DISS. ETH N°

On a Mathematical Understanding of Deep Neural Networks

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

AMIR JOUDAKI

MSc ETH in Computer Science, ETH Zurich

born on 25.06.1989

Accepted on the recommendation of

Prof. Dr. Gunnar Rätsch (ETH Zurich), examiner
Prof. Dr. Thomas Hoffman (ETH Zurich), co-examiner
Prof. Dr. Francis Bach (INRIA Paris), co-examiner

2024

Abstract

This thesis investigates the fundamental challenges in training deep neural networks, focusing on signal propagation through network depth. It examines how various architectural choices, such as fully connected layers, weight initialization, normalization layers, and non-linear activations, affect forward and backward passes in deep architectures. The research addresses critical issues such as rank collapse, gradient stability, and their impact on training dynamics and network performance.

Leveraging tools from mean field theory, random matrix theory, and Markov chain theory, we develop a mathematical framework for analyzing signal propagation in deep networks. We characterize conditions leading to rank collapse and gradient instability and provide theoretical insights into the effectiveness of normalization techniques and initialization schemes, suggesting avenues for improving signal propagation and training dynamics in very deep networks. Fundamentally, this thesis's findings are a step toward mathematical principles underlying the success of modern neural network architectures.

Acknowledgments

I would like to thank my advisor, Prof. Gunnar Ratsch, for his guidance, and mentorship throughout my Ph.D. studies. I would also like to thank my co-advisor, Prof. Francis Bach, for his insightful comments and feedback on my research, which have been instrumental in my academic development.

During the years, I have had the pleasure of working with and learning from many brilliant researchers. I am particularly thankful for the collaboration with Hadi Daneshmand, who has been a great friend and collaborator throughout my Ph.D. studies. I am also thankful to Dr. Andre Kahles, who supervised and mentored me during the earlier stages of my Ph.D. studies on topics related to bioinformatics. I am thankful to Prof. Thomas Hoffman and Prof. Francesco Orabonoa for their insightful discussions and conversations.

I would also like to thank former and current members of the biomedical informatics group (BMI) at ETH Zurich, who have provided a warm, supportive, and intellectually stimulating environment. I am grateful to Stefan and Natalia for their friendship and support during the most challenging times. I am also grateful to my friends and colleagues, Gideon, Harun, Alex Immer, Ragnar, Vincent, Victor, Matthias, Natalie, Francesco Locattelo, and the rest of the BMI community for their friendship and various insightful discussions.

Besides my research network, I am grateful for the friendship and intellectual discussions with several friends. In particular, I am grateful for my conversations with David, Stefan, Alex Meterez, Aran, Hoda, Hadi, Atiye, and Andishe, and Claudia. I am also grateful for my friend, Meysam, who passed away, but left a memory of a great friend with me that will always be cherished.

Last but not least, I am grateful for their ceassless and unconditoinal love and support of my sisters Atefeh and Afagh, my mother Farzaneh, and father Hossein, and my fiancée Alice. I am There is no overstting that without their love and support, I would not have been able to embark and complete this journey.

This thesis is dedicated to my father, who left us shortly before finishing this thesis.

Contents

1

Introduction

Deep neural networks have revolutionized the field of artificial intelligence, achieving unprecedented performance in a wide range of tasks, from image recognition [KSH12] to natural language processing [Dev+18]. Despite their remarkable success, these models often remain enigmatic, functioning as "black boxes" that transform inputs into outputs through a complex series of non-linear operations [AB16]. This lack of theoretical understanding poses significant challenges for researchers and practitioners, as it hinders our ability to better understand and optimize these systems.

At the heart of the deep learning paradigm lies signal propagation—the journey of information as it flows through the layers of a neural network during both forward and backward passes [GB10a]. Understanding this process is crucial for several reasons. It provides insights into how neural networks process and transform information, potentially illuminating the principles underlying their underlying processes. A deeper understanding of signal propagation can guide the design of better network initialization [GB10a], more effective network architectures [He+16a], and more efficient optimization algorithms [KB14]. Overall, a theoretical understanding contributes to the broader goal of making neural networks more principled and efficient design and understanding of neural networks.

The importance of signal propagation becomes particularly evident when considering the challenges associated with training deep neural networks. As networks grow in depth, they gain the potential for increased expressivity and the ability to learn more complex representations [BSF94]. However, this increased depth also introduces significant obstacles to stable training, many of

which are directly related to how signals propagate through the network [GB10a; He+15]. This thesis aims to bridge some of these gaps in our understanding.

1.1 Research objectives and scope

The primary objective of this thesis is to demystify certain behaviors of neural networks by conducting a thorough investigation into the effects of various neural network components on signal propagation through depth. Specifically, we aim to address the following central research question:

How do forward and backward passes evolve as signals propagate through the layers of a deep neural network?

To approach this question comprehensively, we focus on several key aspects of neural network design: fully connected layers, weight initialization, normalization techniques, and non-linear activations. Fully connected layers serve as fundamental building blocks of neural networks and provide a starting point for our analysis [SMG13a]. The choice of initial weights can dramatically affect a network's training dynamics, and we investigate various initialization strategies and their impact on signal propagation [SMG13a; GB10a; He+15]. Second, normalization layers such as Batch Normalization (BN) [IS15], and Layer Normalization (LN) [BKH16] have been crucial in enabling the training of very deep networks, and we analyze how these techniques influence signal flow and stability. Third, the choice of non-linear activation function significantly affects a network's representational capacity and training dynamics, so we examine popular choices such as ReLU [NH10] and hyperbolic tangent, exploring their effects on signal propagation [GBB11; MHN13; CUH15; He+15; RZL17].

From a mathematical perspective, our analysis focuses on two key operations within neural networks: matrix products and element-wise activations. Matrix products, which occur in linear layers, transform representations between layers and affect how signals propagate through the network [SMG13a]. Non-linear activation functions introduce crucial non-linearities into the network, and we investigate how different activation functions shape the distribution of activations and gradients [Kla+17; PSG17; PSG18]. While normalization layers like BN and LN are non-linear operations, they do not act elementwise. Somewhat surprisingly, we find that we can study them also as a special type of matrix product, where one of the matrices is diagonal and is proportional to the standard deviation of the activations in feature or batch space [Dan+20; DJB21].

Our analysis primarily focuses on networks at initialization, as this initial state plays a critical role in determining the subsequent optimization trajectory and the network's ultimate performance [SMG13a; Xia+18; FC18; PSG17]. This enables us to leverage tools from random matrix theory and Markov chain theory to analyze how layer representations evolve stochastically.

1.2 Challenges in training deep neural networks

As neural networks have grown deeper, achieving state-of-the-art performance across various tasks [He+16a; Dev+18], two main challenges have emerged in training these architectures effectively [PMB13]. In the backward pass, the vanishing and exploding gradients issues become significant [BSF94; PMB13; Han18]. In the forward pass, the problem of representation collapse occurs, where different input samples map to increasingly similar representations as depth increases [Dan+20; Noc+22b]. Both representation collapse and gradient instability substantially affect training dynamics and network performance [Han18].

1.2.1 Explosion and vanishing gradients

The problems of exploding and vanishing gradients have been long-standing challenges in training deep neural networks [Hoc91; BSF94]. These issues arise during the backward pass of the backpropagation algorithm and can severely hinder the learning process.

Gradient explosion occurs when the gradients grow as they propagate backward through the network layers, which can lead to numerical instability, causing the training process to diverge [PMB13]. Conversely, vanishing gradients occur when the gradients become exponentially small, effectively preventing the network from learning long-range dependencies [Hoc98].

The vanishing gradient problem is particularly detrimental when it affects a network's first or last layers. In the case of first-layer vanishing gradients, the network fails to capture important features from the input data, leading to a loss of crucial information at the beginning of the network [GB10a]. When gradients vanish in the last layers, the network struggles to propagate error signals back to earlier layers, resulting in poor fine-tuning of the overall network [He+15].

Moreover, vanishing gradients can cause a deep network to behave like a much shallower one, negating the potential benefits of deep architectures in learning hierarchical representations [SGS15]. This "effective shallowness" limits the network's ability to learn complex, non-linear mappings that deep learning is renowned for [BL07].

The vanishing and exploding gradient problems are intimately related to depth, weight initialization, and activation functions [GB10a; He+15]. From a mathematical perspective, gradients can be represented as an extended chain of matrix products, a consequence of the chain rule in calculus. The primary challenge arises in maintaining a stable gradient flow as this product chain grows [SMG13a; PMB13]. To combat these issues, researchers have devised several strategies. These include meticulous weight initialization techniques [SMG13a; GB10a], strategic selection of activation functions [NH10; CUH15; Kla+17], and innovative architectural designs such as skip connections in residual networks [He+16a]. These approaches collectively aim to mitigate the adverse effects of deep network architectures on gradient propagation.

1.2.2 Rank collapse

Rank collapse refers to the phenomenon where the outputs of deep neural networks become increasingly correlated as the depth increases, leading to a loss of expressivity [Dan+20]. This issue is particularly prevalent in networks with standard initialization schemes and can severely impede the network's ability to learn complex representations [Dan+20]. In mathematical terms, rank collapse manifests as a decrease in the rank of the Gram matrix of hidden representations as signals propagate through the network [Dan+20]. This reduction in rank effectively limits the network's capacity to represent diverse features, which is shown to be hard to recover during training [Dan+20; DJB21].

Recent studies have shown that rank collapse is not limited to fully connected networks but also affects other architectures such as convolutional neural networks [Xia+18] and transformers [DCL21; Noc+22b]. Addressing rank collapse is crucial for enabling the training of very deep networks and fully leveraging their potential representational power.

Rank collapse and the vanishing gradient are closely interconnected phenomena stemming from the challenge of proper signal propagation through the network [Dan+20; Han18]. Informally, both issues can be viewed as the network "losing" information as it propagates through successive layers. While vanishing gradients pertain to the backward pass, rank collapse is a statement about the forward pass, making their formal relationship less apparent [PSG17]. Consequently, further theoretical investigations are necessary to elucidate the connection between these two phenomena and develop strategies to address them concurrently [Yan+19; HR18].

1.2.3 Impact on Training Dynamics

The phenomena of rank collapse and gradient instability substantially affect the training dynamics and overall performance of deep neural networks. These challenges manifest in several interconnected ways, significantly impacting the efficiency and effectiveness of the learning process. Networks suffering from rank collapse or gradient issues often require significantly more training iterations to achieve comparable performance [IS15; San+18; Dan+20; DJB21]. This increased training time can be a major bottleneck, especially for large-scale models and datasets.

Moreover, while very deep networks should, in theory, be capable of learning highly abstract and hierarchical features [BL07; ZF14], these issues can prevent the network from fully utilizing its depth [He+16a; Hua+17]. This limitation undermines one of the primary advantages of deep architecture. Gradient instability also makes networks more sensitive to choices of learning rate, initialization scheme, and other hyperparameters [ZDM19; Luo+19]. This increased sensitivity complicates the training process and can lead to inconsistent results across different runs.

1.2.4 Addressing challenges of depth

To address the issues with gradient stability and rank collapse, researchers have proposed various techniques, including careful initialization schemes [GB10a; He+15], normalization layers [IS15; BKH16], skip connections [He+16a; Hua+17], and gradient clipping [PMB13; Zha19]. While these methods have shown empirical success, a deeper theoretical understanding of their effects on signal propagation is crucial for developing more principled approaches to network design and optimization [SMG13a; PSG17; Yan+19].

Despite these advances, many open questions remain regarding the optimal strategies for mitigating rank collapse and gradient instability across different network architectures and tasks [HR18; Yan+19]. Future research in this area will likely focus on developing a unified theory that explains how these various techniques interact and how they can be combined optimally to improve neural network training and performance [PSG18; Yan20].

1.3 Thesis structure and contributions

This dissertation explores the challenges in training deep neural networks and proposes novel approaches to address these issues. The chapters are given in the same chronological order in which they were written and published. The chapters are as follows:

- **Chapter 2:** Batch Normalization Orthogonalizes Representations. This chapter presents a novel theoretical analysis of batch normalization, demonstrating how it orthogonalizes representations in deep neural networks. Relevant publications: Hadi Daneshmand, Amir Joudaki, and Francis Bach. "Batch normalization orthogonalizes representations in deep random networks." In *Advances in Neural Information Processing Systems*, vol. 34, pp. 4896-4906, 2021 [DJB21].

- **Chapter 3:** Bridging Mean field and Finite Width gap. Here, we extend the analysis of batch normalization to bridge the gap between mean field theory and finite-width networks. We present concentration bounds for mean field predictions with batch normalization. Relevant publication: Amir Joudaki, Hadi Daneshmand, and Francis Bach. "On Bridging the Gap between Mean Field and Finite Width in Deep Random Neural Networks with Batch Normalization." In *International Conference on Machine Learning*, 2023 [JDB23a].

- **Chapter 4 & Chapter 5:** Discuss how normalization and activation functions can lead to isometry of representations in deep neural networks. Relevant publication: Amir Joudaki, Hadi Daneshmand, and Francis Bach. "On the impact of activation and normalization in obtaining isometric embeddings at initialization." In *Advances in Neural Information Processing Systems*, vol. 36, pp. 39855-39875, 2023 [JDB23b].

- **Chapter 6:** Batch Normalization without Gradient Explosion. This chapter presents how a theoretically inspired weight initialization can prevent gradient explosion with batch normalization. Relevant publication: Alexandru Meterez*, Amir Joudaki*, Francesco Orabona, Alexander Immer, Gunnar Ratsch, and Hadi Daneshmand. "Towards Training Without Depth Limits: Batch Normalization Without Gradient Explosion." In *International Conference on Learning Representations*, 2024 [Met+24].

- **Chapter 7:** Conclusion and Future Directions. This chapter summarizes the main contributions of the dissertation and outlines potential avenues for future research.

Each chapter improves our understanding of deep neural network training dynamics and provides novel techniques for addressing depth challenges in these models.

Personal Retrospective. For this dissertation, I decided to focus on the theoretical aspects of the contributions and leave most of the contributions of a purely empirical nature out of the dissertation. While the empirical results are important and interesting, I aimed to provide a

more coherent and consistent narrative throughout the dissertation by focusing on the theoretical aspects. The empirical results are available in the corresponding publications.

2

Batch Normalization Orthogonalizes Representations

Batch Normalization (BN) [IS15] enhances training across a wide range of deep network architectures and experimental setups [He+16a; Hua+17; Sil+17]. The practical success of BN has inspired research into the underlying mechanism of BN [San+18; KAA19; ALL19; Bjo+18]. BN influences first-order optimization methods by avoiding the rank collapse in deep representation [Dan+20], direction-length decoupling of optimization [Koh+18], influencing the convergence of the steepest descent [ALL19; Bjo+18], and smoothing the optimization objective function [San+18; KAA19]. However, the benefits of BN go beyond its critical role in optimization. For example, [FSM21] shows that BN networks with random weights also achieve surprisingly high performance after only minor adjustments of their weights. This striking result motivates us to study the representational power of random networks with BN.

We study hidden representations across layers of a laboratory random BN with linear activations. Consider a batch of samples passing through consecutive BN and linear layers with Gaussian weights. The representations of these samples are perturbed by each random linear transformation, followed by a non-linear BN. At first glance, the deep representations appear unpredictable after many stochastic and non-linear transformations. Yet, we show that these transformations orthogonalize the representations. To prove this statement, we introduce the notion of "orthogonality gap", defined in Section 2.2, to quantify the deviation of representations from a perfectly

orthogonal representation. Then, we prove that the orthogonality gap decays exponentially with the network depth and stabilizes around a term inversely related to the network width. More precisely, we prove

$$\mathbb{E}\left[\text{orthogonality gap}\right] = \mathcal{O}\left((1-\alpha)^{\text{depth}} + \frac{\text{batchsize}}{\alpha\sqrt{\text{width}}}\right)$$

holds for $\alpha > 0$ that is an absolute constant under a mild assumption. In probability theoretic terms, we prove stochastic stability of the Markov chain of hidden representations [Kus67; KY03; Kha11]. The orthogonality of deep representations allows us to prove that the distribution of the representations after linear layers contracts to a Wasserstein-2 ball around isotropic Gaussian distribution as the network depth grows. Moreover, the radius of the ball is inversely proportional to the network width. Omitting details, we prove the following bound holds:

$$\text{Wasserstein}_2(\text{representations}, \text{Gaussian})^2 = \mathcal{O}\left((1-\alpha)^{\text{depth}}\,(\text{batchsize}) + \frac{(\text{batchsize})^2}{\alpha\sqrt{\text{width}}}\right).$$

The above equation shows how depth, width, and batch size, interact with the Gaussian approximation of the representations. Since the established rate is exponential with depth, the distribution of the representations stays in a Wasserstein ball around isotropic Gaussian distribution after a few layers. Thus, BN not only stabilizes the distribution of the representations, which is its main promise [IS15], but also enforces Gaussian isotropic distribution in deep layers.

There is growing interest in bridging the gap between neural networks, as the most successful parametric methods for learning, and Gaussian processes and kernel methods, as well-understood classical models for learning [JGH18; GM+18; Lee+19a; BM19; Hua+14]. This link is inspired by studying random neural networks in the asymptotic regime of infinite width. The seminal work by [Nea96] sparks that a single-layer network resembles a Gaussian process as its width goes to infinity. However, increasing the depth may significantly shift the distribution of the representations away from Gaussian [IS15]. This distributional shift breaks the link between Gaussian processes and deep neural networks. To ensure Gaussian representations, [GM+18] suggests increasing the width of the network proportional to the network depth. Here, we show that BN ensures Gaussian representations even for deep networks with *finite* width. This result bridges the link between deep neural networks and Gaussian processes in the regime of finite width. Many studies rely on deep Gaussian representations in an infinite width setting [Yan+19; Sch+17; PSG18; Kla+17; DPKL19]. Our non-asymptotic Gaussian approximation can be incorporated into their analysis to extend these results to the regime of finite width.

Since training starts from random networks, representations in these networks directly influence training. Hence, recent theoretical studies has investigated the interplay between initial hidden

representations and training [Dan+20; Bjo+18; FSM21; Sch+17; SMG13a; Bah+20]. In particular, it is known that hidden representations in random networks *without BN* become correlated as the network grows in depth, thereby drastically slowing training [Dan+20; He+16a; Bjo+18; SMG13a]. On the contrary, we prove that deep representations in networks *with BN* are almost orthogonal. We experimentally validate that initial orthogonal representations can save training time that would otherwise be needed to orthogonalize them. By proposing a novel initialization scheme, we ensure the orthogonality of hidden representations. Such an initialization effectively avoids the training slowdown with depth for vanilla networks, with no need for BN. This observation further motivates studying the inner workings of BN to replace or improve it for deep learning.
Theoretically, we made the following contributions:

1. For MLPs with batch normalization, linear activation, and Gaussian weights, we prove that representations across layers become increasingly orthogonal up to a constant inversely proportional to the network width.

2. Leveraging the orthogonality, we prove that the distribution of the representations contracts to a Wasserstein ball around a Gaussian distribution as the depth grows. Up to the best of our knowledge, this is the first *non-asymptotic* Gaussian approximation for deep neural networks with finite width.

Experimentally, we made the following contribution[1]:

3. Inspired by our theoretical understanding, we propose a novel weight initialization for standard neural networks that ensure orthogonal representations without BN. Experimentally, we show that this initialization effectively avoids training slowdown with depth in the absence of BN.

2.1 Preliminaries

2.1.1 Notations

Akin to [Dan+20], we focus on a Multi-Layer Perceptron (MLP) with batch normalization and linear activation. Theoretical studies of linear networks is a growing research area [SMG13a; Dan+20; BHL19; Aro+19a]. When weights are initialized randomly, linear and non-linear

[1]Implementations are available at https://github.com/hadidaneshmand/batchnorm21.git

networks share similar properties such as the rank collapse issue studied in [Dan+20]. For ease of analysis, we assume activations are linear.

We use n to denote batch size, and d to denote the width across all layers, which we further assume is larger than the batch size $d \geq n$. Let $H_\ell \in \mathbb{R}^{d\times n}$ denote representations for n samples in layer ℓ, with $H_0 \in \mathbb{R}^{d\times n}$ corresponding to n input samples in the batch with d features. Successive representations are connected by Gaussian weight matrices $W_\ell \sim \mathcal{N}(0, I_d/d)$ as

$$H_{\ell+1} = \frac{1}{\sqrt{d}} BN(W_\ell H_\ell), \quad BN(M) = \mathrm{diag}(MM^\top)^{-1/2} M, \tag{2.1}$$

where $\mathrm{diag}(.)$ zeros out off-diagonal elements of its input matrix, and the scaling factor $1/\sqrt{d}$ ensures that all matrices $\{H_k\}_k$ have unit Frobenius norm (see Section 2.1). The BN function in Eq. (2.1) differs slightly from the commonly used definition for BN as the mean correction is omitted. However, [Dan+20] shows this difference does not change the network properties qualitatively.

2.1.2 The linear independence of hidden representations

[Dan+20] observe that if inputs are linearly independent, then their hidden representations remain linearly independent in all layers as long as $d = \Omega(n^2)$. Under technical assumptions, [Dan+20] establishes a lower-bound on the average of the rank of hidden representations over infinite layers. Based on this study, we assume that the linear independence holds and build our analysis upon that. This avoids restating the technical assumptions of [Dan+20] and also further technical refinements of their theorems. The next assumption presents the formal statement of the linear independence property.

Assumption $\mathcal{A}_1(\alpha, \ell)$. There exists an absolute positive constant α such that the minimum singular value of H_k is greater than (or equal to) α for all $k = 1, \ldots, \ell$.

The linear independence of the representations is a shared property across all layers. However, the representations constantly change when passing through random layers. In this chapter, we mathematically characterize the dynamics of the representations across layers.

2.2 Orthogonality of deep representations

2.2.1 A warm-up observation

To illustrate the difference between BN and vanilla networks, we compare hidden representations of two input samples across the layers of these networks. Figure 2.1 plots the absolute value of cosine similarity of these samples across layers. This plot shows a stark contrast between vanilla and BN networks: While representations become increasingly orthogonal across layers of a BN network, they become increasingly aligned in a vanilla network. More specifically, we observe that BN is able to orthognalized almost aligned representations; while, the vanilla network provides almost align representation of two orthogonal samples in deep layers. While this behaviour has been theoretically studied for vanilla networks [Dan+20; Bjo+18; SMG13a], up to the best of our knowledge, it is not theoretically analyzed for BN networks for networks with finite width regime. In the following section, we formalize and prove this orthogonalizing property for BN networks with finite widths.

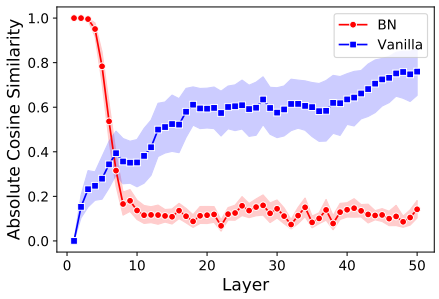


Figure 2.1: **Orthogonality: BN vs. vanilla networks.** The horizontal axis shows the number of layers, and the vertical axis shows the absolute value of cosine similarity between two samples across the layers ($d = 32$). Mean and 95% confidence intervals of 20 independent runs.

2.2.2 Theoretical analysis

The notion of orthogonality gap plays a central role in our analysis. Given the hidden representation $H \in \mathbb{R}^{d\times n}$, matrix $H^\top H$ constitutes inner products between representations of different samples. Note that $H^\top H \in \mathbb{R}^{n\times n}$ is different form the covariance matrix $HH^\top/n \in \mathbb{R}^{d\times d}$. The orthogonality gap of H is defined as the deviations of $H^\top H$ from identity matrix, after proper

scaling. More precisely, define $V : \mathbb{R}^{d\times n} \setminus \mathbf{0} \to \mathbb{R}_+$ as

$$V(H) := \left\| \left(\frac{1}{\|H\|_F^2} \right) H^\top H - \left(\frac{1}{\|I_n\|_F^2} \right) I_n \right\|_F. \tag{2.2}$$

The following theorem establishes a bound on the orthogonality of representation in layer ℓ.

Theorem 2.1. *Under Assumption $\mathcal{A}_1(\alpha, \ell)$, the following holds:*

$$\mathbb{E}\left[V(H_{\ell+1})\right] \le 2\left(1 - \frac{2}{3}\alpha\right)^\ell + \frac{3n}{\alpha\sqrt{d}}. \tag{2.3}$$

Assumption $\mathcal{A}_1$ is studied by [Dan+20] who note that $\mathcal{A}_1(\alpha, \infty)$ holds as long as $d = \Omega(n^2)$. If $\mathcal{A}_1$ does not hold, one can still prove that there is a function of representations that decays with depth up to a constant (see Section A.3).

The above result implies that BN is an approximation algorithm for orthogonalizing the hidden representations. If we replace $\mathrm{diag}(M)^{-1}$ by $(M)^{-1}$ in BN formula, in Eq. equation 2.1, then all the hidden representation will become exactly orthogonal. However, computing the inverse of a *non-diagonal* $d \times d$ matrix is computationally expensive, which must repeat for all layers throughout training, and differentiation must propagate back through this inversion. The diagonal approximation in BN significantly reduces the computational complexity of the matrix inversion. Since the orthogonality gap decays at an exponential rate with depth, the approximate orthogonality is met after only a few layers. Interestingly, this yields a desirable cost-accuracy trade-off: For a larger width, the orthogonality is more accurate, due to term $1/\sqrt{d}$ in the orthogonality gap, and also the computational gain is more significant.

From a different angle, Theorem 2.1 proves the stochastic stability of the Markov chain of hidden representations. In expectation, stochastic processes may obey an inherent Lyapunov-type of stability [Kus67; KY03; Kha11]. One can analyze the mixing and hitting times of Markov chains based on the stochastic stability of Markov chains,[KS76; Ebe09]. In our analysis, the orthogonality gap is a Lyapunov function characterizing the stability of the chain of hidden representations. This stability opens the door to more theoretical analysis of this chain, such as studying mixing and hitting times. For, example the stability can be used for mixing analysis of the chain $\{H_1, \ldots, H_\ell, \ldots\}$. Let π denote the stationary distribution of this chain. Under a particular stochastic stability condition called geometric drift condition,

$$\left|\mathbb{E}\left[\varphi(H_\ell)\right] - \mathbb{E}_{H\sim\pi}\left[\varphi(H)\right]\right| \le \alpha^\ell \left|\mathbb{E}\left[\varphi(H_0)\right] - \mathbb{E}_{H\sim\pi}\left[\varphi(H)\right]\right|$$

holds for $\alpha \in (0,1)$ and measurable function $\varphi : R^{d\times n} \to R$ (see Theorem 3.6 of [Hai10]). The drift condition holds if there exists a Lyapunov function $L : R^{d\times n} \to [0,1]$ and constant $K \geq 0$ such that the following holds:

$$E\left[L(H_{\ell+1})|H_\ell\right] \leq \gamma L(H_\ell) + K.$$

We prove that the above condition holds in Eq equation A.6. However, such theoretical analyses may not be of interest to the machine learning community, hence we focus on the implications of these results for understanding the underpinnings of neural networks.

