NTK-Erf}$ (c) EoS (d) 2L-FCN

Figure 12: $\mathcal{NC}_1(\mathbf{H})$ of the limiting kernels, adaptive kernel (EoS) with final annealing factor $d_1 = 500$ and 2L-FCN with $d_1 = 500$ and `Erf` activation. The dimension d_0 on the x-axis is chosen from $\{1, 2, 8, 32, 128\}$. For a tuple $n_c = (n_1, n_2, n_3, n_4)$ such that $n_1 + n_2 + n_3 + n_4 = N = 1024$ we sample the dataset as per $\mathcal{D}_2(N, d_0)$ (145)

Appendix I. Limitations and Future Work

In certain cases, we have observed that none of the kernel methods approximate the 2L-FCN reasonably. One such instance is the following, where we sample $\mathbf{x}^{1,i} \sim \mathcal{N}(-2 * \mathbf{1}_{d_0}, 4 * \mathbf{I}_{d_0}), y^{1,i} = -1, i \in [N/2]$ for class 1 and $\mathbf{x}^{2,j} \sim \mathcal{N}(2 * \mathbf{1}_{d_0}, 4 * \mathbf{I}_{d_0}), y^{2,j} = 1, j \in [N/2]$ for class 2 of our dataset. Essentially, these are scenarios where there is a significant overlap between samples of the two classes. First, we note that we had to increase the learning rate of our 2L-FCN from 10^{-3} to $5 \cdot 10^{-3}$ and run GD for 2000 epochs for convergence. For dimensions $d_0 = \{8, 16, 32\}$, the EoS reasonably approximates the 2L-FCN but for $d_0 = \{64, 128\}$, the $\mathcal{NC}_1(\mathbf{H})$ values for 2L-FCN turned out to be almost twice as large as the EoS (see Figure 9). To this end, we leave modifications to the EoS for handling such noisy data cases and different activation functions as future work.

Additionally, we highlight the difficulties in the theoretical/empirical analysis of NC1 with EoS. The primary bottleneck is a lack of rigorous study on the existence and uniqueness of solutions (As also highlighted by Seroussi et al. (2023)). Since we deviate from the lazy regime and deal with kernels in the feature learning setup, we cannot expect simpler closed-form solutions like the limiting NNGP/NTK for the EoS. However, analytical solutions to the EoS can sometimes be time-consuming and require a manual selection of the annealing schedule. This is a tradeoff that can be improved with future research. Furthermore, the role of scaling N, d_0, d_1 on NC1 is yet to be fully understood and we hope that our analysis lays the groundwork for such efforts. Finally, we point the reader to Appendix F for a discussion on a relative NC1 metric that explicitly incorporates the variability collapse of the data vectors into the NC1 metric. In particular, we aim to differentiate between settings where the neural network learned meaningful features and learned to classify complex datasets or was simply able to leverage the already collapsed data vectors. Our results showcase that in higher dimensions, the data vectors are 'more' collapsed than the activations themselves. Thus showcasing the limitations of the current NC1 metrics and encouraging the reader to explore richer variants.