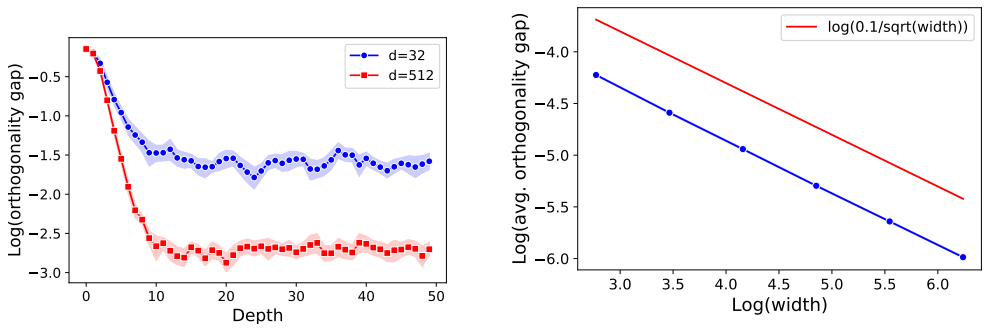
It is helpful to compare the orthogonality gap in BN networks to studies on vanilla networks [Bjo+18; Dan+20; SMG13a]. The function implemented by a vanilla linear network is a linear transformation as the product of random weight matrices. The spectral properties of the product of i.i.d. random matrices are the subject of extensive studies in probability theory [Bou+12]. Invoking these results, one can check that the orthogonality gap of the hidden representations in vanilla networks rapidly increases since the rank of representations converges to a one matrix as the depth grows [Dan+20].

Remarkably, [AAK21] proves if activation functions obey a self-normalization property, then a specific kernel of hidden representations becomes well-condition as the network depth grows. However, it is not clear how to impose the self-normalization property. Here, we establish the whitening for an explicitly defined normalization used in practice.

2.2.3 Experimental validations

Our experiments presented in Fig. 2.2a validate the exponential decay rate of V with depth. In this plot, we see that $\log(V_\ell)$ linearly decreases for $\ell = 1, \ldots, 20$, then it wiggles around a small constant. Our experiments in Fig. 2.2b suggest that the $\mathcal{O}(1/\sqrt{d})$ dependency on width is almost tight. Since $V(H_\ell)$ rapidly converges to a ball, the average of $V(H_\ell)$ over layers estimates the radius of this ball. This plot shows how the average of $V(H_\ell)$, over 500 layers, changes with the network width, validating the $\mathcal{O}(1/\sqrt{d})$ dependency implied by Theorem 2.1.

Consider an input matrix $H_0 \in \mathbb{R}^{d\times n}$ containing n samples in $\mathbb{R}^d$. Assuming that elements of this matrix are i.i.d. zero-mean and unit variance random variables, the minimum singular value of H_0 is greater than $\sqrt{d} - \sqrt{n}$. In practical applications, the batchsize used for normalization is relatively smaller than the network width, hence $\mathcal{A}_1(\Theta(1), 0)$ holds. For the intermediate representations, we also observed that $\mathcal{A}_1(\alpha,\ell)$ holds –for an α independent from ℓ– as long as $d \gg n$. Let α_0 denote the minimum singular value of H_0. For different values of d and



(a) Orthogonality vs. depth (b) Orthogonality vs. width

Figure 2.2: **Orthogonality gap vs. depth and width.** Left: $\log(V(H_\ell))$ vertically versus ℓ horizontally. Right: $\log(\frac{1}{500}\sum_{\ell=100}^{600} V(H_\ell))$ vertically versus $\log(d)$ horizontally. The chain starts from a diagonal H_0 with one relatively large diagonal value. This structure imposes a large orthogonality gap for H_0. Mean and 95% confidence interval of 20 independent runs.

n, we check whether $\mathcal{A}_1(0.1\alpha_0, 1000)$ holds. Fig. 2.3 illustrates that this assumption hold for $d = \Omega(n^2)$, which is also confirmed by [Dan+20].

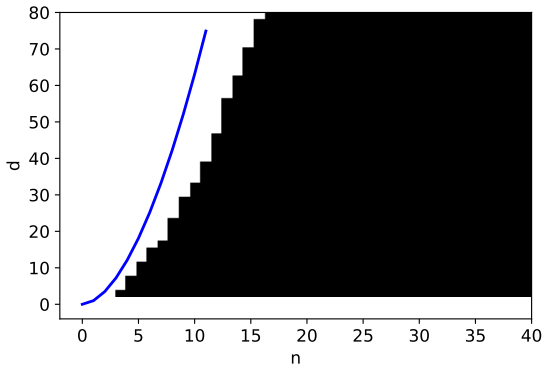


Figure 2.3: **Validations for $\mathcal{A}_1$** Pixel (n, d) marks whether $\mathcal{A}_1$ holds in all 10 independent runs: The black color indicates $\mathcal{A}_1(0.1\alpha_0, 1000)$ failed in at least once (where α_0 is the minimum singular value of H_0). The blue curve marks $d = (n - 28)^{1.8}$ highlighting $\mathcal{A}_1$ holds for $d = \Omega(n^2)$.

2.3 Gaussian approximation

2.3.1 Orthogonality yields Gaussian approximation

The established result on the orthogonality gap allows us to show that the representations after linear layers are approximately Gaussian. If H_ℓ is orthogonal, the random matrix $W_\ell H_\ell$ is equal in distribution to a standard Gaussian matrix due to the invariance of standard Gaussian distribution under linear orthogonal transformations. We can formalize this notion of Gaussianity by bounding the Wasserstein-2 distance between the distribution of $W_\ell H_\ell$ and standard Gaussian distribution. Let $\mathcal{W}_2(R_1, R_2)$ denote the Wasserstein-2 distance between probability distributions of two random variables R_1, and R_2. The next lemma formally establishes the link between the orthogonality and the distribution of the representations.

Lemma 2.2. *Given $G \in \mathbb{R}^{d\times n}$ with i.i.d. zero-mean $1/d$-variance Gaussian elements, the following Gaussian approximation holds:*

$$\mathcal{W}_2\left(W_\ell H_\ell, G/\sqrt{n}\right)^2 \leq 2n\mathbb{E}\left[V(H_\ell)\right]. \tag{2.4}$$

Combining the above result with Theorem 2.1 yields the result presented in the next corollary.

Corollary 2.3. *For $G \in \mathbb{R}^{d\times n}$ with i.i.d. zero-mean $1/d$-variance Gaussian elements,*

$$\mathcal{W}_2\left(W_\ell H_\ell, G/\sqrt{n}\right)^2 \leq 4n\left(1-\frac{2}{3}\alpha\right)^\ell + \frac{6n^2}{\alpha\sqrt{d}} \tag{2.5}$$

holds under Assumption $\mathcal{A}_1(\alpha, \ell)$.

In other words, the distribution of the representations contracts to a Wasserstein 2 ball around an isotropic Gaussian distribution as the depth grows. The radius of the Wasserstein 2 ball is at most $\mathcal{O}(1/\sqrt{\text{width}})$. As noted in the last section, $\mathcal{A}_1$ is extensively studied by [Dan+20] where it is shown that $\mathcal{A}_1(\alpha > 0, \infty)$ holds as long as $d = \Omega(n^2)$.

2.3.2 Deep neural networks as Gaussian processes

Leveraging BN, Corollary 2.3 establishes the first *non-asymptotic* Gaussian approximation for deep random neural networks. For vanilla networks, the Gaussianity is guaranteed only in the *asymptotic* regime of infinite width [GARA19; Nea96; Lee+19a; Nea96; HJ15]. Particularly,

[GM+18] links vanilla networks to Gaussian processes when their width is infinite and grows in successive layers, while our Gaussian approximation holds for networks with finite width across layers. Table 2.1 briefly compares Gaussian approximation for vanilla and BN networks.

Network	Width	Depth	Distribution of Outputs
Vanilla MLP	infinite	finite	Converges to Gaussian as width $\to \infty$
Vanilla Convnet	infinite	finite	Converges to Gaussian as width $\to \infty$
BN MLP (Cor. 2.3)	(in)finite	(in)finite	In a $\mathcal{O}(\text{width}^{-1/4})$-$\mathcal{W}_2$ ball around Gaussian

Table 2.1: **Distribution of representations in random vanilla and BN networks.** For the convolutional network, the width refers to the number of channels. Results for Vanilla MLPs and Vanilla convolution networks are establish by [GM+18], and [GARA19], respectively. Remarkably, Corollary 2.3 holds for MLP with linear activations.

The link between Gaussian processes and infinite-width neural networks has inspired several studies to rely on Gaussian representations in deep random networks [Kla+17; DPKL19; PSG18; Sch+17; Yan+19]. Assuming the representations are Gaussian, [Kla+17] designed novel activation functions that improve the optimization performance, [DPKL19] studies the sensitivity of random networks, [PSG18] highlights the spectral universality in random networks, [Sch+17] studies information propagating through the network layers, and [Yan+19] studies gradients propagation through the depth. Indeed, our analysis implies that including BN imposes the Gaussian representations required for these analyses.

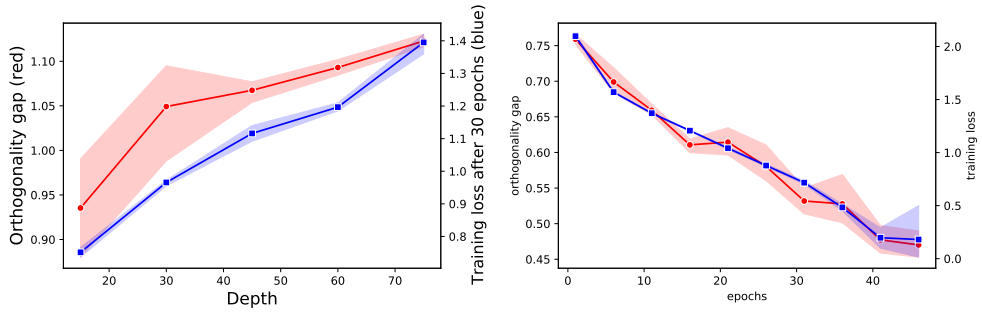
2.4 The orthogonality and optimization

In the preceding sections, we elaborated on the theoretical properties of BN networks in controlled settings. In this section, we demonstrate the practical applications of our findings. In the first part, we focus on the relationship between depth and orthogonality. Increasing depth drastically slows the training of neural networks with BN. Furthermore, we observe that as depth grows, the training slowdown highly correlates with the orthogonality gap. This observation suggests that SGD needs to orthogonalize deep representations in order to start classification. This intuition leads us to the following question: If orthogonalization is a prerequisite for training, can we save optimization time by starting from orthogonal representations? To test this experimentally, we devised a weight initialization that guarantees orthogonality of representations. Surprisingly, even in a network without BN, our experiments showed that this initialization avoids the training slow down, affirmatively answering the question.

Throughout the experiments, we use vanilla MLP (without BN) with a width of 800 across all hidden layers, ReLU activation, and used Xavier's method for weights intialization [GB10a]. We use SGD with stepsize 0.01 and batch size 500 and for training. The learning task is classification with cross entropy loss for CIFAR10 dataset [KH+09, MIT license]. We use PyTorch [Pas+19, BSD license] and Google Colaboratory platform with a single Tesla-P100 GPU with 16GB memory in all the experiments. The reported orthogonality gap is the average of the orthogonality gap of representation in the last layer.

2.4.1 Orthogonality correlates with optimization performance

In the first experiment, we show that the orthogonality of representations at the initialization correlates with optimization speed. For networks with 15, 30, 45, 60, and 75 widths, we register training loss after 30 epochs and compare it with the initial orthogonality gap. Figure 2.4a shows the training loss (blue) and the initial orthogonality gap (red) as a function of depth. We observe that representations are more entangled, i.e., orthogonal, when we increase depth, coinciding with the training slowdown. Intuitively, the slowdown is due to the additional time SGD must spend to orthogonalize the representations before classification. In the second experiment, we validate this intuitive argument by tracking the orthogonality gap during training. Figure 2.4b plots the orthogonality gap of output and training loss for a network with 20 layers. We observe that SGD updates are iteratively orthogonalizing representations, marked by the reduction in the orthogonality gap.



(a) gap and loss vs. depth (b) gap and loss during training

Figure 2.4: **Orthogonality and Optimization** Left: the orthogonality gap at initialization (red, left axis) and the training loss after 30 epochs (blue, right axis) with depth. Right: the orthogonality gap (red, left axis) and the training loss in each epoch (blue, right axis). Mean and 95% confidence interval of 4 independent runs.

2.4.2 Learning with initial orthogonal representations

We have seen that the slowdown in SGD for deeper networks correlates with the orthogonality gap before training. Here we show that by preemptively orthogonalizing representations, we avoid the slowdown with depth. While in MPL with linear activations, hidden representations remain orthogonal simply by taking orthogonal matrices as weights [PSG18; SMG13a], the same does not hold for networks with non-linear activations, such as ReLU. To enforce the orthogonality in the absence of BN, we introduce a dependency between weights of successive layers that ensures deep representations remain orthogonal. More specifically, we incorporate the SVD decomposition of the hidden representation of each layer into the initialization of the subsequent layer. To emphasis this dependency between layers and to distinguish it from purely orthogonal weight initialization, we refer to this as *iterative orthogonalization*.

We take a large batch of samples $n \geq d$, as the input batch for initialization. Let us assume that weights are initialized up to layer $W_{\ell-1}$. To initialize W_ℓ, we compute SVD decomposition of the representations $H_\ell = U_\ell \Sigma_\ell V_\ell^\top$ where matrices $U_\ell \in \mathbb{R}^{d\times d}$ and $V_\ell \in \mathbb{R}^{n\times d}$ are orthogonal. Given this decomposition, we initialize W_ℓ by

$$W_\ell = \frac{1}{\|\Sigma_\ell^{1/2}\|_F} V_\ell' \Sigma_\ell^{-1/2} U_\ell^\top, \tag{2.6}$$

where $V_\ell' \in \mathbb{R}^{d\times d}$ is an orthogonal matrix obtained by slicing $V_\ell \in \mathbb{R}^{n\times d}$. Notably, the inverse in the above formula exists when n is sufficiently larger than d [2]. It is easy to check that $V(W_\ell H_\ell) < V(H_\ell)$ holds for the above initialization (see Section A.4), similar to BN. By enforcing the orthogonality, this initialization significantly alleviates the slow down of training with depth (see Fig. 2.5), with no need for BN. This initialization is not limited to MLPs. In Section A.5, we compare iterative orthogonalization with a BN-replacement method recently proposed by [Bro+21].

2.5 Discussion

To recap, we proved the recurrence of random linear transformations and BN orthogonalizes samples. Our experiments underline practical applications of this theoretical finding: starting from orthogonal representations effectively avoids the training slowdown with depth for MLPs. Based on our experimental observations, a proper initialization ensuring the orthogonality of

[2] We may inductively assume that H_ℓ is almost orthogonal by the choice of $W_1, \ldots, W_{\ell-1}$. Thus, Σ_ℓ is invertible.

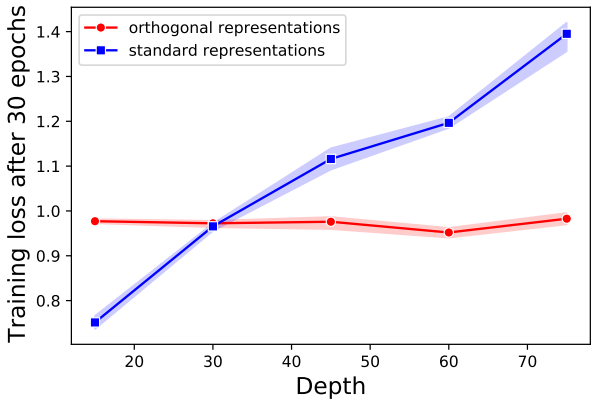


Figure 2.5: **Iterative orthogonalization.** Horizontal axis: depth. Vertical axis: the training loss after 30 epochs for Xavier's initialization (blue), our initialization (red). Mean and 95% confidence interval of 4 independent runs.

hidden representations may replace BN in neural architectures. This future research direction has the potentials to boost the training of deep neural networks and change benchmarks in deep learning.

Although our theoretical bounds hold for MLPs with linear activations, our experiments confirm similar orthogonal stability for various neural architectures.

[LDT21] experimentally compares properties of different normalization techniques including layer normalization (LN) [BKH16]. According to this study, LN does not necessarily orthogonalize the outputs of deep neural networks. Hence, more theoretical studies are required to understand the essence of different normalization techniques in deep learning.

3

Briding mean field and finite width analysis

There is a growing demand for a theoretical understanding of neural networks to improve their safety, robustness, computational and statistical effectiveness. Originating in statistical mechanics for investigating complex systems with interacting particles, this theory has been repurposed in recent years for exploring neural network dynamics under the regime of infinite width. By going beyond the microscopic changes of individual neurons, mean field analysis has revealed the collective neuronal behaviors that emerge at initialization [PSG18; Yan+19; PW17], throughout training [JGH18; CB18; Lee+19b], and after training [CB20; Ba+19].

In this chapter, we delve into the role of mean field theory at initialization when network weights are allocated randomly. Pioneering works by Glorot and Bengio [GB10b] and Saxe, McClelland, and Ganguli [SMG13a] underscored the impact of initialization on training, paving the way for mean field theory to uncover a wealth of insights. Examples include identifying connections between wide neural networks and Gaussian processes [Nea96; GM+18; JGH18], examining the concentration of singular values of input-output Jacobians [PSG18; Fen+22], and designing activation functions [Kla+17; RZL17; LNR22]. Remarkably, Xiao et al. [Xia+18] introduced an initialization scheme capable of training convolutional networks comprising 10000 layers.

A common thread among these studies is the dynamics of inner products between hidden representations, encoded in their Gram matrix. Mean field theory models these dynamics via a

difference equation derived from the infinite-width limit of the network. However, mean field analysis is inherently prone to approximation errors when dealing with networks of finite width. As G. Matthews et al. [GM+18] observed, this error grows with depth, ultimately leading to vacuous error bounds in the infinite depth limit. To tackle this issue, they propose to increase the network width proportional to depth. This idea is echoed in other studies which propose maintaining a constant ratio between depth and width [HN19; LNR21], a regime in which Hanin [Han22] confirmed a constant concentration bound for mean field estimates.

Can we achieve a bounded mean field error when the width is finite? We answer this question affirmatively for MLPs that are endowed with batch normalization. In particular, we show that under some technical assumption on the underlying dynamics (as formally expressed in Theorem 3.2), the mean field estimation error for Gram matrices remains bounded at infinite depth. Specifically, we demonstrate that this error is limited by $\text{width}^{-1/2}$ with high probability. This contrasts with the vacuous concentration bounds at infinite depth observed in the absence of normalization [LNR22]. Our results highlight the importance of existing mean field analyses of batch normalization by Yang et al. [Yan+19], and demonstrate their high accuracy in the finite width scenarios that are relevant for practical applications.[1]

3.1 Related works

Numerous studies [SMG13a; Fen+22; Yan+19] have provided valuable insights into training neural networks by studying input-output Jacobians of neural networks with and without normalization at initialization. For example, Feng et al. [Fen+22] have shown that the rank of the input-output Jacobian of neural networks without normalization at initialization diminishes exponentially with depth, while Yang et al. [Yan+19] have shown that batch normalization avoids this exponential diminishing.

The spectrum of Jacobians is intimately related to the spectra of Gram matrices. A Gram matrix (G-matrix) contains inner products of samples within a batch (equation 3.2). Thus, a degenerate G-matrix for the penultimate layer implies that the outputs are insensitive to the inputs [Fen+22; LNR22]. Rank collapse in the last hidden layer occurs in various neural network architectures, including MLPs [SMG13a], convolutional networks [Dan+20], and transformers [DCL21], and leads to ill-conditioning of the input-output Jacobian, which slows training [DJB21; PSG18; Yan+19]. Saxe, McClelland, and Ganguli [SMG13a] have shown that avoiding rank collapse can

[1] Codes available at https://github.com/ajoudaki/meanfield-normalization.

accelerate the training of deep linear networks, making it a focus of theoretical and experimental research [PSG18; Dan+20; DJB21].

A recent line of research [Dan+20] postulates that batch normalization can enhance the training of deep neural nets by avoiding the rank collapse. This claim has been supported by empirical evidence [Yan+19; Dan+20], as well as theoretical studies for neural networks with infinite widths [Yan+19] and linear activation [DJB21]. It has been shown that batch normalization prevents degenerate representations at initialization [Dan+20], and orthogonalizes representations [DJB21]. However, these results are limited to linear activations. The present study extends these findings to neural networks with finite widths and non-linear activations, under an assumption from Markov chain theory.

3.2 Problem settings and background

Notation and terminology. I_n denotes the identity matrix of size $n \times n$ and 1_n denotes the all ones vector in $\mathbb{R}^n$. $\otimes$ refers to Kronecker product. μ_X refers to the probability measure of the random variable X. We use $f \lesssim g, g \gtrsim f$ and $f = O(g)$ to denote the existence of an absolute constant c such that $f \leq c\, g$. $\|v\|$ for a vector v denotes the L^2 norm. $\|C\|$ for matrix C denotes the L^2 operator norm $\|C\| = \sup_{x \in \mathbb{R}^n} \|Cx\|/\|x\|$, $\|C\|_F$ denotes the Frobenius norm. We use $\kappa(C)$ to denote the ratio of largest to smallest eigenvalue. Both $h_{r\cdot}$ and $\mathrm{row}_r(h)$ denote row-vector representation of the r-th row of h. Finally, $X \sim \mathcal{N}(\mu, \sigma^2)^{n \times m}$ denotes $X \in \mathbb{R}^{n \times m}$ is a Gaussian matrix whose elements are drawn i.i.d. from $\mathcal{N}(\mu, \sigma^2)$.

Setup. Let $h_\ell \in \mathbb{R}^{d \times n}$ denote the hidden representation at layer ℓ, where n corresponds to the size of the mini-batch, and d denotes the width of the network that is kept constant across all layers. The sequence $\{h_\ell\}$ is a Markov chain as

$$h_{\ell+1} := W_\ell \phi \circ \textsc{Bn}(h_\ell), \qquad W_\ell \sim \mathcal{N}(0, 1/d)^{d \times d}, \tag{3.1}$$

where $h_0 \in \mathbb{R}^{d \times n}$ is the input batch, ϕ is the element-wise activation function, and $\textsc{Bn}$ is the batch normalization [IS15], which applies row-wise centering and scaling by standard deviation:

$$\textsc{Bn}(x) = \frac{x - \mathrm{mean}(x)}{\sqrt{\mathrm{Var}(x)}}, \qquad \forall r : \mathrm{row}_r(\textsc{Bn}(h)) = \textsc{Bn}(\mathrm{row}_r(h)).$$

3.3 Mean field models and fixed-point analyses

3.3.1 Mean field Gram dynamics

The Gram matrix G_ℓ is defined as the matrix of inner products of hidden representations at layer ℓ as seen in the equation below:

$$G_\ell := \frac{1}{d}(\phi \circ \text{BN}(h_\ell))(\phi \circ \text{BN}(h_\ell))^\top. \tag{3.2}$$

Understanding the dynamics of G_ℓ is a significant challenge in deep learning theory, and has been the subject of several studies [Yan+19; PSG18; PW17]. Due to the randomness of weights, determining the trajectory of this random process proves to be arduous. By tending width d to infinity, i.e., the mean field regime, we can approximate these stochastic dynamics with a deterministic dynamics as below:

$$\overline{G}_{\ell+1} = \mathbb{E}_{h\sim\mathcal{N}(0,\overline{G}_\ell)}\left[\phi\left(\frac{\sqrt{n}Mh}{\|Mh\|}\right)^{\otimes 2}\right], \tag{3.3}$$

Where $\overline{G}_0 = G_0$ serves as the input G-matrix and $M = I_n - \frac{1}{n}1_n^{\otimes 2}$ applies mean reduction on the preactivations. The mean field approach in this context assists in elucidating the analysis of Gram matrices.

3.3.2 Fixed point analysis for infinitely deep and wide Networks

The fixed points of the mean field dynamics, as expressed in equation 3.3 may help elucidate the properties of $\overline{G}_\ell$ as $\ell \to \infty$. Yang et al. [Yan+19] provide a comprehensive characterization of these fixed-points, denoted by G_*, for neural networks with batch normalization. For networks with linear activations, Yang et al. [Yan+19] establish a global convergence to these well-conditioned fixed-points. While they empirically observe convergence to these well-conditioned fixed-point for networks with non-linear activations, that is not established theoretically. In other words, it is challenging to describe the properties of $\overline{G}_\ell$ for finite width and depth, and it is unclear how the fixed-point Gram matrix G_* can inform us about G_ℓ.

3.3.3 An observation

Through an empirical observation we can demonstrate that G_* may not always provide an accurate estimate for G_ℓ. We observe that for a network without batch normalization and linear activations

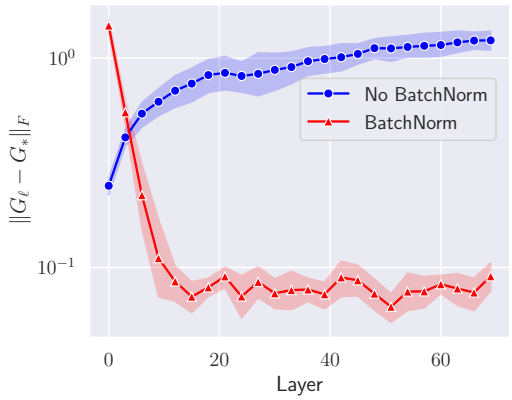


Figure 3.1: *Mean field error amplification with(out) batch-normalization.* The horizontal axis represents the number of layers ℓ (linear), while the vertical axis (log-scale) shows $\|G_\ell - G_*\|_F$, for networks with $n = 5, d = 1000$. The traces show mean and shades indicate 90% confidence intervals over 10 independent simulations.

(when $\phi \circ \text{BN} = identity$ in equation 3.1), the Frobenius distance between G_ℓ and G_* increases with ℓ. In contrast, G_ℓ converges to a neighborhood of G_* when the network includes batch normalization layers. These observations suggest that the mean field estimate G_* from Yang et al. [Yan+19] accurately represents G_ℓ when batch normalization is present.

3.3.4 The challenge of depth for mean field theory

Mean field analysis suffers from a systematic estimation error that increases with depth. Assuming that $\overline{G}_0 = G_0$, then an error of $O(d^{-1/2})$ is observed between $\overline{G}_1$ and G_1 due to the concentration of empirical covariance. Consequently, the mean field dynamics in equation 3.3 incur an error of $O(d^{-1/2})$ at each layer [LNR22]. As depth ℓ grows, these errors are amplified, and the bounds on $\|G_\ell - \overline{G}_\ell\|_F$ become vacuous, thus raising questions about the practical applicability of these fixed-point analyses when width is finite.

Several studies strive to refine the mean field model to enhance its predictive accuracy [GM+18; HN19; LNR22]. Li, Nica, and Roy [LNR22] propose using a stochastic differential equation to model the layer-wise $O(d^{-1/2})$ estimation error for mean field Gram dynamics. This approach allows for accurate predictions of Gram dynamics for MLPs with activation functions but only in the infinite-width-and-depth regime. Our observations, however, suggest that for networks with batch normalization, the deterministic model of Gram matrices provides a surprisingly accurate estimate.

Daneshmand, Joudaki, and Bach [DJB21] established this observation for multilayer perceptrons (MLPs) with batch normalization (BN) and linear activations, subject to specific conditions. They demonstrated that as ℓ increases, batch normalization progressively aligns the Gram matrices G_ℓ with the identity matrix, which coincides with G_* for such networks [Yan+19]. Yang et al. [Yan+19] further proved a concentration bound for $\|G_\ell - G_*\|_F$ in networks with batch normalization and linear activations. However, both these findings are limited to linear activations. Our objective is to extend these results to networks incorporating non-linear activations.

3.4 Concentration bounds for Mean field Predictions with Batch Normalization

3.4.1 Geometric ergodic assumption

The chain of hidden representations obeys a non-linear stochastic recurrence. Despite this non-linearity, the distribution associated with the representation obeys a linear fixed-point iteration determined by the Markov kernel K associated with the chain h_ℓ. The distribution of h_ℓ, denoted by μ_ℓ, obeys

$$\mu_{\ell+1} = T(\mu_\ell), \quad T(\mu) := \int K(x,y) d\mu(y). \tag{3.4}$$

The fixed-points of the above equation are invariant distributions of the chain, which we denote by μ_*. Recall that the total variation for distributions over $d \times n$ matrices can be defined as $\|\mu_X - \mu_Y\|_{tv} := \sup_{A \subseteq \mathbb{R}^{d\times n}} |\mu_X(A) - \mu_Y(A)|$. Notably, the above recurrence is non-expansive in total variation, hence $\|\mu_\ell - \mu_*\|_{tv} \leq \|\mu_{\ell-1} - \mu_*\|_{tv}$ holds for all ℓ. However, we assume the chain obeys a strong property ensuring the convergence to a unique invariant distribution.

Assumption 3.1 (Geometric ergodicity)**.** We assume the chain of hidden representations admits a unique invariant distribution. Furthermore, there is constant α ($\alpha > 0$) such that

$$\|\mu_\ell - \mu_*\|_{tv} \leq (1-\alpha)^\ell \|\mu_0 - \mu_*\|_{tv},$$

holds almost surely for all h_0.

The geometric ergodic property is established for various Markov chains, such as the Gibbs sampler, state-space models [Ebe09], hierarchical Poisson models [Ros95], and Markov chain Monte Carlo samplers [Jon01]. Doeblin [Doe38] provides weak conditions that ensure geometric

ergodicity. Doeblin's condition holds when the Markov chain can explore the entire state space [Ebe09]. This condition may hold under weak assumption on the input matrix for the chain of hidden representations. In particular, when h_ℓ has full rank, the Gaussian product $W_\ell h_\ell$ may explore the entire $\mathbb{R}^{d\times n}$.

3.4.2 Main result

The next theorem proves fixed-point G_* provides an estimate for Gram matrices of sufficiently deep neural networks with a finite width.

Theorem 3.2 (BN-MLP Concentration)**.** *Assume the Markov chain of representations* $\{h_\ell\}$ *obeys Assumption 3.1 with* $\alpha > 0$*, and has non-degenerate fixed-point* G_**. If the activation* ϕ *is uniformly bounded* $|\phi(x)| = O(|x|)$*, then Gram matrix deviation* $\|G_* - G_\ell\|_F$ *is bounded by*

$$\kappa(G_*)O\left((1-\alpha)^{\frac{\ell}{2}} + \frac{n}{\sqrt{d}}\alpha^{-\frac{1}{2}}\ln^{\frac{1}{2}}(\frac{d}{n})\right), \tag{3.5}$$

with high probability in d *and* ℓ*.*

Theorem 3.2 quantifies the accuracy of our mean field predictions in terms of batch size, width, depth, and conditioning of G_*. Notably, almost all commonly used activations, e.g., ReLU and hyperbolic tangent, satisfy the uniform bounded condition $|\phi(x)| = O(|x|)$. Under Assumption 3.1, this theorem proves the fixed-point Gram matrix G_* accurately estimates G_ℓ for a sufficiently large ℓ. According to this theorem, $\|G_\ell - G_*\|_F$ decays with depth at an exponential rate. Thus, approximately after a logarithmic number of layers $\ell \approx \log(\text{width}/\text{batch-size})$, the term $O(n/\sqrt{d})$ dominates the distance.

Remarkably, this is a considerable improvement compared to the concentration bounds for neural networks without batch normalization that become vacuous as the depth increases [HN19; Han22]. The established bound in the last theorem holds jointly for all ℓ in that we do not need to apply union bound.

Let us remark that if the fixed-point Gram matrix is degenerate, i.e., if $\kappa(G_*)$ is unbounded, the bound of the Theorem becomes vacuous. Therefore, Theorem 3.2 reinforces the necessity for a well-conditioned fixed-point for the mean field errors to remain within bounds. As long as the fixed-point Gram is well-conditioned, the Gram matrices G_ℓ's stay within an $O(\text{batch}/\text{width}^{1/2})$ proximity with constant probability.

When contrasting Theorem 3.2 with the activation shaping approach by Li, Nica, and Roy [LNR22], we observe that while activation shaping necessitates solving a stochastic differential

equation to track the dynamics of the Gram matrix, BN-MLP relies solely on the mean field prediction G_*, which can be computed in closed-form Yang et al. [Yan+19].

Proof Sketch of Theorem 3.2. We first construct an approximate invariant distribution, associated with T as defined in equation 3.4. For the construction of such distribution, we utilize the mean field Gram matrix to form an input $\hat{h} \in \mathbb{R}^{d\times n}$, with rows drawn i.i.d. from $\text{row}_r(\hat{h}) \sim \mathcal{N}(0, G_*)$. The next lemma proves that the law of $\hat{h}$, denoted by $\hat{\mu}$, does not significantly change under T.

Lemma 3.3. *Assuming uniformly bounded activation* $|\phi(x)| = O(|x|)$, *we have*

$$\|T(\hat{\mu}) - \hat{\mu}\|_{tv} \lesssim \|G_*^{-1}\| \frac{n^2}{d} \ln(d/n). \tag{3.6}$$

The proof of the last lemma is based on the fixed-point property of G_* (see Appendix for the detailed proof). Using the last lemma together with Assumption 3.1, we prove that $\hat{\mu}$ is in a tv-ball around the invariant distribution μ_*. Under this assumption, we have

$$\begin{aligned} \|T(\hat{\mu}) - T(\mu_*)\|_{tv} &= \|T(\hat{\mu}) - \mu_*\|_{tv} \\ &\leq (1-\alpha)\|\hat{\mu} - \mu_*\|_{tv}, \end{aligned} \tag{3.7}$$

where we used the invariant property of μ_* in the above equation. Using triangular inequality, we get

$$\begin{aligned} \|T(\hat{\mu}) - \hat{\mu}\|_{tv} &= \|T(\hat{\mu}) - \mu_* + \mu_* - \hat{\mu}\|_{tv} \\ &\geq \|\mu_* - \hat{\mu}\|_{tv} - \|T(\hat{\mu}) - \mu_*\|_{tv} \\ &\geq \alpha\|\hat{\mu} - \mu_*\|_{tv}. \end{aligned} \tag{3.8}$$

Plugging the bound from the last lemma into the above inequality concludes $\hat{\mu}$ lies within a radius $n^2\|G_*^{-1}\|/d\alpha$ of μ_*. This concludes the proof: Since the chain is geometric ergodic, the distribution μ_ℓ converges to an tv-ball around $\hat{\mu}$ at an exponential rate. This allows us to characterize the moments of μ_ℓ using those of $\hat{\mu}$.

3.4.3 Validation of main theoretical results

Our principal finding suggests a link between the Gram matrices of hidden representations with independent weights. Assuming $\kappa(G_*) = O(1)$ this is captured by the relation:

$$\|G_\ell - G_*\|_F = O\left((1-\alpha)^{\ell/2} + \frac{n}{\sqrt{d}}\right). \tag{3.9}$$

We test this relationship by numerically estimating G_* by tending d and ℓ to sufficiently large values. We then plot the left-hand side of the above equation versus depth, width, and batch size in the following figures. These plots illustrate how the difference in Gram matrices changes with respect to depth, width, and batch size. This supports our theoretical results and showcases their potential implications for practical settings.

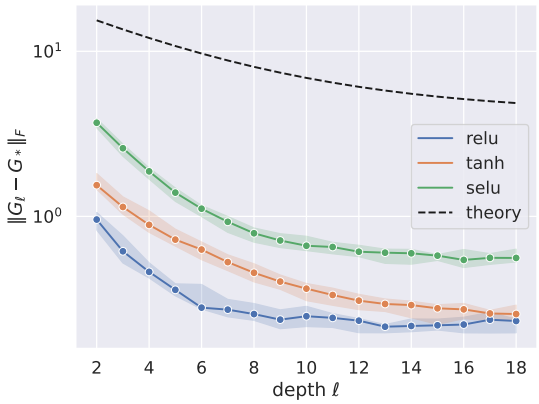


Figure 3.2: $\|G_\ell - G_*\|_F$ vs. depth, $\ell = 1, 2, ..., 20$, with a fixed width of $d = 1000$ and a batch size of $n = 10$. The dashed line shows the theoretical upper bound of Theorem 3.2.

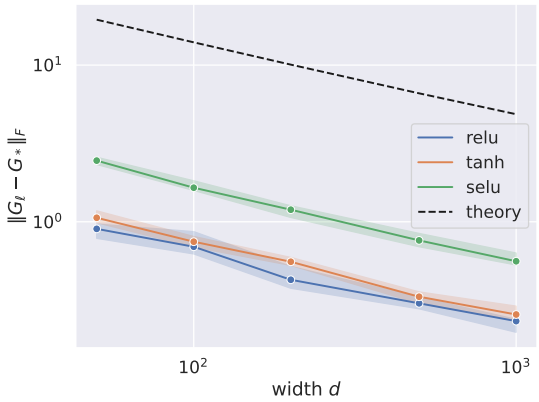


Figure 3.3: $\|G_\ell - G_*\|_F$ vs. width, $d = 50, 100, 200, 500, 1000$, with a fixed depth of $\ell = 20$ and a batch size of $n = 10$. The second term $O(n/\sqrt{d})$ is always dominant, as demonstrated in the following log-log plot.

You can find the detailed proofs in Chapter B.

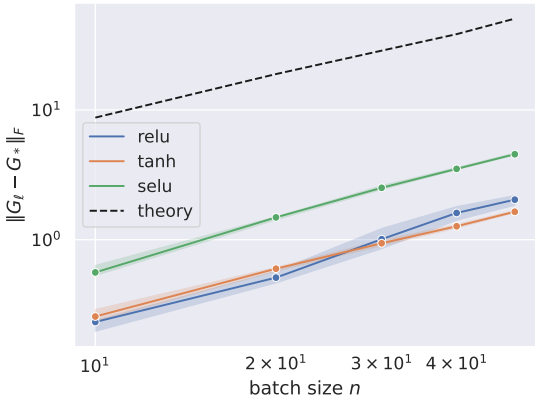


Figure 3.4: $\|G_\ell - G_*\|_F$ vs. batch size, with a fixed width of $d = 1000$ and a depth of $\ell = 20$, and varying batch sizes of $n = 10, 20, 30, 40, 50$. Dashed line shows upper bound given in Theorem 3.2.

3.5 Limitations and Future Directions

In this chapter, we presented a theoretical framework that bridges the gap between the mean field theory of neural networks with finite and infinite widths, with a focus on batch normalization at initialization. Many questions that were out of the scope for this study, suggesting directions for new lines of inquiry.

Rapidly mixing assumption. One limitation of our work is the rapidly mixing assumption that was used to establish the concentration of our results. While our experiments validated our results based on this assumption, it would be beneficial to prove that this assumption holds for a wide range of neural networks with batch normalization.

Training and optimization. While our focus of the current work was on random neural networks. In an elegant observation, Feng et al. [Fen+22] demonstrate that the rank of input-output Jacobian of neural networks without normalization at initialization diminishes at an exponential rate with depth (Theorem 5), which implies changes in the input does not change the direction of outputs. In a remarkable observation, Yang et al. [Yan+19] show the exact opposite for BN-MLP using a mean field analysis (Theorem 3.10): any slight changes in the input lead to unbounded changes in the output. These results naturally raise the following question: Can we arrive at non-trivial results about input-output Jacobian at the infinite depth finite width regime?

The mean field approach is also used to analyze the training mechanism. In particular, Chizat and Bach [CB18] prove that gradient descent globally converges when optimizing single-layer neural networks in the limit of an infinite number of neurons. Although the global convergence does not hold for standard neural networks, insights from this mean field analysis can be leveraged in understanding the training mechanism.

Exploring other normalizations. More research is needed for other normalization techniques, such as weight normalization [SK16] or layer normalization [BKH16] to understand the impact of these normalization techniques on the robustness and generalization of neural networks. Our findings highlight the power of mean field theory for analyzing neual networks with normalization layers.

Extending to other architectures Our analyses are limited to MLPs. Extending our work to convolutional neural networks and transformers would enable us to analyze and enhance initialization for these neural networks. In particular, recent studies have shown that transformers suffer from the rank collapse issue when they grow in depth [Noc+22a]. A non-asymptotic mean field theory may enable us to tackle this issue by providing a sound understanding of representation dynamics in transformers.

Overall, our results demonstrate that depth is not necessarily a curse for mean field theory, but can even be a blessing when neural networks have batch normalization. The inductive bias provided by batch normalization controls the error propagation of mean field approximations, enabling us to establish non-asymptotic concentration bounds for mean field predictions. This result underlines the power of mean field analyses in understanding the behavior of deep neural networks, thereby motivating the principle development of new initialization and optimization techniques for neural networks based on mean field predictions.

4

Obtaining isometry with normalization

Normalization layers, such as batch normalization [IS15] and layer normalization [BKH16], are essential components of neural architecture design. They have been shown to improve the training stability and speed of deep neural networks [He+16a; Dev+18]. In this chapter, we explore the isometry properties of normalization layers. While in Chapter 2, we focused on the orthogonality properties of normalization layers, here we investigate the isometry properties of these layers, which will be formalized later. This chapter is dedicated to one of the primary results of this thesis, which is the isometry bias of normalization layers. While conceptually orthogonal and isometry properties are related, the striking property of isometry is that it holds deterministically for all matrices, while orthogonality properties discussed in Chapter 2 hold in expectation, with respect to a particular initialization of the weights.

4.1 Gram matrices and isometry

Given n data points $\{x_i\}_{i\leq n} \in \mathbb{R}^d$, the Gram matrix G^ℓ of the feature vectors $x_1^\ell, \ldots, x_n^\ell \in \mathbb{R}^d$ at layer ℓ of the network is defined as

$$G^\ell := \left[\langle x_i^\ell, x_j^\ell \rangle\right]_{i,j\leq n}, \qquad \ell = 1, \ldots, L. \tag{4.1}$$

Intuitively, an isometric Gram matrix implies that the network preserves the distances and angles between the input data points after mapping them to the feature space. Isometry of Gram matrices

can be quantified using the eigenvalues of G^ℓ. One possible way to formulate isometry is to use the ratio of the volume and scale of the parallelepiped spanned by the feature vectors $x_1^\ell, \ldots, x_n^\ell$. For example, consider two points on a plane $x_1, x_2 \in \mathbb{R}^2$ with lengths $a = |x_1|, b = |x_2|$ and angle $\theta = \angle(x_1, x_2)$. The ratio is given by $ab\sin(\theta)/(a^2 + b^2)$, which is maximized when $a = b$ and $\theta = \pi/2$. This is shown for $n = 2$ and $n = 3$ feature vectors in Figure 4.1.

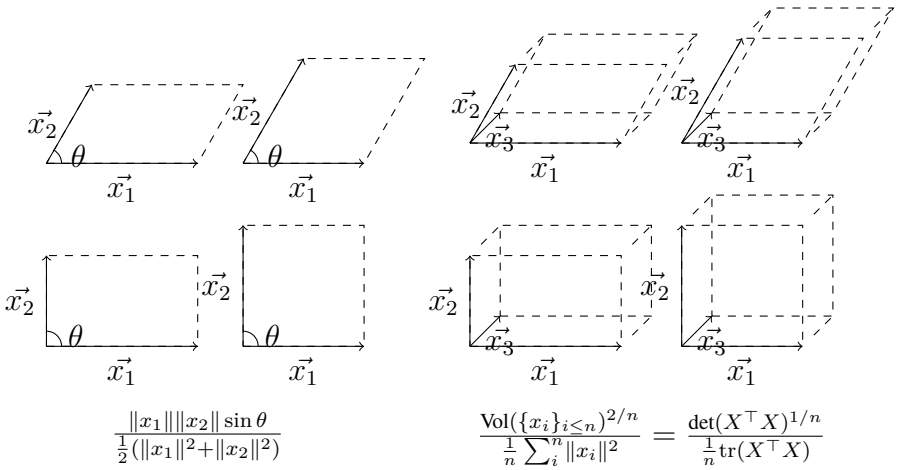


Figure 4.1: A geometric interpretation of isometry: higher volume corresponds to higher isometry.

Inspired by this intuition, we can define the isometry of the Gram matrix.

Definition 4.1. Let M be an $n \times n$ positive semi-definite matrix. We define the *isometry* $\mathcal{I}(M)$ of M as the ratio of its normalized determinant to its normalized trace:

$$\mathcal{I}(M) := \frac{\det(M)^{1/n}}{\frac{1}{n}\text{tr}(M)}. \tag{4.2}$$

The function $\mathcal{I}(M)$ defined in equation 4.2 quantifies how well M approximates the identity matrix I_n. We can easily check that $\mathcal{I}(M)$ has some desirable properties (see Lemma 4.2 for formal statements and proofs):

- Scale-invariance: Multiplying M by a constant does not affect $\mathcal{I}(M)$.
- Isometry-preserving: $\mathcal{I}(M)$ ranges between 0 and 1, with 0 and 1 corresponding to degenerate and identity matrices respectively.
- Isometry gap: $-\log \mathcal{I}(M)$ lies between 0 and ∞, with 0 and ∞ indicating identity and degenerate matrices respectively.

These properties suggest that $\mathcal{I}(M)$ is a suitable function for measuring how close a matrix is to being an isometry, i.e., a transformation that preserves distances between metric spaces. Moreover, there is a clear link between isometry and normalization, which we will explore in the next section. We will often measure how far a matrix is from isometry by its negative log $-\log \mathcal{I}(M)$, which we will call *isometry gap*.

Basic properties of isometry It is straightforward to check isometry obeys the following basic isometry-preserving properties:

Lemma 4.2. *For PSD matrix M, the isometry defined in equation 4.2 obeys the following properties: 1) scale-invariance $\mathcal{I}(cM) = \mathcal{I}(M)$, 2) only takes value in the unit range $\mathcal{I}(M) \in [0, 1]$ 3) it takes its maximum value if and only if M is identity $\mathcal{I}(M) = 1 \iff M = I_n$, and 3) takes minimum value if and only if M is degenerate $\mathcal{I}(M) = 0$.*

Proof of Lemma 4.2. The scale-invariance is trivially true as scaling M by any constant will scale $\det(M)^{1/n}$ and $\mathrm{tr}(M)$ by the same amount. The proof of other properties is a straightforward consequence of writing the isometry in terms of the eigenvalues $\mathcal{I}(M) = (\prod_i \lambda_i)^{1/n}/(\frac{1}{n}\sum_i \lambda_i)$, where λ_i's are eigenvalues of M. By arithmetic vs geometric mean inequality over the eigenvalues we have $(\prod_i \lambda_i)^{1/n} \leq \frac{1}{n}\sum_i \lambda_i)$, which proves that $\mathcal{I}(M) \in [0, 1]$. Furthermore, the inequality is tight if and only if the values are all equal $\lambda_1 = \cdots = \lambda_n$, which holds only for an identity $M = I_n$. Finally, the isometry is zero if and only if at least one eigenvalue is zero, which is the case for degenerate matrix M. □

4.2 Isometry bias of normalization

This notion of isometry has a remarkable property: if we normalize each point by its Euclidean norm then the isometry of their associated Gram matrix does not decrease. We formalize this property in the following theorem.

Theorem 4.3. *Given n samples $\{x_i\}_{i\leq n} \subset \mathbb{R}^d \setminus \{\mathbf{0}_d\}$, and their projection onto the unit sphere $\widetilde{x}_i := x_i/\|x_i\|$, and their respective Gram matrices $G = \left[\langle x_i, x_j\rangle\right]_{i,j\leq n}$ and $\widetilde{G} = \left[\langle \widetilde{x_i}, \widetilde{x_j}\rangle\right]_{i,j\leq n}$. The isometry of Gram matrices obeys*

$$\mathcal{I}\left(\widetilde{G}\right) \geq \mathcal{I}(G)\left(1 + \frac{\frac{1}{n}\sum_i^n (a_i - \bar{a})^2}{\bar{a}^2}\right), \qquad \textit{where } a_i := \|x_i\|, \bar{a} := \frac{1}{n}\sum_i^n a_i. \tag{4.3}$$

Theorem 4.3 shows a subtle property of normalization: Because the terms $(a_i - \bar{a})^2$ are always non-negative, the left-hand side is always greater than or equal to $\mathcal{I}(G)$. It further quantifies the improvement in isometry as a function of variation of norms. The terms $\bar{a}$ and $\frac{1}{n}\sum_i^n (a_i - \bar{a})^2$ correspond to the sample mean and variance of $a_1, \ldots, a_n$. Thus, the more diverse the norms, the larger the improvement in isometry.

Proof of Theorem 4.3. Define $D := \operatorname{diag}(a_1/\sqrt{d}, \ldots, a_n/\sqrt{d})$. Observe that $C = D\widetilde{G}D$, implying $\det(G) = \det(\widetilde{G})\det(D)^2$. Because $\widetilde{x_i}$'s have norm $\sqrt{d}$, diagonals of Gram after normalization are constant $\widetilde{G}_{ii} = d$, implying $\frac{1}{n}\operatorname{tr}(\widetilde{G}) = d$. We have

$$\frac{\mathcal{I}(\widetilde{G})}{\mathcal{I}(G)} = \frac{\frac{1}{n}\operatorname{tr}(G)}{\frac{1}{n}\operatorname{tr}(\widetilde{G})} \frac{\det(\widetilde{G})^{1/n}}{\det(G)^{1/n}} \tag{4.4}$$

$$= \frac{\frac{1}{n}\sum_i^n a_i^2}{d} \frac{\det(\widetilde{G})^{1/n}}{\det(\widetilde{G})^{1/n}(d^{-n}\prod_i^n a_i^2)^{1/n}} \tag{4.5}$$

$$= \frac{(\frac{1}{n}\sum_i a_i)^2}{(\prod_i^n a_i)^{2/n}} \frac{\frac{1}{d}\sum_i^n a_i^2}{(\frac{1}{n}\sum_i a_i)^2} \tag{4.6}$$

$$= 1 + \frac{\frac{1}{n}\sum_i^n (a_i - \bar{a})^2}{\bar{a}^2}, \qquad \bar{a} := \frac{1}{n}\sum_i^n a_i \tag{4.7}$$

□

4.3 Implications for normalization layers

Theorem 4.3 allows us to highlight the isometry bias of layer and batch normalization.

Corollary 4.4. *Consider n vectors before and after layer-normalization $\{x_i\}_{i\le n} \subset \mathbb{R}^d \setminus \{\mathbf{0}_d\}$ and $\{\widetilde{x}_i\}_{i\le n}, \widetilde{x_i} := \ln x_i$. Define their respective Gram matrices $G := [\langle x_i, x_j\rangle]_{i,j\le n}$, and $\widetilde{G} := [\langle \widetilde{x}_i, \widetilde{x}_j\rangle]_{i,j\le n}$. We have:*

$$\mathcal{I}\left(\widetilde{G}\right) \ge \mathcal{I}(G)\left(1 + \frac{\frac{1}{n}\sum_i^n (a_i - \bar{a})^2}{\bar{a}^2}\right), \qquad \text{where } a_i := \|x_i\|, \bar{a} := \frac{1}{n}\sum_i^n a_i.$$

Observe that we can view the layer normalization as a projection onto the $\sqrt{d}$-sphere, which is equivalent to the unit-norm projection in Theorem 4.3 up to a constant scale factor. Since isometry is scale-invariant, this implies that corollary 4.4 follows directly from Theorem 4.3.

Moreover, corollary 4.4 shows that the isometry of layer normalization is deterministic and does not rely on random weights. This means that layer normalization always preserves or enhances

the isometry of representations, even during training. We provide empirical evidence for this in discussion.

Despite the seemingly vast differences between layer normalization and batch normalization, the following corollary shows an intimate link between them through the prism of isometry.

Corollary 4.5. *Given n samples in a mini-batch before $X \in \mathbb{R}^{d\times n}$, and after normalization $\widetilde{X} = \mathrm{BN}X$ and define covariance matrices $C := XX^\top$ and $\widetilde{C} := XX^\top$. We have:*

$$\mathcal{I}\left(\widetilde{C}\right) \geq \mathcal{I}(C)\left(1 + \frac{\frac{1}{d}\sum_i^d (a_i - \bar{a})^2}{\bar{a}^2}\right), \qquad \textit{where } a_i := \|X_{i\cdot}\|, \bar{a} := \frac{1}{n}\sum_{i=1}^d a_i.$$

Gram matrices of networks with batch normalization have been the subject of many previous studies at network initialization: it has been postulated that BN prevents rank collapse issue [Dan+20] and that it orthogonalizes the representations [DJB21], and that it imposes isometry [Yan+19]. The isometry results implied by Corollaries 4.5 give a geometric interpretation of all these findings: it is straightforward to verify that maximum isometry is achieved with identity Gram matrix, which implies orthogonalization of these results. Rather strikingly, while all the results stated before have been of a probabilistic nature, the improved isometry of the Gram matrix implied by Corollary 4.5 holds deterministically.

5

Fixed-points and global convergence of deep representations

The study of neural networks has increasingly focused on understanding the internal mechanisms that govern their learning and generalization capabilities. A central aspect of this inquiry is how neural networks transform and preserve input data structure as it passes through multiple layers. One powerful approach to studying these transformations is through the lens of kernel methods, which have long been used in machine learning to understand the relationships between data points in high-dimensional spaces [SS02; SS04].

Studying neural networks from the perspective of kernels has been the subject of many theoretical studies. This perspective has led to significant advancements, such as the development of Neural Tangent Kernels (NTKs) [JGH18] and Convolutional Kernel Networks [Mai+14], or in the notion of dual kernel [DFS16]. The idea that neural networks, especially in the infinite-width limit, can be understood through kernels has been explored in various recent works [Lee+19b; Aro+19b; Yan19]. Moreover, the use of kernel methods to measure the similarity between inputs as the network processes them provides a powerful means of understanding the inductive biases and representative capacity of neural networks, particularly in the context of deep learning [Zha+17; Zha+21; CS09].

However, the question of how the inner product between hidden layer representations evolves across layers and whether it converges globally to a fixed point remains an important but underex-

plored area of study [SMG13a; Sch+17; PSG18]. While previous research has provided insights into local behaviors [Yan+19], a global analysis of fixed points and convergence properties of kernel sequences, particularly in the presence of nonlinear activations, still needs to be improved.

This chapter addresses this gap by introducing and analyzing the evolution of the kernel through layers. The kernel sequence tracks the similarity between two inputs. More concretely, we consider the kernel sequence $k(h^\ell(x), h^\ell(y))$, where k denotes some notion of similarity, such as inner product or cosine similarity, and $h^\ell(x)$ and $h^\ell(y)$ denote layer ℓ representation for respective inputs x and y. Understanding whether and how this sequence converges to a fixed point as the network depth increases is crucial for uncovering the inherent implicit biases of deep networks.

Our analysis builds upon foundational work in understanding the role of activation functions in neural networks, such as the use of ReLU and its variants [GBB11; NH10] and the exploration of nonlinear activations in large-scale models [RZL17; CUH15]. Additionally, the interplay between activation functions and normalization techniques, such as layer normalization [BKH16], plays a critical role in shaping the convergence behavior of kernel sequences [Kla+17; HDR19]. Moreover, the study of neural networks through the lens of kernel methods is deeply connected to earlier work on Gaussian processes [WR06] and the dynamics of signal propagation in deep networks [Poo+16; Rag+17]. By analyzing how these kernels evolve across layers, particularly in terms of activation functions, we contribute to a deeper understanding of the dynamics within deep neural networks.

Throughout this chapter, we assume that the network operates under the mean field regime, which simplifies the analysis by tending width to infinity and making stochastic sequences deterministic.

5.1 Preliminaries

In this section, we introduce the fundamental concepts, notations, and definitions used throughout this chapter.

We consider a feed-forward neural network with L layers, each with width d. The network takes an input vector $x \in \mathbb{R}^d$ and maps it to an output vector $h^L(x) \in \mathbb{R}^d$ through a series of transformations. The hidden representations at each layer ℓ are denoted by $h^\ell(x)$. The transformation at each layer is composed of a linear transformation followed by a nonlinear activation function ϕ. Mathematically, the hidden representation at layer ℓ is given by:

$$h^\ell(x) = \phi(W^\ell h^{\ell-1}(x)), \qquad W^\ell \in \mathbb{R}^{d\times d}, \tag{5.1}$$

where $h^0(x) := x$ is set to input, and elements of W^ℓ are drawn i.i.d. from $N(0, 1/d)$. The activation function $\phi : \mathbb{R} \to \mathbb{R}$ denotes the activation applied element-wise. In some variations of the MLP, we will use normalization layers to adjust the activations at each layer, namely Layer Normalization (LN) or Root Mean Squared (RMS) normalization. Finally, the neural kernel between two inputs x and y at layer ℓ is defined as:

$$\rho_\ell = \frac{\langle h^\ell(x), h^\ell(y)\rangle}{\|h^\ell(x)\|\|h^\ell(y)\|}, \tag{5.2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, and ρ_0 corresponds to similiarty between input samples x and y.

The main goal of this chapter is to analyze sequence $\{\rho_\ell\}$, at initialization, when the weights are still random. The motivation for this analysis is to understand if various architectural choices, namely activation or number of layers, lead to specific biases of the kernel towards a certain fixed point. Namely, if there is a bias towards zero, it would imply that at initialization, the representations become more orthogonal. In contrast, a positive fixed point would mean the representations become more similar.

In the current setup, the sequence $\{\rho_\ell\}$ is a stochastic sequence or a Markov chain due to the random weights at each layer. However, we will show that under the mean field regime, the sequence $\{\rho_\ell\}$ converges to a deterministic sequence, and we will analyze the properties of this deterministic sequence.

5.2 Mean field regime

In this section, we conduct a mean field analysis of multilayer perceptrons (MLPs) to explore the neural kernel's fixed-point behavior as the network depth increases. This approach allows us to gain insight into the global dynamics of neural networks, mainly how the similarity between two input samples evolves as they pass through successive network layers. Now, we can state the mean field regime for the kernel sequence, stating that in this regime, the sequence becomes deterministic.

Proposition 5.1. *Under the mean field regime, i.e., $d \to \infty$, the kernel sequence ρ_ℓ of an MLP with activation ϕ that obeys $\mathbb{E}_{X\sim N(0,1)}\phi(X)^2 = 1$. Then the sequence evolves deterministically as follows:*

$$\rho_\ell = \mathbb{E}_{X,Y}[\phi(X)\phi(Y)], \qquad \textit{where } X, Y \sim N(0,1),\ \mathbb{E}XY = \rho_{\ell-1}.$$

where the initial value ρ_0 corresponds to the input.

Historically, conditions similar to $\mathbb{E}\phi(X)^2 = 1$ have been used to prevent the forward pass from vanishing or exploding. For example, applying ReLU will zero out half of the activations, which will lead to a vanishing norm of forward representations, and Kaiming He's initialization He et al. [He+16a] addresses that by scaling weights to maintain consistent forward pass norms across layers. This principle is further refined in a self-normalizing activation Klambauer et al. [Kla+17], ensuring consistent mean and variances between pre- and post-activations.

Proof. The proof is a straightforward application of the law of large numbers, as the mean field regime implies that the sample means converge to the population means. Let us inductively assume that at layer ℓ the it holds that $\frac{1}{d}\|h^\ell(x)\|^2 = 1$ and $\frac{1}{d}\|h^\ell(y)\|^2 = 1$, and $\frac{1}{d}\langle h^\ell(x), h^\ell(y)\rangle = \rho_\ell$. Thus, if X and Y denote the pre-activations for a given unit for the two inputs at layer ℓ, they follow standard Gaussian distribution and have covariance ρ_ℓ. By construction we have $\mathbb{E}\phi(X)^2 = \mathbb{E}\phi(Y)^2 = 1$, and $\mathbb{E}\phi(X)\phi(Y) = \rho_{\ell+1}$. Finally, since each hidden unit is independent of others, based on the law of large numbers, we can conclude the samples means will converge to their expectation $\frac{1}{d}\|h^{\ell+1}(x)\|^2 = 1$, $\frac{1}{d}\|h^{\ell+1}(y)\|^2 = 1$, and $\frac{1}{d}\langle h^{\ell+1}(x), h^{\ell+1}(y)\rangle = \rho_{\ell+1}$. This concludes the proof. □

As you can see, this proposition shows that the kernel sequence ρ_ℓ converges to a deterministic sequence. We can relax the conditions on the weights to allow any distribution with zero mean and variance $1/d$, such as uniform distribution.

Proposition 5.2. *Under the mean field regime, i.e., $d \to \infty$, the kernel sequence ρ_ℓ of an MLP with activation ϕ that obeys $\mathbb{E}\phi(X)^2 = 1$. If each element of the weights is drawn i.i.d. from a distribution with zero mean and $1/d$ variance, then the sequence evolves deterministically as given by Proposition 5.1.*

Proof. The key observation is that the pre-activations to each unit can be written as the sum of i.i.d. elements; by Central Limit Theorem, we can conclude that as $d \to \infty$, their distribution converges to a normal distribution, which allows us to apply Proposition 5.1. □

Propositions 5.1 and 5.2 under the condition that under some normality conditions on activations, showed the sequence evolution is entirely determined by a single parameter $\rho_{\ell-1}$, which evolves deterministically, Inspired by this property, we define it as the mapping between the covariance of pre-activations and the covariance of post-activations.

Definition 5.3. Given two random variables X, Y with covariance ρ, and activation ϕ, define the *kernel map* κ as the mapping between the covariance of pre-activations and the covariance of post-activations:

$$\kappa(\rho) := \mathbb{E}_{X,Y}[\phi(X)\phi(Y)], \qquad \text{where } X, Y \sim N(0,1),\ \rho := \mathbb{E}XY. \tag{5.3}$$

With this definition, we can express the result of Propositions 5.1 and 5.2 in terms of the kernel map. Under the same setting as Proposition 5.2, the kernel sequence $\{\rho_\ell\}$ evolves deterministically according to the kernel map κ:

$$\rho_\ell = \kappa(\rho_{\ell-1}), \tag{5.4}$$

where κ is defined in Definition 5.3. Thus, in order to understand the convergence of sequence $\{\rho_\ell\}$, we can study the fixed-points of the kernel map κ, which are the values of ρ^* that satisfy $\kappa(\rho^*) = \rho^*$. With the assumptoin that $\mathbb{E}\phi(X)^2 = 1$, the kernel map κ is a mapping between $[-1, -1]$ to itself. Thus, Brower's fixed-point theorem implies that the kernel map κ has at least one fixed point ρ^*. But as we will show, there is potentially more than one fixed point, and it will be interesting to understand which ones are locally or globally attractive.

5.3 Hermite expansion of activation functions

Hermite polynomials possess completeness and orthogonality under the Gaussian weight kernel. This means that any function in the space of square-integrable functions with respect to the Gaussian kernel can be expressed as a linear combination of Hermite polynomials. The square integrability of an activation function ensures that it does not lead to exploding activations, as non-square integrable functions suggest heavy-tailed post-activations that lack second moments, and it holds for all activations that are used in practice. We use the *normalized* Hermite polynomials and their coefficients.

Definition 5.4. Define normalized Hermite polynomials $\mathrm{he}_k(x)$ as a scaled version of the probabilist's Hermite polynomials $\mathrm{He}_k(x)$, as follows

$$\mathrm{He}_k(x) := (-1)^k e^{x^2/2} \frac{d^k}{dx^k} e^{-x^2/2}, \qquad \mathrm{he}_k(x) := \frac{1}{\sqrt{k!}} \mathrm{he}_k(x).$$

Here is a list of the first few normalized Hermite polynomials:

Table 5.1: Hermite polynomials and their normalized versions

Order	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
$\mathrm{He}_k(x)$	1	x	(x^2-1)	(x^3-3x)	(x^4-6x^2+3)
$\mathrm{he}_k(x)$	1	x	$\frac{1}{\sqrt{2}}(x^2-1)$	$\frac{1}{\sqrt{6}}(x^3-3x)$	$\frac{1}{\sqrt{24}}(x^4-6x^2+3)$

Most importantly, Hermite polynomials satisfy the orthogonality property:

$$\mathbb{E}_{x\sim N(0,1)}\left[\mathrm{He}_k(x)\,\mathrm{He}_l(x)\right]=k!\delta_{kl},\qquad \mathbb{E}_{x\sim N(0,1)}\left[\mathrm{he}_k(x)\,\mathrm{he}_l(x)\right]=\delta_{kl},\tag{5.5}$$

where δ_{kl} is the Dirac delta. Here we can see the motivation behind our normalization, as it cancels out with the factorial term in the orthogonality property.

Based on this property, we can express any function ϕ as a linear combination of Hermite polynomials:

Definition 5.5. Given an activation function ϕ that is square-integrable with respect to the Gaussian kernel $\int_{-\infty}^{\infty}\phi(x)^2e^{-x^2/2}dx<\infty$, the Hermite expansion of ϕ is defined as:

$$\phi(x)=\sum_{k=0}^{\infty}c_k\,\mathrm{he}_k(x),\qquad c_k=\mathbb{E}_{X\sim N(0,1)}\left[\phi(X)\,\mathrm{he}_k(X)\right],$$

where $\mathrm{he}_k(x)$ are the normalized Hermite polynomials and c_k are the Hermite coefficients.

Besides orthogonality, Hermite polynomials have another magical property that is crucial for our later analysis.

Lemma 5.6 (Consequence of Mehler's kernel)**.** *If $X,Y\sim N(0,1)$ with covariance $\mathbb{E}XY=\rho$ we have*

$$\mathbb{E}_{X,Y}\,\mathrm{he}_n(X)\,\mathrm{he}_k(Y)=\rho^n\delta_{nk}$$

where δ_{nk} is the Dirac delta.

This lemma states that given two Gaussian random variables X,Y with covariance ρ,, the expectation of the product of Hermite polynomials is zero unless the indices are equal.

Proof of Lemma 5.6. The property can be deduced from Mehler's formula [Meh66]. The formula states that

$$\begin{aligned}&\frac{1}{\sqrt{1-\rho^2}}\exp\left(-\frac{\rho^2(x^2+y^2)-2xy\rho}{2(1-\rho^2)}\right)\\&=\sum_{m=0}^{\infty}\mathrm{he}_m(x)\,\mathrm{he}_m(y),\end{aligned}$$

where the $m!$ factor difference is due to the definition of Hermite polynomials with an additional $1/\sqrt{m!}$ compared to the one used in Mehler's kernel. Observe that the left-hand side is equal to $p(x,y)/p(x)p(y)$, where $p(x,y)$ is the joint PDF of (X,Y), and $p(x), p(y)$ are PDF of X and Y respectively. Therefore, we can take the expectation using the expansion

$$\begin{aligned}\mathbb{E}_{X,Y}\left[\operatorname{he}_n(X)\operatorname{he}_k(Y)\right] &= \int \operatorname{he}_n(x)\operatorname{he}_k(y)p(x,y)dxdy \\ &= \sum_{m=0}^{\infty}\rho^m \int \operatorname{he}_n(x)\operatorname{he}_k(y)\operatorname{he}_m(x)\operatorname{he}_m(y)dp(x)dp(y) \\ &= \sum_{m=0}^{\infty}\rho^m \mathbb{E}_{X\sim N(0,1)}\left[\operatorname{he}_n(X)\operatorname{he}_m(X)\right]\mathbb{E}_{Y\sim N(0,1)}\left[\operatorname{he}_k(Y)\operatorname{he}_m(Y)\right] \\ &= \rho^n\delta_{nk}\end{aligned}$$

where in the last line we used the orthogonality property $\mathbb{E}_{X\sim N(0,1)}H_k(x)H_n(X) = \delta_{nk}$. □

Lemma 5.6 is crucial for our theory. Based on this lemma, we can express the kernel map in terms of the Hermite coefficients, showing a particular structure of the kernel map.

Corollary 5.7. *Given $X, Y \sim N(0,1)$ with covariance $\mathbb{E}XY = \rho$, and $\phi(X) = \sum_{k=0}^{\infty} c_k \operatorname{he}_k(X)$, we have*

$$\kappa(\rho) = \mathbb{E}_{X\sim N(0,1)}\left[\phi(X)\phi(X)\right] = \sum_{k=0}^{\infty} c_k^2\rho^k.$$

Proof of Corollary 5.7. Using Lemma 5.6 we have

$$\mathbb{E}_{X\sim N(0,1)}\left[\phi(X)\phi(X)\right] = \mathbb{E}_{X\sim N(0,1)}\left[\sum_{k=0}^{\infty} c_k \operatorname{he}_k(X)\sum_{l=0}^{\infty} c_l \operatorname{he}_l(X)\right] = \sum_{k=0}^{\infty} c_k^2\rho^k.$$

□

These algebraic properteis of kernel map will be crucial in our analysis of the convergence of the kernel sequence to a fixed point. Before we turn our attention to convergence, let us draw some links between the properties of activation functions, its kernel map, and its Hermite coefficients.

5.4 Globally contracting kernel to zero

Recall that for activations that obey $\mathbb{E}\phi(X)^2 = 1$,, the kernel map κ is a power series with non-negative coefficients $\sum_k c_k\rho^k$ that maps $[-1,1]$ to itself, and obeys $\kappa(1) = 1$. This immediately

Table 5.2: Properties of activations in terms of Hermite coefficients and kernel map

Property	Activation ϕ	Hermite coefficients $\{c_k\}_{k\geq 0}$	Kernel map κ
Centered	$\mathbb{E}\phi(X) = 0$	$c_0 = 0$	$\kappa(0) = 0$
Stable	$\mathbb{E}\phi(X)^2 = 1$	$\sum_{k=0}^{\infty} c_k^2 = 1$	$\kappa(1) = 1$
Non-linear	$\phi(x)$ is non-linear	$\sum_{k=2}^{\infty} c_k^2 > 0$	$\kappa(\rho)$ is non-linear

reveals that $\rho = 1$ is a fixed point of the kernel map. This is unsurprising, as it reaffirms the fact that if two inputs are identical with unit variance, the post-activations will also be identical. However, it is far more interesting whether this fixed point is globally attractive and how the kernel map influences the convergence rate to this fixed point. In a similar vein, we can ask if there are other fixed points and how the kernel map influences the convergence rate to these fixed points.

In the same vein as previous chapters, we are primarily interested in seeing which properties in the kernel maps lead to convergence towards orthogonality. In other words, we are primarily interested in conditions that lead to $\rho^* = 0$ being a fixed point and under which conditions this fixed point is globally attractive.

Finally, we can state one of the central results of this chapter, which characterizes the global convergence of the kernel map towards the fixed point $\rho^* = 0$.

Theorem 5.8. *Let ϕ be an activaiton with kernel map κ that is centered $\kappa(0) = 0$, and $\kappa(1) = 1$. Let ρ_ℓ be the kernel sequence $\rho_{\ell+1} = \kappa(\rho_\ell)$, given initial value ρ_0. Then, the sequence contracts towards fixed-point zero $\rho^* = 0$ with rate α:*

$$\frac{|\rho_\ell|}{1 - |\rho_\ell|} \leq \frac{|\rho_0|}{1 - |\rho_0|}\alpha^\ell, \qquad \alpha := \frac{1}{2 - \kappa'(0)}, \tag{5.6}$$

where α is strictly less than one if the activation is non-linear. The only other fixed points can be $\rho^* = 1$ or $\rho^* = -1$,, and none of them is locally or globally attractive.

Note that the statement becomes vacuous if the initial covariance is $\rho_0 = 1$ or $\rho_0 = -1$, as in these cases, the right-hand side goes to infinity, but the theorem is non-vacuous for all $\rho_0 \in (-1, 1)$. This is unsurprising, as the kernel map cannot separate the two inputs if they are identical or anti-identical. In plain words, this states that if the two samples are not identical, by iteratively passing them through a nonlinear activation, the covariance between them will contract with rate α in expectation. While it may not seem immediately obvious why $\alpha < 1$ for non-linear activations, the following remark makes the connection between the two.

Remark 5.9. Let us assume $\kappa(\rho) = \sum_{k=0}^{\infty} c_k^2 \rho^k$, where c_k denote Hermite coefficients. Note that $\kappa'(0) = c_1^2 \leq \sum_{k=0}^{\infty} c_k^2 = \kappa(1) = 1$. Now, if we have $\kappa'(1) = 1$, it implies that $c_k = 0$ for all $k \geq 2$. This is a contradiction with the assumption that the activation is non-linear.

Another observation is that rather than proving convergence directly, we showed contraction under the potential $|\rho|/(1-|\rho|)$. Because $|\rho|/(1-|\rho|)$ is a monotonically increasing with $|\rho|$, the contraction of $|\rho_\ell|/(1-|\rho_\ell|)$ implies contraction of $|\rho_\ell|$ towards zero. We can state the contraction directly in terms of the sequence ρ_ℓ.

Corollary 5.10. *Under the same setting as in Theorem 5.8, the sequence $\rho_\ell = \kappa(\rho_{\ell-1})$ converges to zero with rate $|\rho_\ell| \leq \rho_0/(1-|\rho_0|)\alpha^\ell$.*

The proof follows directly from Theorem 5.8 and the inequality $|\rho_\ell| \leq |\rho_\ell| \leq (1-|\rho_\ell|)$.

Proof idea: The main proof idea of this theorem is to show that upon applying the kernel map κ to the covariance of pre-activations, the potential function $|\rho|/(1-|\rho|)$ decreases with rate α, and applying an induction over ℓ. Condition $\kappa(1) = \mathbb{E}\phi(X)^2 = 1$ ensures that activations are stable and do not explode or vanish, allowing us to apply the single step contraction inductively. The proof of the single step follows from the properties of Hermite polynomials and the orthogonality of the kernel map. The non-linearity condition is essential to have $\alpha < 1$, which rules out identity-like activations that do not change the value. The centered condition $\kappa(0) = \mathbb{E}\phi(X) = 0$ ensures that ρ^* is a fixed point.

Some activations, such as SeLU Klambauer et al. [Kla+17], which is a self-normalizing activation function that has $\mathbb{E}\phi(X)^1 = 0$.

Proof of Theorem 5.8. The main technical part of the proof is to prove that:

$$\Psi(\kappa(\rho)) \leq \alpha \Psi(\rho), \qquad \Psi(\rho) = \frac{|\rho|}{1-|\rho|}.$$

Base don the given assumptions we have $c_0^2 = \kappa(0) = 0$ and based on the assumption $\mathbb{E}\phi^2(X) = 1$ and properties of Hermite polynomials we have $\kappa(1) = \sum_{k=1}^{\infty} c_k^2 = 1$.

First, we will consider positive $\rho \in [0, 1)$, for which we have

$$\begin{aligned}
\frac{\Psi(\kappa(\rho))}{\Psi(\rho)} &= \left(\frac{\kappa(\rho)}{1-\kappa(\rho)}\right)\left(\frac{\rho}{1-\rho}\right)^{-1} \\
&= \frac{\rho^{-1}\sum_{k=1}^{\infty} c_k^2 \rho^k}{(1-\rho)^{-1}\sum_{k=1}^{\infty} c_k^2(1-\rho^k)} \\
&= \frac{\sum_{k=1}^{\infty} c_k^2 \rho^{k-1}}{\sum_{k=1}^{\infty} c_k^2 \left(\frac{1-\rho^k}{1-\rho}\right)} \\
&= \frac{\sum_{k=1}^{\infty} c_k^2 \rho^{k-1}}{\sum_{k=1}^{\infty} c_k^2 \left(\sum_{i=0}^{k-1} \rho^i\right)} \\
&\leq \frac{\sum_{k=1}^{\infty} c_k^2 \rho^{k-1}}{\sum_{k=1}^{\infty} c_k^2 \rho^{k-1} + \sum_{k=2}^{\infty} c_k^2} \\
&\leq \max_{\rho\in[0,1]} \frac{\sum_{k=1}^{\infty} c_k^2 \rho^{k-1}}{\sum_{k=1}^{\infty} c_k^2 \rho^{k-1} + \sum_{k=2}^{\infty} c_k^2} \\
&= \frac{\sum_{k=1}^{\infty} c_k^2}{2\sum_{k=1}^{\infty} c_k^2 - c_1^2} \\
&= \frac{1}{2-\kappa'(0)} =: \alpha.
\end{aligned}$$

Now, we can use the fact that the norm of the sum of values is bounded by the sum of their norms to argue that

$$|\kappa(\rho)| = |\sum_{k=1}^{\infty} c_k^2 \rho^k| \leq \sum_{k=1}^{\infty} c_k^2 |\rho|^k = \kappa(|\rho|)$$

Because $x \mapsto x/(1-x)$ is monotonically increasing for $x \in [0, 1]$, for all $\rho \in [-1, 1]$ we have

$$\frac{|\kappa(\rho)|}{1-|\kappa(\rho)|} \leq \frac{\kappa(|\rho|)}{1-\kappa(|\rho|)} \leq \frac{|\rho|}{1-|\rho|}\alpha.$$

Where we invoked the inequality that was proven for $\rho \in [0, 1]$, Plugging in the definition of Ψ we have proven that for all ρ we have

$$\Psi(\kappa(\rho)) \leq \alpha\Psi(\rho).$$

We can conclude the proof by induction over ℓ.

Other fixed-points: What remains to show is that the only possible fixed points are $\rho^* = 0, 1, -1,$, and only $1, -1$ are neither locally or globally attractive.

First, by contraction result we have so far, for any $\rho \in (-1, 1), \rho \neq 0$, then $|\kappa(\rho)|$ will be strictly smaller than $|\rho|$, which contradicts with it being a fixed point. That only leaves $\{-1, 0, 1\}$ as possible fixed points. Now, we will prove that $-1, 1$ are not locally attractive fixed points.

For $\rho^* = 1$ to be locally attracting, we must have $|\kappa'(\rho^*)| < 1$, and by construction we have $\kappa'(0) < 1$, However, this implies in some small ϵ there is $\rho_0 \in (0, \epsilon)$ and $\rho_1 \in (1-\epsilon, 1)$, such that we have $\kappa(\rho_0) < \rho_0$ and $\kappa(\rho_1) > \rho_1$. Now, by continuity of $\kappa(\rho)$ there must be a point $\rho_2 \in (\rho_0, \rho_1)$ such that $\kappa(\rho_2) = \rho_2$, which contradicts the assumption that $\rho^* = 1$ is the only fixed point. The same argument can be made for $\rho^* = -1$.

Thus, we have shown that only $\{-1, 0, 1\}$ can be fixed-point, and only $\rho^* = 0$ is globally attractive. □

5.5 Convergence of the kernel with general activations

So far, we only discussed the convergence of the kernel map for activations that are centered and the convergence of their kernel sequence towards zero. However, we can extend this analysis to general activations that are not centered. In this section, we will show that for any activation function that is non-linear, there is a unique fixed point ρ^* that is globally attractive.

Theorem 5.11. *Let κ be the kernel map satisfying $\kappa(1) = 1$. Define the kernel sequence $\rho_{\ell+1} = \kappa(\rho_\ell)$, with initial step $\rho_0 \in (-1, 1)$. Assuming that the activation is non-linear, there is a unique ρ^* that is globally contracting, which is necessarily non-negative $\rho^* \in [0, 1]$. The only other fixed-points distinct from ρ^*, could be -1 or 1, neither of which are locally or globally contracting. Furthermore, we have the following contraction rate towards ρ^*:*

1. *If $\kappa(0) = 0$, then $\rho^* = 0$ is an attracting with rate*
$$\frac{|\rho_\ell|}{1 - |\rho_\ell|} \leq \frac{|\rho_0|}{1 - |\rho_0|}(1/\kappa'(1))^\ell \tag{5.7}$$
2. *If $\kappa(0) > 0$ and $\kappa'(1) < 1$, then $\rho^* = 1$ is an attracting, with rate*
$$|\rho_\ell - 1| \leq |\rho_0 - 1|(\kappa'(1))^\ell \tag{5.8}$$
3. *If $\kappa(0) > 0$, and $\kappa'(1) = 1$ then $\rho^* = 1$ is attracting with rate*
$$|\rho_\ell - 1| \leq \frac{|\rho_0 - 1|}{\ell\alpha|\rho_0 - 1| + 1}, \qquad \alpha = 1 - \kappa(0) - \kappa'(0). \tag{5.9}$$
4. *If $\kappa(0) > 0$, and $\kappa'(1) > 1$ then the attracting fixed point is necessarily in the range $\rho^* \in (0, 1)$, we have*
$$|\rho_\ell - \rho^*| \leq \frac{|\rho_0 - \rho^*|}{1 - |\rho_0|}\alpha^\ell \qquad \alpha = \max\left\{1 - \kappa(0), \kappa'(\rho^*), \frac{1 - \rho^*}{2 - \kappa'(\rho^*)}\right\}, \tag{5.10}$$
where α is strictly less than one if the activation is non-linear.

Remark 5.12. One of the most striking results of this theorem is that for any non-linear activation, there is a unique fixed point ρ^* that is locally or globally attractive. Furthermore, the fixed point is necessarily non-negative. This implies that for any MLP with non-linear activations, the covariance of pre-activations will converge towards a fixed point, which is always non-negative.

Remark 5.13. The assumptions laid out in the theorem are nearly tight. For example, the non-linearity assumption is necessary to rule out identity activation, where every point in $[-1, 1]$ is a fixed point. Furthermore, considering odd activations such as $\phi(x) = x^3$, and we can see that if $\rho_0 \in \{-1, 1\}$, the sequence will remain the same $\rho_\ell = \rho_0$, and hence will not converge, which implies the condition that $\rho_0 \in (-1, 1)$ is necessary.

Remark 5.14. Note that all convergence rates are of exponential form $\|\rho_\ell - \rho^*\| = O(\alpha^\ell)$, where $\alpha < 1$ if the activation is non-linear, with the exception of the case where $\kappa'(1) = 1$, where the rate is of the form $O(1/\ell)$.

Proof. We will cases individually, starting with the first case that falls direclty under a previous theorem.

Case 1: $\kappa(0) = 0$

We can observe that the case where $\kappa(0) = 0$, falls directly under Theorem 5.8, and thus there is no need to prove it again.

Cases 2,3: $\kappa(0) > 0$ and $\kappa'(1) \leq 1$

In this part, we jointly consider two cases where $\kappa(0) > 0$ and $\kappa'(1) < 1$, and $\kappa'(1) = 1$. Let us consider the ratio between distances $|\rho_\ell - 1|$:

$$\begin{aligned}
\frac{|\kappa(\rho) - 1|}{|\rho - 1|} &= \frac{1 - \kappa(\rho)}{1 - \rho} \\
&= \frac{\kappa(1) - \kappa(\rho)}{1 - \rho} \\
&= \frac{\sum_{k=1}^{\infty} c_k^2 (1 - \rho^k)}{1 - \rho} \\
&= \sum_{k=1}^{\infty} c_k^2 \sum_{i=0}^{k-1} \rho^i \\
\implies \frac{|\kappa(\rho) - 1|}{|\rho - 1|} &= \kappa'(1) - \kappa'(1) + \sum_{k=1}^{\infty} c_k^2 \sum_{i=k}^{\infty} \rho^i \\
&= \kappa'(1) - \sum_{k=1}^{\infty} c_k^2 \Big(k - \sum_{i=k}^{\infty} \rho^i\Big)
\end{aligned}$$

Clearly, the term $k - \sum_{i=k}^{\infty} \rho^i$ is is always non-negative, implying that if $\kappa'(1) < 1$ we have the contraction

$$\frac{|\kappa(\rho) - 1|}{|\rho - 1|} \leq \kappa'(1) \implies |\rho_\ell - 1| \leq |\rho_0 - 1| \kappa'(1)^\ell.$$

Otherwise, if $\kappa'(1) = 1$, we have

$$\frac{|\kappa(\rho) - 1|}{|\rho - 1|} = 1 - \sum_{k=1}^{\infty} c_k^2 \Big(k - \sum_{i=0}^{k-1} \rho^i\Big).$$

Now, observe that the first term for $k = 1$ is zero. Furthermore, the sequence $k - \sum_{i=0}^{k-1} \rho^i$ is monotonically increasing in k. Thus, the smallest value the weighted sum can achieve is if all of the weights of terms above $k \geq 2$ are concentrated in $k = 2$, which leads to the contraction

$$\frac{|\kappa(\rho) - 1|}{|\rho - 1|} \leq 1 - (1 - c_0^2 - c_1^2)(2 - 1 - \rho) = 1 - (1 - c_0^2 - c_1^2)(1 - \rho).$$

Now, define sequence $x_\ell := 1 - \rho_\ell$, and observe that we have

$$x_{\ell+1} \leq x_\ell (1 - \alpha x_\ell), \qquad \alpha = 1 - c_0^2 - c_1^2,$$

where $\alpha > 0$ if the activation is non-linear. We can prove inductively that

$$x_\ell \leq \frac{x_0}{\ell \alpha x_0 + 1}.$$

If we plug in the definition of x_n we have proven

$$|\rho_\ell - 1| \leq \frac{|\rho_0 - 1|}{\ell \alpha |\rho_0 - 1| + 1}.$$

Case 4: $\kappa(0) > 0$ and $\kappa'(1) > 1$

The main strategy is to prove some contraction of $\kappa(\rho)$ towards ρ^*, under the kernel map κ. In other words, we need to show $|\kappa(\rho) - \rho^*|$ is smaller than $|\rho - \rho^*|$ under some potential. First, we assume there is a ρ^* such that $\kappa'(\rho^*) < 1$, and show this contraction, and later prove its existence and uniqeness.

To prove contraction towards ρ^* when $\kappa'(\rho^*) < 1$, we consider three cases: 1) If $\rho > \rho^*$, 2) If $\rho \in [0, \rho^*]$, and 3) If $\rho < 0$. However, the bounds will be of different potential forms and will have to be combined later. Let $\kappa(\rho) = \sum_{k=0}^{\infty} c_k^2 \rho^k$ be the kernel map with $\kappa(1) = 1$ with fixed-point ρ^* that satisfies $\kappa'(\rho^*) < 1$.

- $\rho \geq \rho^*$**.** we will prove:

$$\frac{|\kappa(\rho) - \rho^*|}{1 - \kappa(\rho)} \leq \frac{|\rho - \rho^*|}{1 - \rho} \kappa'(\rho^*)$$

  We have the series expansion around ρ^*: $\kappa(\rho) = \rho^* + \sum_{k=1}^{\infty} a_k (\rho - \rho^*)^k$. For points $\rho \geq \rho^*$, we will have $\kappa(\rho) \geq \rho^*$, thus we can write

$$
\begin{aligned}
\frac{\kappa(\rho)-\rho^*}{1-\kappa(\rho)} &= \frac{\sum_{k=1}^{\infty} a_k(\rho-\rho^*)^k}{\kappa(1)-\kappa(\rho^*)} \\
&= \frac{\sum_{k=1}^{\infty} a_k(\rho-\rho^*)^k}{\sum_{k=1}^{\infty} a_k(1-\rho^*)^k - \sum_{k=1}^{\infty} a_k(\rho-\rho^*)^k} \\
&= \frac{(\rho-\rho^*)\sum_{k=1}^{\infty} a_k(\rho-\rho^*)^{k-1}}{(1-\rho)\sum_{k=1}^{\infty} a_k(\sum_{i=0}^{k-1}(1-\rho^*)^i(\rho-\rho^*)^{k-1-i})} \\
&= \frac{\rho-\rho^*}{1-\rho}\cdot\frac{\sum_{k=1}^{\infty} a_k(\rho-\rho^*)^{k-1}}{\sum_{k=1}^{\infty} a_k(\sum_{i=0}^{k-1}(1-\rho^*)^i(\rho-\rho^*)^{k-1-i})} \\
&\leq \frac{\rho-\rho^*}{1-\rho}\frac{\sum_{k=1}^{\infty} a_k(\rho-\rho^*)^{k-1}}{\sum_{k=2}^{\infty} a_k(1-\rho^*)^{k-1} + \sum_{k=1}^{\infty} a_k(\rho-\rho^*)^{k-1}} \\
&\leq \max_{\rho\in[\rho^*,1]} \frac{\rho-\rho^*}{1-\rho}\frac{\sum_{k=1}^{\infty} a_k(\rho-\rho^*)^{k-1}}{\sum_{k=2}^{\infty} a_k(1-\rho^*)^{k-1} + \sum_{k=1}^{\infty} a_k(\rho-\rho^*)^{k-1}} \\
&\leq \frac{\rho-\rho^*}{1-\rho}\frac{\sum_{k=1}^{\infty} a_k(1-\rho^*)^{k-1}}{\sum_{k=2}^{\infty} a_k(1-\rho^*)^{k-1} + \sum_{k=1}^{\infty} a_k(1-\rho^*)^{k-1}} \\
&= \frac{\rho-\rho^*}{1-\rho}\frac{\kappa(1)-\rho^*}{2\kappa(1)-\kappa'(\rho^*)} \\
&= \frac{\rho-\rho^*}{1-\rho}\frac{1-\rho^*}{2-\kappa'(\rho^*)}
\end{aligned}
$$

Thus, we have proven that

$$
\rho \geq \rho^* \implies \frac{|\kappa(\rho)-\rho^*|}{1-\kappa(\rho)} \leq \frac{|\rho-\rho^*|}{1-\rho}\frac{1-\rho^*}{2-\kappa'(\rho^*)}
$$

- $0 \leq \rho \leq \rho^*$**.** Consider $\rho \in [0, \rho^*]$. For these $\kappa'(\rho)$ is always monotonically increasing, implying that $\kappa'(\rho) \leq \kappa'(\rho^*) < 1$. Thus, $|\kappa(\rho)-\rho^*| \leq \kappa'(\rho^*)|\rho-\rho^*|$. This implies that in this range $|\kappa(\rho)-\rho^*|$ will contract with a rate $\kappa'(\rho^*)$:

$$
0 \leq \rho \leq \rho^* \implies |\kappa(\rho)-\rho^*| \leq \kappa'(\rho^*)|\rho-\rho^*|
$$

- $-1 \leq \rho \leq 0$**.** Finally, let us consider $\rho \leq 0$. Recall that we have $\kappa(1) = 1$. Thus, we can express $\kappa(\rho) - 1$ as product of $(\rho - 1)$ with some power series $q(\rho)$:

$$
\kappa(\rho) - 1 = (\rho-1)q(\rho), \quad q(\rho) = \sum_{k=0}^{\infty} b_k \rho^k.
$$

In fact, we can expand $\kappa(\rho)$ in terms of these new coefficients

$$
\kappa(\rho) = 1 - b_0 + \sum_{k=0}^{\infty}(b_k - b_{k+1})\rho^k
$$

Due to the non-negativity of coefficients of κ, we can conclude $1 \geq b_0 \geq b_1 \geq ....$ Based on this observation, for $0 < \rho < 1$, we can conclude $q(-\rho) = b_0 - b_1\rho + b_2\rho^2 - ... \leq b_0$. Because we can pair each odd and even term $-b_k\rho^k + b_{k+1}\rho^{k+1}$ for all odd k, and because coefficients $b_k \geq b_{k+1}$ and $\rho^k \geq \rho^{k+1}$ for $\rho \in [0, 1]$, we can argue $q(-\rho) \leq b_0 = 1 - c_0^2$. Now, plugging this value into the kernel map for $0 < \rho < 1$, we have:

$$\begin{aligned}
\kappa(-\rho) &= 1 - (1+\rho)q(-\rho) \\
&\geq 1 - (1+\rho)(1-c_0^2) \\
&= 1 - 1 - \rho - c_0^2(1+\rho) \\
&\implies \kappa(-\rho) + \rho \geq c_0^2(1+\rho) \\
&\implies -\rho + c_0^2(1+\rho) \leq \kappa(-\rho) \leq \kappa(\rho)
\end{aligned}$$

Now, if we assume $\kappa(-\rho) \leq \rho^*$ then

$$\frac{|\kappa(\rho) - \rho^*|}{|-\rho - \rho^*|} = \frac{\rho^* - \kappa(-\rho)}{\rho^* + \rho} \leq \frac{\rho^* + \rho - c_0^2(1+\rho)}{\rho + \rho^*} = 1 - \frac{c_0^2(1+\rho)}{\rho + \rho^*} \leq 1 - c_0^2$$

Now, if we assume $\kappa(-\rho) \geq \rho^*$, knowing that $\kappa(-\rho) \leq \kappa(\rho)$, which necessitates $\rho \geq \rho^*$, which implies $\kappa(\rho) \leq \rho$. Thus, we have

$$\frac{|\kappa(-\rho) - \rho^*|}{|-\rho - \rho^*|} = \frac{\kappa(-\rho) - \rho^*}{\rho + \rho^*} \leq \frac{\kappa(\rho) - \rho^*}{\rho + \rho^*} \leq \frac{\rho - \rho^*}{\rho + \rho^*} \leq \frac{1 - \rho^*}{1 + \rho^*} \leq 1 - \rho^*$$

Combining both cases we have

$$\rho \leq 0 \implies \frac{|\kappa(\rho) - \rho^*|}{|\rho - \rho^*|} \leq 1 - \min(c_0^2, \rho^*) = 1 - \min(\kappa(0), \rho^*).$$

We can further prove that $\rho^* = \kappa(\rho^*) = k(0) +$ non-negative terms, which implies that $\rho^* \geq k(0)$. Thus, we can conclude that

$$\rho \leq 0 \implies |\kappa(\rho) - \rho^*| \leq (1 - k(0))|\rho - \rho^*|$$

Positivity, uniqueness, and existence of a globally attractive fixed-point Here, the goal is to prove there is exactly one point $\rho^* \in [0, 1]$ such that $\kappa(\rho^*) = \rho^*$ and $\kappa'(\rho^*) < 1$. We will prove the properties of positivity, uniqueness, and existence separately.

Positivity: Let us assume that $\rho^* \leq 0$. is a fixed-point. Then, we can apply the contraction rate proven for Case 3, which shows that $\kappa(\rho^*) \geq \rho^* + k(0) > \rho^*$, which is a contradiction.

Uniqueness: Assume that there are two fixed points ρ_1 and ρ_2 that satisfy $\kappa'(\rho_1), \kappa'(\rho_2) < 1$. Let us assume wlog that $\rho_1 < \rho_2$. Then we can invoke the contraction rate proven so far to argue

that all points in $(-1, 1)$, including $\rho \in (\rho_1, \rho_2)$ are attracted towards both ρ_1 and ρ_2, which is a contradiction. Thus, there can be at most one fixed point.

Existence of ρ^*: Because $\kappa(1) = 1$ the set of all fixed-points is non-empty. Let us assume that ρ^* is the first (smallest) fixed point of $\kappa(\rho^*) = \rho^*$, which because of the positivity result is necessarily $\rho^* > 0$. If we assume that $\kappa'(\rho^*) > 1$, then in the small ϵ-neighborhood of it $\rho_1 \in (\rho^* - \epsilon, \rho^*)$ we have $\kappa(\rho_1) < \rho_1$. Because $\kappa(\rho)$ is continuous, and is above identity line at $\rho = 0$ and under identity line $\rho = \rho_1$, there must be a point $0 < \rho_2 < \rho_1$ where it is at identity $\kappa(\rho_2) = \rho_2$, which is a contradiction with assumption that ρ^* is the smallest fixed point. Thus, we must have $\kappa'(\rho^*) \leq 1$. If we assume that $\kappa'(\rho^*) = 1$, then the κ must align with the identity line from ρ^* to 1, which implies that all higher order terms $c_k, k \geq 2$ must be zero, which in turn implies that κ is a linear function. This is a contradiction with the assumption that the activation is non-linear. Thus, we must have $\kappa'(\rho^*) < 1$, which proves the desired existence.

Combining the cases Let us summarize the results so far. We have proven the existence of a unique fixed-point $\rho^* \in [0, 1]$ such that $\kappa'(\rho^*) < 1$, and we have proven contraction rates for each of the three cases.

$$\begin{cases} \frac{|\kappa(\rho)-\rho^*|}{1-\kappa(\rho)} \leq \frac{|\rho-\rho^*|}{1-\rho} \frac{1-\rho^*}{2-\kappa'(\rho^*)} & \text{if } \rho \geq \rho^* \\ |\kappa(\rho) - \rho^*| \leq \kappa'(\rho^*)|\rho - \rho^*| & \text{if } 0 \leq \rho \leq \rho^* \\ |\kappa(\rho) - \rho^*| \leq (1 - k(0))|\rho - \rho^*| & \text{if } \rho \leq 0 \end{cases}$$

Again, we can consider two cases, if $\kappa'(1) < 1$ or $\kappa'(1) > 1$. First, note that $\kappa'(1) = 1$ is impossible, as it would imply $\kappa'(1) = \sum_{k=1}^{\infty} k c_k^2 = \sum_{k=0} c_0^2$, which implies that all $c_k = 0$, which is a contradiction with the assumption that the activation is non-linear.

Let us now define the joint decay rate:

$$\alpha = \max \left\{ 1 - k(0), \kappa'(\rho^*), \frac{1 - \rho^*}{1 - \kappa'(\rho^*)} \right\}$$

In other words, this is the worst-case rate for any of the above cases.

Now, let us assume we are starting from initial ρ_0 and define $\rho_\ell = \kappa(\rho_{\ell-1})$. One important observation is that if we have $\rho_0 \geq \rho^*$ then by monotonicity of κ in the $[0, 1]$ range, it will remain the same range, and similarly if $\rho_0 \in [0, \rho^*]$ it will remain in the same range. Thus, from that index onwards, we can apply the contraction rate of the respective case. The only case that there might be a transition is if $\rho_0 < 0$.

Assuming that $\rho_0 < 0$, let ρ_ℓ be the first index that we have $\rho_\ell \geq 0$. Thus, from ρ_0 to ρ_ℓ we can apply the contraction rate of the third case:

$$|\rho_\ell - \rho^*| \leq |\rho_0 - \rho^*|\alpha^\ell$$

Now, we have two possiblities, either $\rho_\ell \geq \rho^*$ or $\rho_\ell \leq \rho^*$. If $\rho_\ell \geq \rho^*$, we can apply the contraction rate of the first case, and if $\rho_\ell \leq \rho^*$ we can apply the contraction rate of the second case:

$$\begin{cases} |\rho_L - \rho^*| \leq |\rho_\ell - \rho^*|\alpha^{L-\ell} & 0 \leq \rho_\ell \leq \rho^* \\ \frac{|\rho_L - \rho^*|}{1-\rho_L} \leq \frac{|\rho_\ell - \rho^*|}{1-\rho_\ell}\alpha^{L-\ell} & \rho_\ell \geq \rho^* \end{cases}$$

If we plug in our contraction up to step ℓ and use the fact that the norm of the sequence is non-increasing $|\rho_0| \geq \rho_\ell$, we have

$$\begin{cases} |\rho_L - \rho^*| \leq |\rho_0 - \rho^*|\alpha^{L} & 0 \leq \rho_\ell \leq \rho^* \\ \frac{|\rho_L - \rho^*|}{1-\rho_L} \leq \frac{|\rho_0 - \rho^*|}{1-|\rho_0|}\alpha^{L} & \rho_\ell \geq \rho^* \end{cases}$$

We can now take the worst case of these two and conclude that

$$|\rho_L - \rho^*| \leq \frac{|\rho_0 - \rho^*|}{1 - |\rho_0|}\alpha^L$$

So far in the proof, we assumed the existence of ρ^* that obeys $\kappa'(\rho^*) < 1$. We can now prove that such a fixed point exists. It is unique, and it is necessarily in the range $\rho^* \in [0, 1]$.

□

5.6 Considering normalization layers

In this section, we will provide the statement and proof of the most general theorem of this chapter, which characterizes the global convergence towards fixed points of the kernel map, considering the MLPs with normalization layers.

Before that, let us consider some variants of the MLP with normalization layers. We consider two types of normalization layers, with normalization layers before or after the activation:

$$\text{Pre-act normalization:} \qquad h^\ell = \phi\left(\text{norm}\left(W^\ell h^{\ell-1}\right)\right), \tag{5.11}$$

$$\text{Post-act normalization:} \qquad h^\ell = \text{norm}\left(\phi\left(W^\ell h^{\ell-1}\right)\right), \tag{5.12}$$

where norm can be Layer Normalization (LN) or Root Mean Squared (RMS) normalization, defined as:

$$\mathrm{LN}(z) = \frac{z - \bar{z}}{\sqrt{\frac{1}{d}\sum_{i=1}^{d}(z_i - \bar{z})^2}}, \qquad \mathrm{RMS}(z) = \frac{z}{\sqrt{\frac{1}{d}\sum_{i=1}^{d} z_i^2}}. \tag{5.13}$$

Finally, we can prove a more general global contration statement for the kernel map in the mean field regime, with the normalization layers.

Theorem 5.15. *Consider an MLP with activation ϕ that is square-integrable $\mathbb{E}\phi(X)^2 < \infty$, and is nonlinear. For any of the given choices of normalization, the following additional conditions on the activations*

- *No normalization: It holds $\mathbb{E}\phi(X) = 0$, and $\mathbb{E}\phi(X)^2 = 1$.*
- *MLP with pre-act or post-act RMS, or pre-act LN: It holds $\mathbb{E}\phi(X) = 0$.*
- *MLP with post-act LN: No further assumptions on activation than stated.*

the following contraction holds in the mean field regime:

$$\frac{|\rho_\ell|}{1 - |\rho_\ell|} \leq \frac{|\rho_0|}{1 - |\rho_0|}\alpha^\ell, \qquad \alpha = \frac{\kappa(1) - \kappa(0)}{\kappa'(1)}. \tag{5.14}$$

where α is strictly less than one if the activation is non-linear.

Proof. The first part, i.e., the no normalization case, follows directly from Theorem 5.8. The remainder of the proof follows from the following observation:

Let us assume that ϕ is square-integrable $\mathbb{E}\phi(X)^2$. Then, the following hold in the mean field regime and $X \sim N(0, 1)$. We have

- For MLP with post-RMS, pre-RMS, and pre-LN, the normalization layer will converge to $z \to z/\sqrt{\mathbb{E}\phi(X)^2}$.
- In MLP with post-LN, the normalization layer will converge to $z \to (z - \mu)/\phi$, where the $\mu = \mathbb{E}[\phi(X)]$ and $\phi = \sqrt{\mathbb{E}(\phi(X) - m)^2}$.

Let us consider each case separately.

- **Pre-act RMS and pre-act LN:** Let us define $s := \sqrt{\mathbb{E}\phi(X)^2}$. Let us inductively assume that $\frac{1}{d}\|h^{\ell-1}\| = s$. Thus, elements of $a := W^\ell h^{\ell-1}$ are drawn i.i.d/ from $N(0, s^2)$. Thus, in the mean field regime, the centering step of LN will be ineffective, and the normalization step of both LN and RMS will converge to s. Thus, normalization will result in a Gaussian vector with its elements drawn i.i.d. from $N(0, 1)$. Thus, the norm of the activations will be, by definition, for each element, we have $\mathbb{E}\phi(a_i)^2 = s^2$. Finally, in the mean field regime, the sample mean will converge to population mean $\frac{1}{d}\|h^\ell\|^2 = \frac{1}{d}\|\phi(a)\|^2 = s^2$, proving the induction hypothesis. In the process, we have also proved that in the mean field, both pre-act RMS and pre-act LN will converge to $z \to z/s$.

- **Post-act RMS:** Again, let us define $s := \sqrt{\mathbb{E}\phi(X)^2}$. Because $h^{\ell-1}$ is defined after the normalization. we have $\frac{1}{d}\|h^{\ell-1}\| = 1$. Thus, elements of $a := W^\ell h^{\ell-1}$ are drawn i.i.d/ from $N(0, 1)$. Thus, after going through activation $\phi(a^\ell)$, for each element have $\mathbb{E}\phi(a_i)^2 = s^2$. and therefore in the mean field, the sample mean will converge to population mean $\frac{1}{d}\|h^\ell\|^2 = s^2$, which proves the claim. Thus, RMS normalization will converge to $z \to z/s$.

- **Post-act LN:** The analysis is similar to above, except for the last step, layer normalization will subtract sample mean and sample standard deviation. In the mean field, both of these quantities converge to their population counterparts $\mu = \mathbb{E}\phi(X)$, and $\sigma = \sqrt{\mathbb{E}(\phi(X) - \mu)^2}$, proving that the post-act LN will act like $z \to (z - \mu)/\sigma$.

Now that we have established that in all these cases, we can replace the normalization and activation layer with a new activation Psi, that obeys $\mathbb{E}\psi(X)^2 = 1$, up to some absolute constant scale. The only difference is that in post-act LN, this new activation will also be centered $\mathbb{E}\psi(X) = 0$, while in the other three cases, if and only if ϕ is centered:

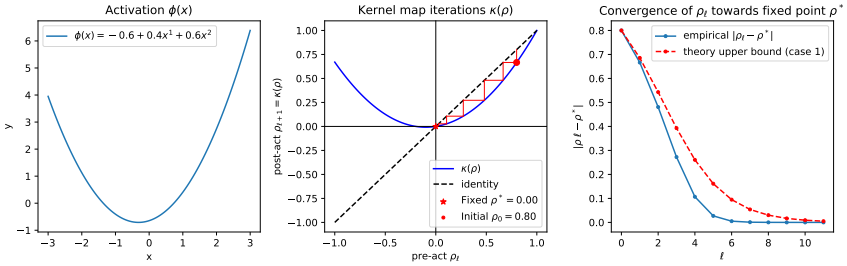
- For pre-act RMS and pre-act LN, and post-act RMS, we can construct a new activation ψ that obeys $\mathbb{E}\psi(X)^2 = 1$, and if ϕ is centered, the new activation is centered $\mathbb{E}\psi(X) = 0$. Thus, assuming that the original activation is centered, we can involve Theorem 5.8 to conclude the proof.

- For post-act LN, we will get a new activation ψ that obeys $\mathbb{E}\psi(X)^2 = 1$, and $\mathbb{E}\psi(X) = 0$. Thus we can invoke Theorem 5.8 to conclude the proof.

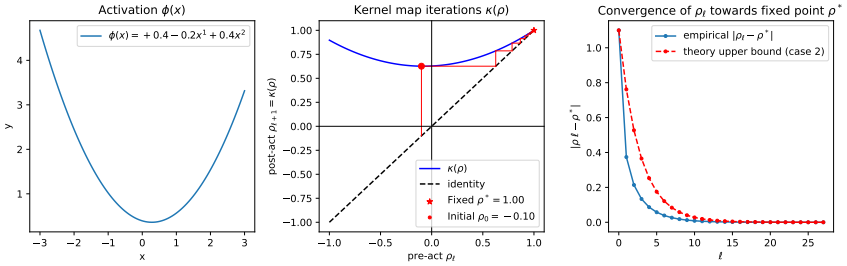
□

5.7 Validation of the global convergence theorem

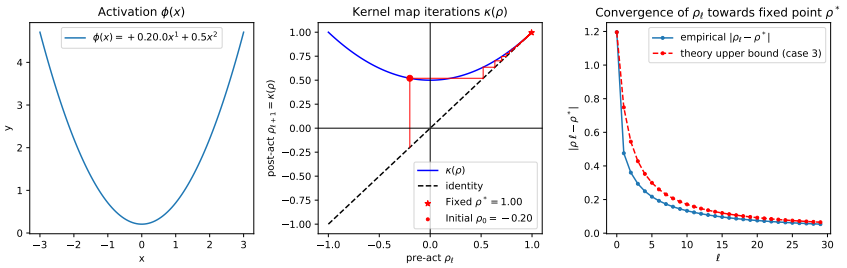
Here we will provide some numerical validation of the global convergence theorem. We will consider the kernel map for some custom made and commonly used activations, and show that the fixed point ρ^* is globally attractive. We will consider the kernel map for the following activations: $\phi(x) = \tanh(x), \phi(x) = \max(0, x), \phi(x) = \exp(x), \phi(x) = \text{GELU}(x)$, which will correspond to the four cases of the theorem. See Figure 5.1 and Figure 5.2 for the results.



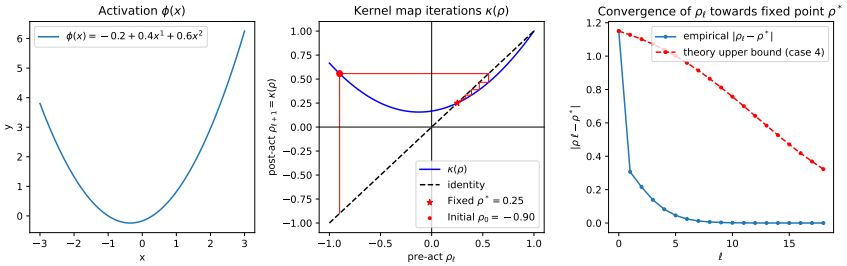
(a) Case 1



(b) Case 2

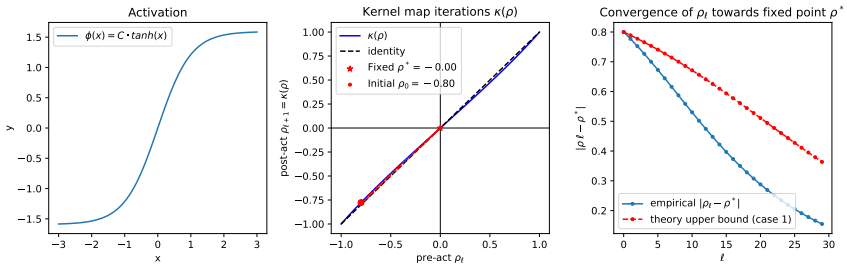


(c) Case 3

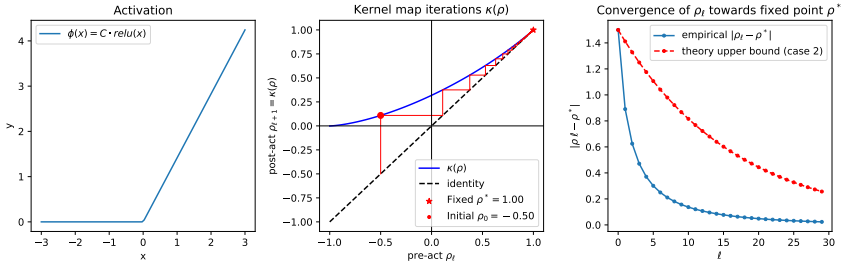


(d) Case 4

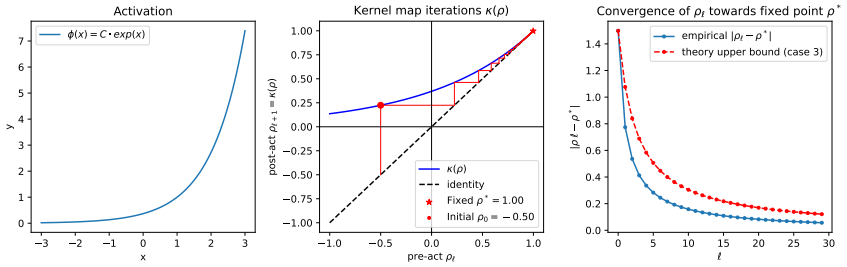
Figure 5.1: Validation of Theorem 5.8, each corresponding to one of the four cases of the theorem. Left column shows the activation ϕ. The middle and right columns show the kernel map show fixed point iteration starting from ρ_0, and applying $\rho_{\ell+1} = \kappa(\rho_\ell)$ for many steps. The middle column shows the kernel map, while the right column shows the distance to the fixed point ρ^*.



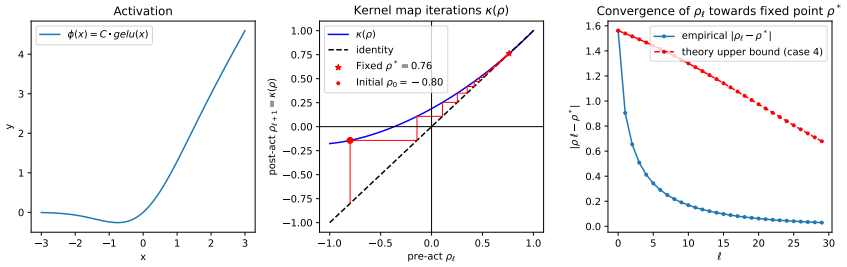
(a) Case 1. Tanh activation.



(b) Case 2: ReLU activation.



(c) Case 3: Exponential activation.



(d) Case 4: GELU activation.

Figure 5.2: Same as Figure 5.1 for some commonly used activations. Note that because the raw activations do not necessariy obey $\mathbb{E}\phi^2(X) = 1$, we have to scale by some contant C to make the activations obey the conditions of the theorem.

6

Batch normalization without gradient explosion

What if we could train even deeper neural networks? Increasing depth empowers neural networks, by turning them into powerful data processing machines. For example, increasing depth allows large language models (LLMs) to capture longer structural dependencies [Dev+18; Liu+19; Bro+20; Goy+21; Raf+20]. Also, sufficiently deep convolutional networks can outperform humans in image classification [Liu+22; Woo+23; Wu+21]. Nevertheless, increasing depth imposes an inevitable computational challenge: deeper networks are harder to optimize. In fact, standard optimization methods exhibit a slower convergence when training deep neural networks. Hence, computation has become a barrier in deep learning, demanding extensive research and engineering.

A critical problem is that deeper networks suffer from the omnipresent issue of rank collapse at initialization: the outputs become collinear for different inputs as the network grows in depth [SMG13a]. Rank collapse is not only present in MLPs and convolutional networks [Fen+22; SMG13a; Dan+20], but also in transformer architectures [DCL21; Noc+22b]. This issue significantly contributes to the training slowdown of deep neural networks. Hence, it has become the focus of theoretical and experimental studies [SMG13a; Fen+22; DJB21; Noc+22b].

One of the most successful methods to avoid rank collapse is Batch Normalization (BN) [IS15], as

Code is available at: https://github.com/alexandrumeterez/bngrad

proven in a number of theoretical studies [Yan+19; Dan+20; DJB21]. Normalization imposes a particular bias across the layers of neural networks [JDB23b]. More precisely, the representations of a batch of inputs become more orthogonal after each normalization [JDB23b]. This orthogonalization effect precisely avoids the rank collapse of deep neural networks at initialization [Yan+19; JDB23b; DJB21; JDB23a].

While batch normalization effectively avoids rank collapse, it causes numerical issues. The existing literature proves that batch normalization layers cause exploding gradients in MLPs in an activation-independent manner [Yan+19]. Gradient explosion limits increasing depth by causing numerical issues during backpropagation. For networks without batch normalization, there are effective approaches to avoid gradient explosion and vanishing, such as tuning the variance of the random weights based on the activation and the network width [He+15; GB10a]. However, such methods cannot avoid gradient explosion in the presence of batch normalization [Yan+19]. Thus, the following important question remains unanswered:

Is there any network with batch normalization without gradient explosion and rank collapse issues?

Contributions. We answer the above question affirmatively by giving a specific MLP construction initialized with orthogonal random weight matrices, rather than Gaussian. To show that the MLP still has optimal signal propagation, we prove that the MLP output embeddings become isometric (equation 6.5), implying the output representations becomes more orthogonal with depth. For a batch of linearly independent inputs, we prove

$$\mathbb{E}\Big[\text{isometry gap}\Big] = \mathcal{O}\left(e^{-\text{depth}/C}\right), \tag{6.1}$$

where C is a constant depending only on the network width and input and the expectation is taken over the random weight matrices. Thus, for sufficiently deep networks, the representations rapidly approach an orthogonal matrix. While Daneshmand, Joudaki, and Bach [DJB21] prove that the outputs converge to within an $\mathcal{O}(\text{width}^{-1/2})$-ball close to orthogonality, we prove that the output representations become perfectly orthogonal in the infinite depth limit. This perfect orthogonalization turns out to be key in proving our result about avoiding gradient explosion. In fact, for MLPs initialized with Gaussian weights and BN, Yang et al. [Yan+19, Theorem 3.9] prove that the gradients explode at an exponential rate in depth. In a striking contrast, we prove that gradients of an MLP with BN and orthogonal weights remain bounded as

$$\mathbb{E}\Big[\log\left(\text{gradient norm for each layer}\right)\Big] = \mathcal{O}\left(\text{width}^5\right). \tag{6.2}$$

Thus, the gradient is bounded by a constant that only depends on the network width where the expectation is taken over the random weight matrices. It is worth noting that both isometry and log-norm gradient bounds are derived *non-asymptotically*. Thus, in contrast to the previously studied mean field or infinite width regime, our theoretical results hold in practical settings where the width is finite.

The limitation of our theory is that it holds for a simplification in the BN module and linear activations. However, our results provide guidelines to avoid gradient explosion in MLPs with non-linear activations. We experimentally show that it is possible to avoid gradient explosion for certain non-linear activations with orthogonal random weights together with "activation shaping" [Mar+21]. Finally, we experimentally demonstrate that avoiding gradient explosion stabilizes the training of deep MLPs with BN.

6.1 Related work

The challenge of depth in learning. Large depth poses challenges for the optimization of neural networks, which becomes slower by increasing the number of layers. This depth related slowdown is mainly attributed to: (i) gradient vanishing/explosion, and (ii) the rank collapse of hidden representations. (i) Gradient vanishing and explosion is a classic problem in neural networks [Hoc98]. For some neural architectures, this issue can be effectively solved. For example, He et al. [He+15] propose a particular initialization scheme that avoids gradient vanishing/explosion for neural networks with rectifier non-linearities while Glorot and Bengio [GB10a] study the effect of initialization on sigmoidal activations. However, such initializations cannot avoid gradient explosion for networks with batch normalization [Yan+19; LDT21]. (ii) Saxe, McClelland, and Ganguli [SMG13a] demonstrate that outputs become independent from inputs with growing depth, which is called the rank collapse issue [Dan+20; DCL21]. Various techniques have been developed to avoid rank collapse such as batch normalization [IS15], residual connections [He+16b], and self-normalizing activations [Kla+17]. A related line of work has shown how signal propagation can be achieved without batch normalization in feed-forward networks [BD19] and ResNets [BDS21]. Other works on CNNs [BGS20] have shown that symmetry breaking is a vital element for achieving signal propagation with stable gradients in deep models. Here, we focus on batch normalization since our primary goal is to avoid the systemic issue of gradient explosion for batch normalization.

Initialization with orthogonal matrices. Saxe, McClelland, and Ganguli [SMG13a] propose initializing the weights with random orthogonal matrices for linear networks without normalization layers. Orthogonal matrices avoid the rank collapse issue in linear networks, thereby enabling a depth-independent training convergence. Pennington, Schoenholz, and Ganguli [PSG17] show that MLPs with sigmoidal activations achieve dynamical isometry when initialized with orthogonal weights. Similar benefits have been achieved by initializing CNNs with orthogonal or almost orthogonal kernels [Xia+18; MM15], and by initializing RNN transition matrices with elements from the orthogonal and unitary ensembles [ASB16; LJH15; HSL16]. Similarly, we use orthogonal random matrices to avoid gradient explosion. What sets our study apart from this literature is that our focus is on batch normalization and the issue of gradient explosion.

Networks with linear activation functions. Due to its analytical simplicity, the identity function has been widely used in theoretical studies for neural networks. Studies on identity activations date back to at least two decades. Fukumizu [Fuk98] studies batch gradient descent in linear neural networks and its effect on overfitting and generalization. Baldi and Hornik [BH95] provide an overview over various theoretical manuscripts studying linear neural networks. Despite linearity, as Saxe, McClelland, and Ganguli [SMG13a] and Saxe, McClellans, and Ganguli [SMG13b] observe, the gradient dynamics in a linear MLP are highly nonlinear. In a line of work, Saxe, McClellans, and Ganguli [SMG13b], Saxe, McClelland, and Ganguli [SMG13a], and Saxe, McClelland, and Ganguli [SMG19] study the training dynamics of deep neural networks with identity activations and introduce the notion of dynamical isometry. Baldi and Hornik [BH89] and Yun, Sra, and Jadbabaie [YSJ17] study the mean squared error optimization landscape in linear MLPs. More recently, the optimum convergence rate of gradient descent in deep linear neural networks has been studied by Arora et al. [Aro+19a] and Shamir [Sha19]. Du and Hu [DH19] prove that under certain conditions on the model width and input degeneracy, linear MLPs with Xavier initialized weights [GB10a] converge linearly to the global optimum. Akin to these studies, we also analyze networks with linear activations. However, batch normalization is a non-linear function, hence the network we study in this chapter is a highly non-linear function of its inputs.

Mean field theory for random neural networks. The existing analyses for random networks often rely on mean field regimes where the network width tends to infinity [PSG17; Yan+19; LNR22; PW17]. However, there is a discrepancy between mean field regimes and the practical regime of finite width. While some analyses attempt to bridge this gap [JDB23a; DJB21], their

results rely on technical assumptions that are hard to validate. In contrast, our non-asymptotic results hold for standard neural networks used in practice. Namely, our main assumption for avoiding rank collapse and gradient explosion is that samples in the input batch are not linearly dependent, which we will show is necessary. To go beyond mean field regimes, we leverage recent theoretical advancements in Weingarten calculus [Wei78; Col03; BCS11; CŚ06; CMN22].

Mean field theory for random neural networks. The existing analyses for random networks often rely on mean field regimes where the network width tends to infinity [PSG17; Yan+19; LNR22; PW17]. However, there is a discrepancy between mean field regimes and the practical regime of finite width. While some analyses attempt to bridge this gap [JDB23a; DJB21], their results rely on technical assumptions that are hard to validate. In contrast, our non-asymptotic results hold for standard neural networks used in practice. Namely, our main assumption for avoiding rank collapse and gradient explosion is that samples in the input batch are not linearly dependent, which we will show is necessary. To go beyond mean field regimes, we leverage recent theoretical advancements in Weingarten calculus [Wei78; Col03; BCS11; CŚ06; CMN22].

6.2 Main results

We will develop our theory by constructing networks that do not suffer from gradient explosion (Sec. 6.2.3) and still orthogonalize (Sec. 6.2.1). The construction is similar to the network studied by Daneshmand, Joudaki, and Bach [DJB21]: an MLP with batch normalization and linear activations. Formally, let $X^\ell \in \mathbb{R}^{d\times n}$ denote the representation of n samples in $\mathbb{R}^d$ at layer ℓ, then

$$X_{\ell+1} = \textsc{Bn}(W_\ell X_\ell), \qquad \ell = 0, \dots, L, \tag{6.3}$$

where $W_\ell \in \mathbb{R}^{d\times d}$ are random weights, n is the mini-batch size and d is the feature dimension. Analogous to recent theoretical studies of batch normalization [DJB21; Dan+20], we define the BN operator $\textsc{Bn} : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ as

$$\textsc{Bn}(X) = \operatorname{diag}(XX^\top)^{-\frac{1}{2}} X, \quad \textsc{Bn}(X)_{ij} = \frac{X_{ij}}{\sqrt{\sum_{k=1}^d X_{ik}^2}}\,. \tag{6.4}$$

Note that compared to the standard BN operator, mean reduction in equation 6.4 is omitted. Our motivation for this modification, similar to Daneshmand, Joudaki, and Bach [DJB21], is purely technical and to streamline our theory. We will experimentally show that using standard BN

modules instead does not influence our results on gradient explosion and signal propagation (for more details see Figure C.10). A second minor difference is that in the denominator, we have omitted a $\frac{1}{n}$ factor. However, this only amounts to a constant scaling of the representations and does not affect our results.

Compared to Daneshmand, Joudaki, and Bach [DJB21], we need two main modifications to avoid gradient explosion: (i) $n = d$, and (ii) W_ℓ are random *orthogonal* matrices. More precisely, we assume the distribution of W_ℓ is the Haar measure over the orthogonal group denoted by $\mathbb{O}_d$ [CŚ06]. Such an initialization scheme is widely used in deep neural networks without batch normalization [SMG13a; Xia+18; PSG17]. For MLP networks with BN, we prove such initialization avoids the issue of gradient explosion, while simultaneously orthogonalizing the inputs.

6.2.1 Tracking signal propagation via orthogonality

As discussed, batch normalization has an important orthogonalization bias that influences training. Without normalization layers, representations in many architectures face the issue of rank-collapse, which happens when network outputs become collinear for arbitrary inputs, hence their directions become insensitive to the changes in the input. In contrast, the outputs in networks with batch normalization become increasingly orthogonal through the layers, thereby enhancing the signal propagation in depth [DJB21]. Thus, it is important to check whether the constructed network maintains the important property of orthogonalization.

Isometry gap. Our analysis relies on the notion of *isometry gap*, $\phi : \mathbb{R}^{d\times n} \to \mathbb{R}$, introduced by Joudaki, Daneshmand, and Bach [JDB23b]. Isometry gap is defined as

$$\phi(X) = -\log\left(\frac{\det(X^\top X)^{\frac{1}{d}}}{\frac{1}{d}\mathrm{tr}(X^\top X)}\right). \tag{6.5}$$

One can readily check that $\phi(X) \geq 0$ and it is zero when X is an orthogonal matrix, i.e., $XX^\top = I_d$. The *isometry* denoted by $\mathcal{I} : \mathbb{R}^{d\times n} \to \mathbb{R}$ is defined as $\mathcal{I}(X) = \exp(-\phi(X))$.

Geometric interpretation of isometry. While the formula for isometry gap may seem enigmatic at first, it has a simple geometric interpretation that makes it intuitively understandable. The determinant $\det(X^\top X) = \det(X)^2$ is the squared volume of the parallelepiped spanned by the columns of X, while $\mathrm{tr}(X^\top X)$ is the sum squared-norms of the columns of X. Thus, the

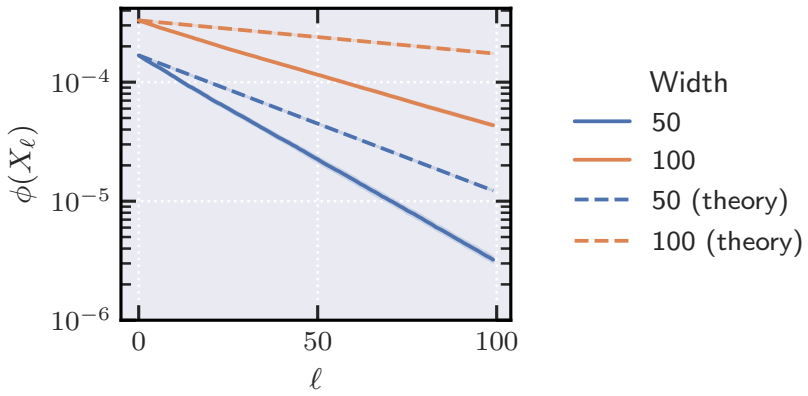


Figure 6.1: Isometry gap (y-axis, log-scale) in depth for an MLP with orthogonal weights, over randomly generated data. As predicted by Theorem 6.1, isometry gap of representations vanishes at an exponential rate. The solid traces are averaged over 10 independent runs, and the dashed traces show the theoretical prediction from Theorem 6.1.

ratio between the two provides a scale-invariant notion of volume and isometry. On the one hand, if there is any collinearity between the columns, the volume will vanish and the isometry gap will be infinity, $\psi(X) = \infty$. On the other hand, $\psi(X) = 0$ implies $X^\top X$ is a scaled identity matrix. We will prove ψ serves as a Lyapunov function for the chain of hidden representations $\{X_\ell\}_{\ell=0}^{\infty}$.

Theory for orthogonalization. The following theorem establishes the link between orthogonality of representations and depth.

Theorem 6.1. *There is an absolute constant C such that for any layer $\ell \leq L$ we have*

$$\mathbb{E}\phi(X_{\ell+1}) \leq \phi(X_0)e^{-\ell/k}, \qquad \textit{where} \qquad k := Cd^2(1 + d\phi(X_0)) \,. \tag{6.6}$$

Theorem 6.1 states that if the samples in the input batch are not linearly dependent, representations approach orthogonality at an exponential rate in depth. The orthogonalization in depth ensures the avoidance of the rank collapse of representations, which is a known barrier to training deep neural networks [Dan+20; SMG13a; Bjo+18].

Figure 6.1 compares the established theoretical decay rate of ψ with the practical rate. Interestingly, the plot confirms that the rate depends on width in practice, akin to the theoretical rate in Theorem 6.1. It is worth mentioning that the condition on the input samples to not be linearly dependent is necessary to establish this result. One can readily check that starting from a rank-deficient input, neither matrix products, nor batch-normalization operations can increase

the rank of the representations. Since this assumption is quantitative, we can numerically verify it by randomly drawing many input mini-batches and check if they are linearly dependent. For CIFAR10, CIFAR100, MNIST and FashionMNIST, we empirically tested that most batches across various batch sizes are full-rank (see Section C.4 for details on the average rank of a batch in these datasets).

Theorem 6.1 distinguishes itself from the existing orthogonalization results in the literature [Yan+19; JDB23b] as it is non-asymptotic and holds for networks with finite width. Since practical networks have finite width and depth, non-asymptotic results are crucial for their applicability to real-world settings. While Daneshmand, Joudaki, and Bach [DJB21] provide a non-asymptotic bound for orthogonalization, the main result relies on an assumption that is hard to verify.

Proof idea of Theorem 6.1. We leverage a recent result established by Joudaki, Daneshmand, and Bach [JDB23b], proving that the isometry gap does not decrease with BN layers. For all non-degenerate matrices $X \in \mathbb{R}^{d\times d}$, the following holds

$$\mathcal{I}(\textsc{Bn}(X)) \geq \left(1 + \frac{\text{variance}\{\|X_{j\cdot}\|\}_{j=1}^{d}}{(\text{mean}\{\|X_{j\cdot}\|\}_{j=1}^{d})^2}\right)\mathcal{I}(X)\,.$$

Using the above result, we can prove that matrix multiplication with orthogonal weights also does not decrease isometry as stated in the next lemma.

Lemma 6.2 (Isometry after rotation)**.** *Let $X \in \mathbb{R}^{d\times d}$ and $W \in \mathbb{R}^{d\times d}$ be an orthogonal matrix and $X' = WX$; then,*

$$\mathcal{I}(\textsc{Bn}(X')) \geq \left(1 + \frac{\text{variance}\{\|X'_{j\cdot}\|\}_{j=1}^{d}}{(\text{mean}\{\|X'_{j\cdot}\|\}_{j=1}^{d})^2}\right)\mathcal{I}(X)\,. \tag{6.7}$$

It is straightforward to check that there exists at least an orthogonal matrix W for which $\mathcal{I}(\textsc{Bn}(WX)) = 1$ (see Corollary C.3). Thus, $\mathcal{I}(\cdot)$ strictly increases for some weight matrices, as long as X is not orthogonal. When the distribution of W is the Haar measure over the orthogonal group, we can leverage recent developments in Weingarten calculus [Wei78; BCS11; CŚ06; CMN22] to calculate a rate for the isometry increase in expectation:

Theorem 6.3. *Suppose $W \sim \mathbb{O}_d$ is a matrix drawn from $\mathbb{O}_d$ such that the distribution of W and UW are the same for all orthogonal matrices U. Let $\{\lambda_i\}_{i=1}^{d}$ be the eigenvalues of $XX^\top$. Then,*

$$\mathbb{E}_W\left[\mathcal{I}(\textsc{Bn}(WX))\right] \geq \left(1 - \frac{\sum_{k=1}^{d}(\lambda_k - 1)^2}{2d^2(d+2)}\right)^{-1}\mathcal{I}(X) \tag{6.8}$$

holds for all $X = \textsc{Bn}(\cdot)$, with equality for orthogonal matrices.

The structure in X induced by BN ensures its eigenvalues lie in the interval $(0, 1]$, in that the multiplicative factor in the above inequality is always greater than one. In other words, $\mathcal{I}(\cdot)$ increases by a constant factor in expectation that depends on how close X is to an orthogonal matrix.

The connection between Theorem 6.3 and the main isometry gap bound stated in Theorem 6.1 is established in the following Corollary (recall $\psi = -\log \mathcal{I}$).

Corollary 6.4 (Isometry gap bound)**.** *Suppose the same setup as in Theorem 6.3, where* $X' = WX$*. Then, we have:*

$$\mathbb{E}_W[\phi(X')|X] \leq \phi(X) + \log\left(1 - \frac{\sum_k (\lambda_k - 1)^2}{2d^2(d+2)}\right). \tag{6.9}$$

Notice that the term $\frac{\sum_{k=1}^{d}(\lambda_k-1)^2}{2d^2(d+2)} = \mathcal{O}(\frac{1}{d})$, yielding $\log\left[1 - \frac{\sum_{k=1}^{d}(\lambda_k-1)^2}{2d^2(d+2)}\right] \leq 0$.

The rest of proof is based on an induction over the layers, presented in Appendices C.1 and C.2.

6.2.2 Orthogonalization and gradient explosion

There is a subtle connection between orthogonalization and gradient explosion. Suppose the input batch is rank-deficient, i.e., degenerate. As elaborated above, since all operations in our MLP can be formulated as matrix products, they cannot recover the rank of the representations, which thus remain degenerate. By perturbing the input such that it becomes full-rank, the output matrix becomes orthogonal, hence non-degenerate at an exponential rate in depth as proven in Theorem 6.1.

Thus, a slight change in inputs leads to a significant change in outputs from degeneracy to orthogonality. Considering that the gradient measures changes in the loss for infinitesimal inputs changes, the large changes in outputs potentially lead to gradient explosion. While this is only an intuitive argument, we observe that in practice the gradient does explode for degenerate inputs, as shown in Figure 6.2.

Nonetheless, in Figure 6.2 we observe that for non-degenerate inputs the gradient norm does not explode. In fact, we observe that inputs are often non-degenerate in practice (see Table C.1 for details). Thus, an important question is whether the gradient norm remains bounded for non-degenerate input batches. Remarkably, we can not empirically verify that for *all degenerate inputs* the gradient norm remains bounded. Therefore, a theoretical guarantee is necessary to ensure avoiding gradient explosion.

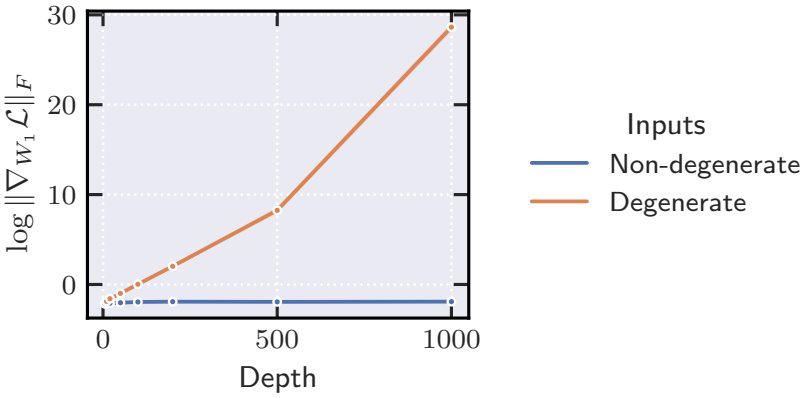


Figure 6.2: Logarithmic plot for the gradient norm of the first layer for networks with different number of layers evaluated on degenerate (orange) and non-degenerate (blue) inputs. The degenerate inputs contain repeated samples from CIFAR10 in the batch, measured at initialization for MLPs of various depths. While gradients explode for degenerate inputs, there is no explosion for non-degenerate inputs. Traces are averaged over 10 independent runs.

6.2.3 Avoiding gradient explosion in depth

So far, we have proven that the constructed network maintains the orthogonalization property of BN. Now, we turn our focus to the gradient analysis. The next theorem proves that the constructed network does not suffer from gradient explosion in depth for non-degenerate input matrices.

Theorem 6.5. *Let the loss function $\mathcal{L} : \mathbb{R}^{d\times d} \to \mathbb{R}$ be $\mathcal{O}(1)$-Lipschitz, and input batch X_0 be non-degenerate. Then, there exists an absolute constant C such that for all $\ell \leq L$ it holds*

$$\mathbb{E}\left[\log \|\nabla_{W_\ell} \mathcal{L}(X_\ell)\|\right] \leq C d^5 (\psi(X_0)^3 + 1) \tag{6.10}$$

where the expectation is over the random orthogonal weight matrices.

Remark 6.6. For degenerate inputs $\psi(X_0) = \infty$ holds, in that the bound becomes vacuous.

Remark 6.7. The $\mathcal{O}(1)$-Lipschitz condition holds in many practical settings. For example, in a classification setting, MSE and cross entropy losses obey the $\mathcal{O}(1)$-Lipschitz condition (see Lemma C.11).

Note that the bound is stated for the expected value of log-norm of the gradients, which can be interpreted as bits of precision needed to store the gradient matrices. Thus, the fact that depth does not appear in any form in the upper bound of Theorem 6.5 points out that training arbitrarily

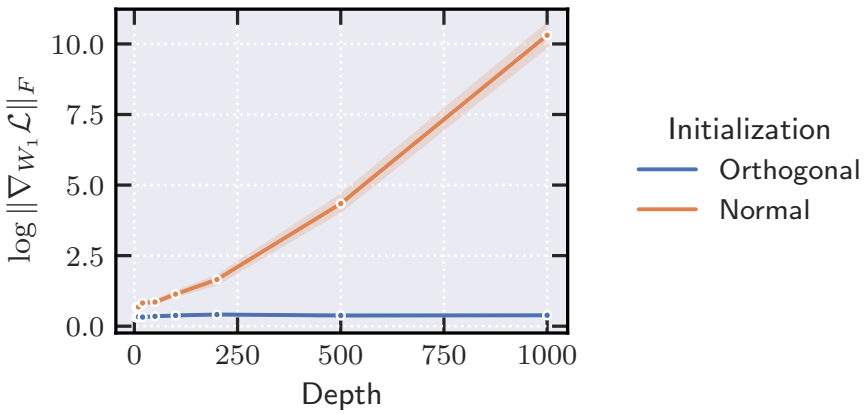


Figure 6.3: Logarithmic plot for the gradient norm of the first layer for networks with different number of layers evaluated on CIFAR10. For Gaussian weights (orange) the gradient-norm grows at an exponential rate, as predicted by Yang et al. [Yan+19, Theorem 3.9], while for orthogonal weights (blue) gradients remain bounded by a constant, validating Theorem 6.5. Traces are averaged over 10 runs and shaded regions denote the 95% confidence intervals.

deep MLPs with orthogonal weights will not face numerical issues that arise with Gaussian weights [Yan+19] as long as the inputs are non-degenerate. Such guarantees are necessary to ensure backpropagation will not face numerical issues.

Theorem 6.5 states that as long as the input samples are not linearly dependent, the gradients remain bounded for any arbitrary depth L. As discussed in the previous section and evidenced in Figure 6.2, this is necessary to avoid gradient explosion. Therefore, the upper bound provided in Theorem 6.5 is tight in terms of inputs constraints. Furthermore, as mentioned before, random batches sampled from commonly used benchmarks, such as CIFAR10, CIFAR100, MNIST, and FashionMNIST, are non-degenerate in most practical cases (see Section C.4 for more details). Thus, the assumptions and thereby assertions of the theorem are valid for all practical purposes.

To the best of our knowledge, Theorem 6.5 is the first non-asymptotic gradient analysis that holds for networks with batch normalization and finite width. Previous results heavily rely on mean field analyses in asymptotic regimes, where the network width tends to infinity [Yan+19]. While mean field analyses have brought many insights about the rate of gradient explosion, they are often specific to Gaussian weights. Here, we show that non-Gaussian weights can avoid gradient explosion, which has previously been considered "unavoidable" [Yan+19]. Figure 6.3 illustrates this pronounced discrepancy.

Proof idea of Theorem 6.5. The first important observation is that, due to the chain rule, we can

bound the log-norm of the gradient of a composition of functions, by bounding the summation of the log-norms of the input-output Jacobian of each layer, plus two additional terms corresponding to the loss term and the gradient of the first layer in the chain. If we discount the effect of the first and last terms, the bulk of the analysis is dedicated to bounding the total sum of log-norms of per layer input-output Jacobian, i.e., the fully connected and batch normalization layers. The second observation is that because the weights are only rotations, their Jacobian has eigenvalues equal to 1. Thus, the log-norm of gradients corresponding to fully connected layers vanish. What remains is to show that for any arbitrary depth ℓ, the log-norm of gradients of batch normalization layers also remains bounded. The main technical novelty for proving this step is showing that the log-norm of the gradient of BN layers is upper bounded by the isometry gap of pre-normalization matrices. Thus, we can invoke the exponential decay in isometry gap stated in Theorem 6.1 to establish a bound on the log-norm of the gradient of these layers. Finally, since the decay in isometry gap is exponentially fast, the bound on the total sum of log-norm of the gradients amounts to a geometric sum that remains bounded for any arbitrary depth ℓ.

6.3 Implications on training

In this section, we experimentally validate the benefits of avoiding gradient explosion and rank collapse for training. Thus far, we have proved that our constructed neural network with BN does not suffer from gradient explosion in Theorem 6.5, and does not have the rank collapse issue in depth via the orthogonalization property established in Theorem 6.1. We find that the constructed MLP is therefore less prone to numerical issues that arise when training deep networks.

By avoiding gradient explosion and rank collapse in depth, we observe that the optimization with vanilla minibatch Stochastic Gradient Descent (SGD) exhibits an almost depth-independent convergence rate for linear MLPs. In other words, the number of iterations to get to a certain accuracy does not vary widely between networks with different depths. Figure 6.4 (c) shows the convergence of SGD for CIFAR10, with learning rate 0.001, for MLPs with width $d = 100$ and batch size $n = 100$. While the SGD trajectory strongly diverges from the initial conditions that we analyze theoretically, Figure 6.4 shows that the gradients remain stable during training, as well as the fact that different depths exhibit largely similar accuracy curves.

While the empirical evidence for our MLP with linear activation is encouraging, non-linear activations are essential parts of feature learning [NH10; Kla+17; HG16; MHN13]. However, introducing non-linearity violates one of the key parts of our theory, in that it prevents represen-

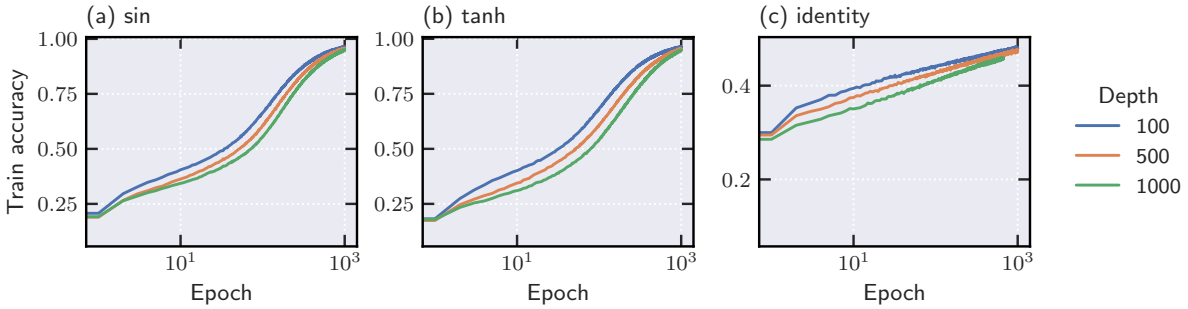


Figure 6.4: Contrasting the training accuracy of MLPs with BN and shaped sin, shaped tanh and identity activations, on the CIFAR10 dataset. The identity activation performs much worse than the nonlinearities, confirming that the sin and tanh networks are not operating in the linear regime. The networks are trained with vanilla SGD and the hyperparameters are width 100, batch size 100, learning rate 0.001.

tations from reaching perfect isometry (see Figure C.9 for details on the connection between non-linearities and gradient explosion in depth). Intuitively, this is due to the fact that non-linear layers, as opposed to rotations and batch normalization, perturb the isometry of representations and prevent them from reaching zero isometry gap in depth. This problem turns out to be not just a theoretical nuisance, but to play a direct role in the gradient explosion behavior. While the situation may seem futile at first, it turns out that activation shaping [LNR22; ZBM22; Mar+21; He+23; Noc+23] can alleviate this problem, which is discussed next. For the remainder of this section, we focus on the training of MLPs with non-linear activations, as well as standard batch normalization and fully connected layers.

6.4 Activation shaping based on the theoretical analysis

In recent years, several works have attempted to overcome the challenges of training very deep networks by parameterizing activation functions. In a seminal work, Martens et al. [Mar+21] propose *deep kernel shaping*, which is aimed at facilitating the training of deep networks without relying on skip connections or normalization layers, and was later extended to LeakyReLU in *tailored activation transformations* [ZBM22]. In a similar direction, Li, Nica, and Roy [LNR22] propose *activation shaping* in order to avoid a degenerate output covariance. While the mechanism proposed by Li, Nica, and Roy [LNR22] covers both smooth and non-smooth activations, we focus on their result for LeakyReLU, which consists of shaping the negative slope of the activation towards identity to ensure that the output covariance matrix remains

non-degenerate when the networks becomes very deep.

Since kernel and activation shaping aim to replace normalization, they have not been used in conjunction with normalization layers. In fact, in networks with batch normalization, even linear activations have non-degenerate outputs [DJB21; Yan+19] and exploding gradients [Yan+19]. Thus, shaping activations towards identity in the presence of normalization layers may seem fruitless. Remarkably, we empirically demonstrate that we can leverage activation shaping to avoid gradient explosion in depth by using a pre-activation gain at each layer.

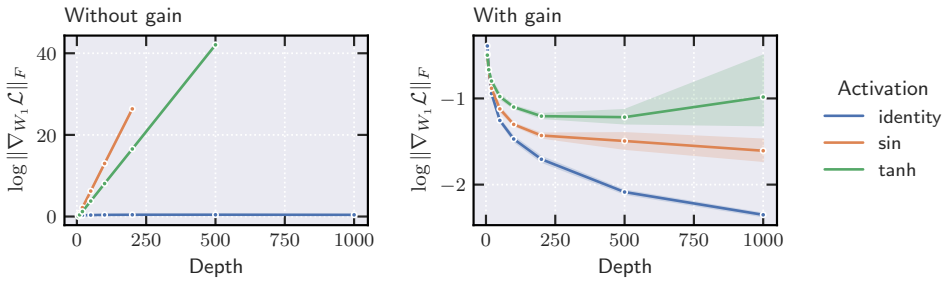


Figure 6.5: Logarithmic plot contrasting the effect of gain on the gradient at initialization of the first layer, for networks with different number of layers initialized with orthogonal weights, BN and different activations, evaluated on CIFAR10. The networks have hyperparameters width 100, batch size 100. Traces are averaged over 10 independent runs, with the shades showing the 95% confidence interval.

Inspired by our theory, we develop a novel activation shaping scheme for networks with BN. The main strategy consists of shaping the activation function towards a linear function across the layers. Our activation shaping consists of tuning the gain of the activation, i.e., tuning α for $\sigma(\alpha x)$. We consider non-linear activations $\sigma \in \{\tanh, \sin\}$.

The special property that both $\tanh$ and $\sin$ activations have in common is that they are centered, $\phi(0) = 0$, are differentiable around the origin $\phi'(0) = 1$, and have bounded gradients $\phi'(x) \leq 1, \forall x$. Therefore, by tuning the per-layer pre-activation gain α_ℓ towards 0, the non-linearities behave akin to the identity function. This observation inspires us to study the relationship between the rate of gradient explosion for each layer as a function of the gain parameter α_ℓ. Formally, we consider an MLP with shaped activations using gain α_ℓ for the ℓth layer, that has the update rule

$$X_{\ell+1} = \phi(\alpha_\ell \textsc{Bn}(W_\ell X_\ell)) \,. \tag{6.11}$$

Since the gradient norm has an exponential growth in depth, as shown in Figure 6.5, we can compute the slope of the linear growth rate of log-norm of gradients in depth. We define the

rate of explosion for a model of depth L and gain α_ℓ at layer ℓ as the slope of the log norm of the gradients $R(\ell, \alpha_\ell)$. We show in Figure 6.5 that by tuning the gain properly, we are able to reduce the exponential rate of the log-norm of the gradients by diminishing the slope of the rate curve and achieve networks trainable at arbitrary depths, while still maintaining the benefits of the non-linear activation. The main idea for our activation shaping strategy is to have a bounded total sum of rates across layers, by ensuring faster decay than a harmonic series (see App. C.5 for more details on activation shaping). Figure 6.5 illustrates that this activation shaping strategy effectively avoids gradient explosion while maintaining the signal propagation and orthogonality of the outputs in depth. Furthermore, Figure 6.4 shows that the training accuracy remains largely depth-independent. For further experiments using activation shaping, see Section C.7.

6.5 Discussion

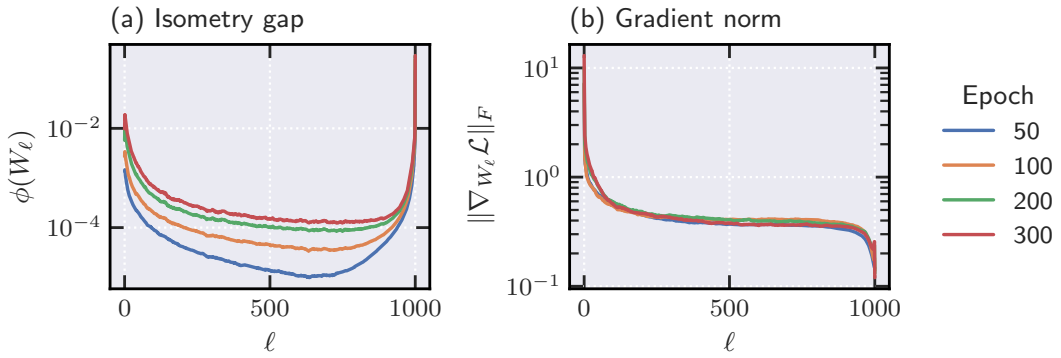


Figure 6.6: **Implicit orthogonality bias of SGD.** Training an MLP with width $d = 100$, batch size $n = 100$, and depth $L = 1000$, activation $\tanh$, using SGD with $\text{lr} = 0.001$ (a) Isometry gap (y-axis; log-scale) of weight matrices across all layers throughout training. (b) Gradient norms at each layer during training.

Implicit bias of SGD towards orthogonality in optimization. Optimization over orthogonal matrices has been an effective approach for training deep neural networks. Enforcing orthogonality during training ensures that the spectrum of the weight matrices remains bounded, which prevents gradient vanishing and explosion in depth. Vorontsov et al. [Vor+17] study how different orthogonality constraints affect training performance in RNNs. For example Lezcano-Casado and Martınez-Rubio [LCMR19] leverage the exponential map on the orthogonal group, Jose, Cissé, and Fleuret [JCF18] decompose RNN transition matrices in Kronecker factors and impose

soft constraints on each factor and Mhammedi et al. [Mha+17] introduce a constraint based on Householder matrices.

While these studies *enforce* orthogonality constraints, one of our most striking empirical observations is that when our MLP grows very deep, the middle layers remain almost orthogonal even after many steps of SGD. As shown in Figure 6.6, for 1000 layer networks, the middle layers remain orthogonal during training. One could hypothesize that this is due to small gradients in these layers. In Figure 6.6, we observe that the gradients of these middle layers are not negligible. Thus, in our MLP construction, both with linear activation and with activation shaping, the gradient dynamics have an *implicit bias* to optimize over the space of orthogonal matrices. The mechanisms underlying this implicit orthogonality bias will be an ample direction for future research.

7

Discussion and Conclusion

This thesis has made significant contributions to advancing the theoretical understanding of deep neural networks, with a particular focus on the behavior of MLPs at initialization. Despite neural networks' complex and non-linear nature, the research presented here offers several key insights into their mathematical properties. These contributions enhance the growing body of research to establish a solid theoretical foundation for deep neural networks.

Contributions. One of the primary contributions of this work is the demonstration that BN induces orthogonality in deeply hidden representations within mini-batches. This result was elaborated in Chapter 2. Additionally, this thesis has shown that when BN is applied, the mean field approximation provides a reliable predictor of behavior even in finite-width networks, as detailed in Chapter 3. Another significant finding is the layer and batch normalization tendency to bias activations toward a more isometric distribution, as discussed in Chapter 4. Furthermore, non-linear activations promote isometry in activations, explored in Chapter 5. Finally, the work demonstrates that the problem of gradient explosion, often encountered in networks utilizing BN, can be effectively mitigated through orthogonal weight initialization, as shown in Chapter 6.

From a mathematical perspective, this thesis provided two significant theoretical insights. First, we identified emergent behaviors in matrix products that arise from fully connected and normalization layers, including an intriguing property of normalization layers in Chapter 4. This property of normalization layers was later used in Chapter 6 to analyze gradients. Second, the application of Hermite polynomials in Chapter 5 proved instrumental in shedding light on the effects of

non-linear activations on signal propagation. This novel application of Hermite polynomials deepens our understanding of how non-linearity impacts signal propagation.

Limiations of this thesis. A core challenge throughout this work has been to develop a theory that is both mathematically rigorous and closely aligned with realistic neural network configurations. To achieve tractability, we made several simplifications, such as focusing on linear activations in Chapters 2 and 6, employing mixing-type assumptions in Chapters 2 and 3, assuming that batch size and network width are of the same order in Chapter 6, and using mean field approximations in Chapters 3 and 5. Finally, the primary limitation of this dissertation is its focus on the behavior of networks at initialization with randomly chosen weights. While this assumption was critical to the tractability of the analysis, it naturally restricts the direct applicability of the results to trained networks. Extending theoretical insights to trained networks will require developing new analytical tools and methods.

Despite these necessary simplifications, the theoretical results presented here align well with observed behavior in more realistic settings where such assumptions are relaxed. This demonstrates that theoretical insights can retain practical relevance, even when derived under simplifying conditions. At the same time, the research points to certain limitations, which will be discussed in the following section.

Future avenues for research Looking ahead, several avenues of research offer promising opportunities for further exploration.

Analyzing training dynamics in a deep network with non-linear dynamics remains theoretically challenging. However, analyzing gradients at initialization can move us one step closer to that goal. For example, analyzing the backward gradients, particularly in terms of non-linear activations, can be highly rewarding and impactful. In particular, if we combine insights from forward and backward passes of neural networks, we can arrive at a quantitative characterization of the neural target kernel for a particular choice of activation function.

One of the most important future directions involves extending the findings of this thesis to other network architectures, such as transformers and recurrent neural networks. For example, given that feedforward layers are one of the main components of transformer architecture, the insights developed in this thesis for feedforward settings can be insightful for studying how layer normalization and non-linear activation functions behave in transformers. As another example, the problem of learning long-range dependencies in recurrent settings can be viewed

as a vanishing gradient problem. As we argued in several chapters, batch normalization is an effective module for dealing with vanishing gradients in a feedforward setup. While these insights are not directly applicable to recurrent settings, finding analogs of batch normalization in a recurrent setting is worth exploring. Alternatively, insights on the importance of initialization in Chapter 6 and Chapter 2 can provide new perspectives on leveraging initialization to control gradients in recurrent architectures.

As mentioned earlier in limitations, the mean field assumption that the width of the network tends to infinity was a necessary simplification in some of our theoretical analyses. While this assumption has considerable benefits, it also raises questions as to the accuracy of the theoretical predictions. While we saw in Chapter 3 that in some cases, these predictions are accurate, developing non-asymptotic results remains a significant theoretical challenge. A middle-ground approach that has shown promise and higher accuracy is tending the width and depth of the network to infinity while keeping their ratio constant.

Concluding remarks. In conclusion, this dissertation advances our mathematical understanding of deep neural networks, particularly in signal propagation, normalization techniques, and initialization strategies. By applying a fresh mathematical perspective, this work has laid the groundwork for future research to bridge the gap between theoretical understanding and practical application in deep learning. The need for robust theoretical foundations becomes increasingly critical as neural networks grow in complexity and capability. This work contributes to that foundation, offering both immediate insights and promising avenues for future exploration in theory of neural networks.

A

Detailed proofs of Chapter 2

A.1 Preliminaries

Let $v, w \in \mathbb{R}^k$ then $v \odot w \in \mathbb{R}^n$ with coordinates

$$[v \odot w]_i = v_i w_i \tag{A.1}$$

Furthermore $v^{\otimes 2} \in \mathbb{R}^{k \times k}$ with entities

$$[v^{\otimes 2}]_{ij} = v_i v_j. \tag{A.2}$$

In Table A.1, we summarize notations introduced previously.

The Markov chain of hidden representations

Recall the chain of the hidden representations, denoted by $\{H_\ell \in \mathbb{R}^{d \times n}\}$, obeying the following recurrence:

$$H_{\ell+1} = \frac{1}{\sqrt{d}} BN(W_\ell H_\ell), \quad BN(M) = \Big(\mathrm{diag}(MM^\top)\Big)^{-1/2} M, \tag{A.3}$$

Notation	Type	Definition
ℓ	integer	number of layers
n	integer	batch size
d	integer	width of the network
k	integer	output dimension
X	$\mathbb{R}^{d\times n}$	input matrix
H_ℓ	$\mathbb{R}^{d\times n}$	hidden representations at ℓ (obeying Eq. equation 2.4)
BN	$\mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$	batch normalization layer (defined in Eq.equation 2.4)
Law(X)		the law of random matrix X
$\sigma_i(M)$	$\mathbb{R}^{k_1\times k_2} \to \mathbb{R}_+$	the ith largest singular value of matrix M
I_k	$\mathbb{R}^{k\times k}$	Identity matrix of size k
$\mathbf{1}_k$	$\mathbb{R}^n$	all-ones vector

Table A.1: Summary of notations used in this chapter.

where $W_\ell \in \mathbb{R}^{d\times n}$ are random weight matrices with i.i.d. zero-mean Gaussian elements. It is easy to check that the Frobenius norm of H_ℓ is one due to the row-wise normalization:

$$\begin{aligned}\operatorname{tr}\left(BN(H)BN(H)^\top\right) &= \operatorname{tr}\left(\operatorname{diag}(HH^\top)^{-1/2}HH^\top\operatorname{diag}(HH^\top)^{-1/2}\right)\\ &= \operatorname{tr}\left(\operatorname{diag}(HH^\top)^{-1}HH^\top\right)\\ &= d.\end{aligned} \tag{A.4}$$

Lyapunov function characterizing the orthogonality

We introduce the following function serving as a Lyapunov function $\widehat{V} : \mathbb{R}^{d\times n}$ characterizing the orthogonality of the hidden representations:

$$\widehat{V}(H) = \frac{1}{n} - (\sigma_n(H))^2, \tag{A.5}$$

where $\sigma_n(H)$ is the minimum singular value of matrix H. Next lemma proves $\widehat{V}(H_\ell)$ bounds the orthogonality gap.

Lemma A.1. *For all hidden representations H_ℓ, the following holds:*

$$V(H_\ell) \le 2n\widehat{V}(H_\ell).$$

Proof. Let $\sigma_1, \ldots, \sigma_n$ be the singular values of H_ℓ. Given these singular values, one can compute $V(H_\ell)$ as

$$(V(H_\ell))^2 = \sum_{i=1}^{n} \left(\sigma_i^2 - \frac{1}{n}\right)^2.$$

According to Eq. equation A.4, $\sum_{i=1}^{2} \sigma_i^2 = 1$ holds. The proof is an immediate consequence of this propery.

$$\begin{aligned} V^2(H_\ell) &= \sum_{i=1}^{n} \sigma_i^4 - 2 \underbrace{\left(\sum_{i=1}^{n} \sigma_i^2\right)}_{=1} \frac{1}{n} + \frac{1}{n} \\ &= \sum_{i=1}^{n} \sigma_i^4 - \frac{1}{n}. \end{aligned}$$

For a freeze σ_n, the maximum of $\sum_{i=1}^{n} \sigma_i^4 - \frac{1}{n}$ subject to $\sum_{i=1}^{n} \sigma_i^2 = 1$ is met when $\sigma_1^2 = 1 - (n-1)\sigma_n^2$ and $\sigma_2 = \cdots = \sigma_n$. Using this optimal values, we get the following bound:

$$V^2(H_\ell) \leq 2(n-1)^2 \underbrace{\left(\frac{1}{n} - \sigma_n^2\right)^2}_{=\widehat{V}^2(H_\ell(X))}.$$

Taking the square root of both sides concludes the proof. □

A.2 Proof of Theorem 2.1

The proof of Theorem 2.1 relies on the following Theorem that characterizes the change of $\widehat{V}$ in consecutive layers.

Theorem A.2. *The sequence* $\{H_\ell\}$ *obeys*

$$\mathbb{E}\left[\widehat{V}(H_{\ell+1})|H_\ell\right] \leq \left(1 - \frac{2}{3}\left(\frac{1}{n} - \widehat{V}(H_\ell)\right)\right)\widehat{V}(H_\ell) + \frac{1}{\sqrt{d}}.$$

Notably, the above result does not rely on Assumption $\mathcal{A}_1$. Assuming that Assumption $\mathcal{A}_1$ holds, we complete the proof of Theorem 2.1. Combining the last Theorem by this assumption, we get

$$\mathbb{E}\left[\widehat{V}(H_{\ell+1})\right] \leq \left(1 - \frac{2}{3}\alpha\right)\mathbb{E}\left[\widehat{V}(H_\ell)\right] + \frac{1}{\sqrt{d}} \tag{A.6}$$

Induction over ℓ yields

$$\begin{aligned}\mathbb{E}\left[\widehat{V}(H_{\ell+1}(X))\right] &\leq \left(1-\frac{2}{3}\alpha\right)^{\ell}\mathbb{E}\left[\widehat{V}(H_1)\right] + \left(\sum_{k=1}^{\ell}(1-\frac{2}{3}\alpha)^k\right)\frac{1}{\sqrt{d}} \\ &\leq \left(1-\frac{2}{3}\alpha\right)^{\ell}\mathbb{E}\left[\widehat{V}(H_1)\right] + \frac{3}{2\alpha\sqrt{d}}\end{aligned}$$

An application of Lemma A.1 completes the proof:

$$\begin{aligned}\mathbb{E}\left[V(H_{\ell+1})\right] &\leq 2n\mathbb{E}\left[\widehat{V}(H_{\ell+1})\right] \\ &\leq 2\left(1-\frac{2}{3}\alpha\right)^{\ell} + \frac{3n}{2\alpha\sqrt{d}}\end{aligned}$$

Stability analysis without technical assumption

Using Theorem A.2, we can prove stability of the chain $\{H_\ell\}$ without Assumption $\mathcal{A}_1$. After rearrangement of terms in Theorem A.2, we get

$$\mathbb{E}\left[\widehat{V}(H_{\ell+1})|H_\ell\right] - \widehat{V}(H_\ell) \leq -\frac{2}{3}\widehat{V}(H_\ell)\left(\frac{1}{n} - \widehat{V}(H_\ell)\right) + \frac{1}{\sqrt{d}}$$

Taking the expectation over H_ℓ and average over ℓ yields

$$\mathbb{E}\left[\frac{1}{\ell}\sum_{k=1}^{\ell}\widehat{V}(H_k)\left(\frac{1}{n} - \widehat{V}(H_k)\right)\right] \leq \left(\frac{3\mathbb{E}\left[\widehat{V}(H_0)\right]}{2\ell}\right) + \frac{3}{2\sqrt{d}}$$

A.3 Proof of Theorem A.2

Spectral decomposition.

Consider the SVD decomposition of H_ℓ as $H_\ell = U\mathrm{diag}(\sigma)V^\top$ where U and V are orthogonal matrices. Given this decomposition, we get

$$W_\ell H_\ell = \underbrace{W_\ell U}_{W}\mathrm{diag}(\sigma)V^\top \tag{A.7}$$

Since W_ℓ is Gaussian and U is an orthogonal matrix, entities of W are also i.i.d. standard normal. We will repeatedly use the above decomposition in our analyses.

Concentration analysis.

Consider matrix $C_{\ell+1} := H_{\ell+1}^\top H_{\ell+1}$ whose eigenvalues are $\sigma_1^2, \dots, \sigma_n^2$. The SVD decomposition of H_ℓ in Eq. equation A.7 allows us to write $C_{\ell+1}$ as

$$C_{\ell+1} = \frac{1}{d}\sum_{i=1}^{d}\left(\frac{w_i \odot \sigma}{\|w_i \odot \sigma\|_2}\right)^{\otimes 2}$$

where $w_i \in \mathbb{R}^n$ is the i-th row of W, and $\sigma \in \mathbb{R}^n$ is the vector of singular values of H_ℓ. Thus, conditioned on σ, $C_{\ell+1}$ is an empirical average of i.i.d. random vectors. This allows us to prove that this empirical average is concentrated around its expected value. The next lemma states this concentration.

Lemma A.3. *The following concentration always holds*

$$\mathbb{E}_{W_\ell}\|C_{\ell+1} - \mathbb{E}_{W_\ell}\left[C_{\ell+1}\right]\|^2 \leq 1/d$$

where

$$\mathbb{E}_{W_\ell}\left[C_{\ell+1}\right] = \operatorname{diag}(p_1(\sigma), \dots, p_n(\sigma)), \quad p_i(\sigma) := \mathbb{E}\left[\frac{\sigma_n^2 w_n^2}{\sum_{k=1}^n \sigma_k^2 w_k^2}\right], \quad w_i \overset{i.i.d.}{\sim} \mathbb{N}(0,1) \tag{A.8}$$

The concentration of $C_{\ell+1}$ allows us to prove that the Lyapunov function $\widehat{V}(H_{\ell+1})$ is concentrated around $1/n - p_n(\sigma)$.

Lemma A.4. *The following holds*

$$\mathbb{E}_{W_\ell}\left[\left(\widehat{V}(H_{\ell+1}) - (1/n - p_n(\sigma))\right)^2\right] \leq \frac{1}{d}$$

.

The last lemma allows us to predict the value of random variable $\widehat{V}(H_{\ell+1})$ by deterministic term $1/n - p_n(\sigma)$.

Contraction.

The decay in $\widehat{V}(H_{\ell+1})$ with ℓ is due to term $1/n - p_n(\sigma)$ in the last lemma. This term is less than (or equal to) $V(H_\ell)$.

Lemma A.5. *For $p_n(\sigma)$ defined in Eq. equation A.8, the following holds:*

$$\left(\frac{1}{n} - p_n(\sigma)\right) \leq \left(1 - \frac{2}{3}\left(\frac{1}{n} - \widehat{V}(H_\ell)\right)\right)\widehat{V}(H_\ell).$$

Combining the last lemma by Lemma A.4 concludes the proof of Theorem A.2:

$$\mathbb{E}_{W_\ell}\left[\widehat{V}(H_{\ell+1})\right] \leq \left(1 - \frac{2}{3}\left(\frac{1}{n} - \widehat{V}(H_\ell)\right)\right)\widehat{V}(H_\ell) + \frac{1}{\sqrt{d}}. \tag{A.9}$$

To complete the proof, we prove Lemmas A.3, A.4, and A.5.

Proof of Lemma A.3

Given the spectral decomposition of H_ℓ in Eq. equation A.7, we compute element ij of $C_{\ell+1}$, which is denoted by $[C_{\ell+1}]_{ij}$:

$$[C_{\ell+1}]_{ij} = [A^\top A]_{ij} = \frac{1}{d}\sum_{k=1}^{d} A_{ki}A_{kj}, \quad A_{ki} = W_{ki}\sigma_i/\sqrt{v_k}$$

where $v_k = \sum_{m=1}^{n} W_{km}^2\sigma_m^2$. Since W_{km} are zero mean and unit variace, we get

$$\begin{aligned}
\mathbb{E}[C_{\ell+1}]_{ij} &= 0 \\
\mathbb{E}[C_{\ell+1}]_{ij}^2 &= \frac{1}{d^2}\sum_{k=1}^{d} A_{ki}^2A_{kj}^2 + \frac{1}{d^2}\sum_{k,k'}^{d}\underbrace{\mathbb{E}\left[A_{ki}^2A_{kj}^2A_{k'i}^2A_{k'j}^2\right]}_{=0} \\
&= \frac{1}{d^2}\sum_{k=1}^{d} A_{ki}^2A_{kj}^2
\end{aligned}$$

holds for $i \neq j$. By summing up over $i \neq j$, we get

$$\sum_{i\neq j}\mathbb{E}[C_{\ell+1}]_{ij} = \frac{1}{d^2}\left(\sum_k\left(\sum_i A_{ki}^2\right)^2 - \sum_{ik} A_{ki}^4\right)$$

For the diagonal elements, we get

$$\begin{aligned}
\mathbb{E}\left[C_{\ell+1}\right]_{ii} &= \frac{1}{d}\sum_{k=1}^{d}\mathbb{E}A_{ki}^2 \\
&= \frac{1}{d}\sum_{k=1}^{d}\mathbb{E}A_{ki}^2 \\
&= \underbrace{\frac{1}{d}\sum_{k=1}^{d}\mathbb{E}\left[\frac{W_{ki}^2\sigma_i^2}{\sum_{j=1}^{n}W_{kj}^2\sigma_j^2}\right]}_{p_i(\sigma)}
\end{aligned}$$

The variance of $[C_{\ell+1}]_{ii}$ is bounded as

$$\begin{aligned}\text{var}([C_{\ell+1}]_{ii}) &= \mathbb{E}\left(\frac{1}{d}\sum_{k=1}^{d}(A_{ki}^2 - p_i(\sigma))\right)^2 \\ &= \frac{1}{d^2}\sum_{k=1}^{d}(A_{ki}^2 - p_i(\sigma))^2 \\ &\leq \frac{1}{d^2}\sum_{k=1}^{d} A_{ki}^4\end{aligned}$$

Combining results for the diagonal and off-diagonal elements yields

$$\begin{aligned}\mathbb{E}\|C_{\ell+1} - \mathbb{E}_{W_\ell}[C_{\ell+1}]\|_F^2 &= \sum_{ij}\text{var}([C_{\ell+1}]_{ij}) \\ &\leq \frac{1}{d^2}\left(\sum_i A_{ki}^2\right)^2 = 1/d\end{aligned}$$

Proof of Lemma A.4

Notably, the eigenvalues of $C_{\ell+1}$ are squared singular values of $H_{\ell+1}$. Let $\lambda_n(C)$ denote the nth largest eigenvalue of matrix C.

$$\begin{aligned}\mathbb{E}_{W_\ell}\left[\left(\widehat{V}(C_{\ell+1}) - \left(\frac{1}{n} - p_n(\sigma)\right)\right)^2\right] &\leq \mathbb{E}_{W_\ell}\left[\left(\lambda_n(C_{\ell+1}) - \lambda_n\left(\mathbb{E}_{W_\ell}[C_{\ell+1}]\right)\right)^2\right] \\ &= \mathbb{E}_{W_\ell}\left[\|C_{\ell+1} - \mathbb{E}_{W_\ell}[C_{\ell+1}]\|_F^2\right] \\ &\leq \frac{1}{d}\end{aligned}$$

where the last inequality relies on Lemma A.3.

Proof of Lemma A.5

The proof is based an application of moment generating function that allows us to compute expectations of ratios of random variables.

Lemma A.6 (Lemma 1 in [Saw72]). *Let X_1 be a random variable that is positive with probability one and X_2 be an arbitrary random variable. Suppose that there exists a joint moment generating function of X_1 and X_2:*

$$\phi(\theta_1, \theta_2) = \mathbb{E}\left[\exp(\theta_1 X_1 + \theta_2 X_2)\right]$$

for $\theta_1 \leq \epsilon$ and $|\theta_2| < \epsilon$ where ϵ is some positive constant. Then

$$\mathbb{E}\left[\frac{X_2}{X_1}\right] = \int_{-\infty}^{0} \left[\frac{\partial \phi(\theta_1, \theta_2)}{\partial \theta_2}\right]_{\theta_2=0} d\theta_1$$

To estimate $p_n(\sigma)$, we set $X_2 := \sigma_i^2 w_i^2$ and $X_1 = \sum_j \sigma_j^2 w_j^2$, which obtains

$$\begin{aligned}
&\phi(\theta_1, \theta_2) \\
&\quad = \mathbb{E}\left[\exp(\theta_1 X_1 + \theta_2 X_2)\right] \\
&\quad = (2\pi)^{-n/2} \int_{-\infty}^{\infty} \exp((\theta_1+\theta_2)\sigma_i^2 w_i^2 + \sum_{j\neq i} \theta_1 \sigma_j^2 w_j^2) \exp(-\sum_k w_k^2/2) dw \\
&\quad = (2\pi)^{-n/2} \int_{-\infty}^{\infty} \exp(\left(-0.5 + (\theta_1+\theta_2)\sigma_i^2\right) w_i^2) dw_i \left(\prod_{j\neq i} \int_{-\infty}^{\infty} \exp(\left(-0.5 + \theta_1 \sigma_j^2\right) w_j^2) dw_j\right) \\
&\quad = \frac{1}{\sqrt{1-2(\theta_1+\theta_2)\sigma_i^2}} \left(\prod_{j\neq i} \frac{1}{\sqrt{1-2\theta_1\sigma_j^2}}\right).
\end{aligned}$$

Taking derivative with respect to θ_2 yields

$$\frac{\partial \phi}{\theta_2}(\theta_1, 0) = \frac{\sigma_i^2}{(1-2\theta_1\sigma_i^2)^{3/2}} \left(\prod_{j\neq i} \frac{1}{\sqrt{1-2\theta_1\sigma_j^2}}\right)$$

Using the result of the last lemma, we get

$$p_i(\sigma) = \int_{-\infty}^{0} \frac{\sigma_i^2}{(1-2\theta\sigma_i^2)} \left(\prod_{j} \frac{1}{\sqrt{1-2\theta\sigma_j^2}}\right) d\theta$$

Therefore,

$$p_n(\sigma) = \sigma_n^2 f_n(\sigma), \quad f_n(\sigma) := \int_{-\infty}^{0} \frac{d\theta}{(1-2\theta\sigma_n^2)^{3/2} \prod_{j\neq n}(1-2\theta\sigma_j^2)^{1/2}} \tag{A.10}$$

Since $\sum_{i=1}^n \sigma_i^2 = 1$ holds (see Eq. equation A.4), $f_n(\sigma)$ in the above formulation is minimized when the σ_j^2s are all equal for all $j \neq n$. Let $\sigma_n^2 := 1/n - \delta$ and $\sigma_j^2 := 1/n + \delta/(n-1)$ for all $j \neq i$. This allows us to establish a lowerbound on $f_n(\sigma)$ as

$$f_n(\sigma) \geq \int_0^{\infty} \underbrace{\left(1 + 2\theta\left(\frac{1}{n} - \delta\right)\right)^{-\frac{3}{2}} \left(1 + 2\theta\left(\frac{1}{n} + \frac{\delta}{n-1}\right)\right)^{-\frac{n-1}{2}}}_{g(\delta):=} d\theta \tag{A.11}$$

Next lemma proves that $g(\delta)$ is a convex function for $\delta \in [0, 1/n]$.

Lemma A.7. *The function $g(\delta)$, which is defined in Eq. equation A.11, is a convex function on domain $\delta \in [0, 1/n]$.*

The convexity of $g(\delta)$ yields

$$g''(\delta) \geq 0 \; \forall \delta \implies g(\delta) \geq g(0) + \delta g'(0)$$

The above bound allow us to bound $f_n(\sigma)$ as

$$\begin{aligned} f_n(\sigma) &\geq \int_0^\infty g(0) d\theta + \delta \int_0^\infty 2\theta \left(1 + \frac{2\theta}{n}\right)^{-\frac{n}{2}-2} d\theta \\ &= 1 + \frac{2\delta n}{n+2} \end{aligned}$$

Note that we use integration by parts to compute the above integrals. Recall $\delta = \frac{1}{n} - \sigma_n^2$. Combining the above inequality by Eq. equation A.10 concludes the proof of the Lemma A.5:

$$\frac{1}{n} - p_n(\sigma) \leq \left(1 - \frac{2n}{(n+2)}\sigma_n^2\right)\left(\frac{1}{n} - \sigma_n^2\right)$$

Proof of Lemma A.7

We can show convexity of $g(\delta)$ by showing $g''(\delta) \geq 0$ for all $\delta \in [0, 1/n]$. To this end, we define function $h_1(\delta)$ (for the compactness of notations) as

$$h_1(\delta) := \left(1 + 2\theta\left(\frac{1}{n} - \delta\right)\right)^{-5/2} \left(1 + 2\theta\left(\frac{1}{n} + \frac{\delta}{n-1}\right)\right)^{-\frac{n-1}{2}-1}.$$

Given h_1, the derivative of g reads as

$$\begin{aligned} g'(\delta) &= h_1(\delta)\left(-\frac{3}{2}(-2\theta)\left(1 + 2\theta\left(\frac{1}{n} + \frac{\delta}{n-1}\right)\right) - \frac{n-1}{2}\left(\frac{2\theta}{n-1}\right)\left(1 + 2\theta\left(\frac{1}{n} - \delta\right)\right)\right) \\ &= \theta h_1(\delta)\left(3 + \theta\frac{6}{n} + \theta\delta\frac{6}{n-1} - 1 - \frac{2\theta}{n} + 2\theta\delta\right) \\ &= \theta h_1(\delta)\underbrace{\left(2 + \theta\frac{4}{n} + \theta\delta\frac{2n+4}{n-1}\right)}_{h_2(\delta) :=} \\ &= \theta h_1(\delta) h_2(\delta). \end{aligned}$$

One can readily check that $h_2'(\delta) \geq 0$. Hence, $h_1'(\delta) \geq 0$ ensures the convexity of $g(\delta)$. Consider the following auxiliary function

$$h_3(\delta) := \left(1 + 2\theta\left(\frac{1}{n} - \delta\right)\right)^{-7/2} \left(1 + 2\theta\left(\frac{1}{n} + \frac{\delta}{n-1}\right)\right)^{-\frac{n-1}{2}-2}.$$

Given h_3, we compute h_1' as

$$\begin{aligned}
h_1'(\delta) &= h_3(\delta)\left(\left(-\frac{5}{2}\right)(-2\theta)\left(1 + 2\theta\left(\frac{1}{n} + \frac{\delta}{n-1}\right)\right) + \left(-\frac{n-3}{2}\right)\frac{2\theta}{n-1}\left(1 + 2\theta\left(\frac{1}{n} - \delta\right)\right)\right) \\
&= h_3(\delta)\theta\left(5(1 + 2\theta/n + \frac{2\theta\delta}{n-1}) - \frac{n-3}{n-1}(1 + 2\theta/n - 2\theta\delta)\right) \\
&= h_3(\delta)\theta\underbrace{\left(\frac{4n-2}{n-1} + 2\theta\frac{4n-2}{n(n-1)} + 2\theta\delta\frac{n+2}{n-1}\right)}_{h_4(\delta):=} \\
&= \theta h_3(\delta)h_4(\delta).
\end{aligned}$$

Clearly we have $h_3(\delta), h_4(\delta) \geq 0$ for $\delta \in [0, \frac{1}{n}]$. Therefore, the proof is complete.

To compare networks with and without BN, next lemma formally establishes the contraction of orthogonality gap to a large value for networks withou BN.

Lemma A.8. *Let $S_\ell = W_\ell \dots W_1$. Then, there exists a positive constant δ such that the following holds*

$$\lim_{\ell\to\infty} \frac{1}{2\ell} \log\left(\left|V(S_\ell) - \sqrt{\frac{n-1}{n}}\right|\right) \leq -\delta. \tag{A.12}$$

In other words, the gap $V(S_\ell)$ converges to $\sqrt{(n-1)/n}$ with asymptotic rate $\exp(-\delta\ell)$ for the identity inputs. While, Thm. 2.1 proves the gap for BN networks converges to $n/\sqrt{d}$ with an exponential rate. For a sufficiently large d, $n/\sqrt{d} \ll \sqrt{(n-1)/n}$ holds.

Proof. Let $\sigma_1(\ell) \geq \sigma_2(\ell) \geq \cdots \geq \sigma_n(\ell)$ denote singular values of matrix S_ℓ, then it is known that

$$\lim_{\ell\to\infty} \frac{1}{\ell} \log(\sigma_i^2(\ell)) = \frac{1}{2}\left(\log(2) + \Psi\left(\frac{d-i+1}{2}\right)\right) \tag{A.13}$$

holds where Ψ is digamma function [New86], which a monotonically decreasing function. Therefore,

$$\lim_{\ell\to\infty}\frac{1}{\ell}\Big(\log(\sigma_2^2(\ell))-\log(\sigma_1^2(\ell))\Big)=-\delta<0 \tag{A.14}$$

holds for $\delta>0$ that can be exactly computed using function Ψ. The above inequality implies that $\sigma_1^2(\ell)$ increases (or decreases) faster than $\sigma_2^2(\ell)$ with an exponential rate. Using this result, we get the following limit for $j\neq 1$:

$$\lim_{\ell\to\infty}\frac{1}{\ell}\log\left(\frac{\sigma_j^2(\ell)}{\sum_i\sigma_i^2(\ell)}\right)\leq\lim_{\ell\to\infty}\frac{1}{\ell}\log\left(\frac{\sigma_2^2(\ell)}{\sigma_1^2(\ell)}\right)=-\delta \tag{A.15}$$

Furthermore,

$$\lim_{\ell\to\infty}\frac{1}{2\ell}\log\left|\frac{\sigma_1^2(\ell)}{\sum_i\sigma_i^2(\ell)}-1\right|\leq\lim_{\ell\to\infty}\frac{1}{\ell}\log\left|\frac{n\sigma_2^2(\ell)}{\sigma_1^2(\ell)}\right|\leq-\delta+\lim_{\ell\to\infty}\log(n)/\ell\leq-\delta \tag{A.16}$$

Let $\sigma(\ell)=(\sigma_1^2(\ell),\ldots,\sigma_n^2(\ell))\in\mathbb{R}^n$ and 1_n is the all one vector in $\mathbb{R}^n$ and $e_1=(1,0,\ldots,0)\in\mathbb{R}^n$. Using triangular inequality, we get

$$\left|V(S_\ell)-\sqrt{(n-1)/n}\right|=\left|\left\|\frac{\sigma(\ell)}{\|\sigma(\ell)\|_1}-\frac{1}{n}1_n\right\|-\left\|e_1-\frac{1}{n}1_n\right\|\right|\leq\left\|\frac{\sigma(\ell)}{\|\sigma(\ell)\|_1}-e_1\right\| \tag{A.17}$$

Therefore, we get

$$\begin{aligned}\lim_{\ell\to\infty}\frac{1}{\ell}\log&\left(\left|V(S_\ell)-\sqrt{\frac{n-1}{n}}\right|\right)\\&\leq\lim_{\ell\to\infty}\left(\frac{1}{4\ell}\log\left(2n\left(\frac{\sigma_2^2(\ell)}{\sum_j\sigma_j^2(\ell)}\right)\right)+\frac{1}{4\ell}\log\left(2\left(\frac{\sigma_1^2(\ell)}{\sum_j\sigma_j^2(\ell)}-1\right)^2\right)\right)\end{aligned} \tag{A.18}$$

which is bounded by $-\delta$ according to the established bounds in Eq. equation A.15, and Eq. equation A.16. □

Proof of Lemma 2.2

The main idea is based on a particular coupling of random matrices $W_\ell H_\ell$ and G. Consider the truncated SVD decomposition of H_ℓ as $H_\ell=U\mathrm{diag}(\sigma)V^\top$ where $U\in\mathbb{R}^{d\times n}$ and $V\in\mathbb{R}^{n\times n}$ are

orthogonal matrices. Due to the orthogonality, the law of $W_\ell U$ is the same as those of GV. By coupling $W_\ell U = GV$, we get

$$\begin{aligned}
\left(\mathcal{W}_2(W_\ell H_\ell, G/\sqrt{n})\right)^2 &= \inf_{\text{all the couplings}} \mathbb{E}\|W_\ell U \mathrm{diag}(\sigma) V^\top - GVV^\top/\sqrt{n}\|_F^2 \\
&\leq \mathbb{E}\|GV\left(\mathrm{diag}(\sigma) - I/\sqrt{n}\right)V^\top\|_F^2 \\
&= \mathbb{E}\mathrm{tr}(GV\left(\mathrm{diag}(\sigma) - I/\sqrt{n}\right)V^\top V\left(\mathrm{diag}(\sigma) - I/\sqrt{n}\right)V^\top G^\top) \\
&= \mathrm{tr}(V\left(\mathrm{diag}(\sigma) - I/\sqrt{n}\right)\left(\mathrm{diag}(\sigma) - I/\sqrt{n}\right)V^\top \mathbb{E}\left[G^\top G\right]) \\
&= \mathbb{E}\mathrm{tr}(\left(\mathrm{diag}(\sigma) - I/\sqrt{n}\right)\left(\mathrm{diag}(\sigma) - I/\sqrt{n}\right)V^\top V) \\
&= \mathbb{E}\|\left(\mathrm{diag}(\sigma) - I/\sqrt{n}\right)\|_F^2 \\
&= \mathbb{E}\left[\sum_{i=1}^n (\sigma_i - 1/\sqrt{n})^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n \left(\sigma_i^2 - 1/n\right)^2 / \left(\sigma_i + 1/\sqrt{n}\right)^2\right] \\
&\leq n\mathbb{E}\left[\sum_{i=1}^n \left(\sigma_i^2 - 1/n\right)^2\right] \\
&\leq n\mathbb{E}\left[V^2(H_\ell)\right] \\
&\leq 2n\mathbb{E}\left[V(H_\ell)\right].
\end{aligned}$$

A.4 Orthogonality gap for the iterative initialization

Recall the proposed initialization for weights based on SVD decomposition $H_\ell = U_\ell \Sigma_\ell V_\ell^\top$.

$$W_\ell = \frac{1}{\|\Sigma_\ell^{1/2}\|_F} V_\ell' \Sigma_\ell^{-1/2} U_\ell^\top.$$

Here, we show that

$$V(H_\ell) > V(W_\ell H_\ell) \tag{A.19}$$

holds as long as $V(H_\ell) \neq 0$ and $\mathcal{A}_1(\alpha, \ell)$ holds. Given singular values H_ℓ, we get

$$\begin{aligned}
V(H_\ell) &= \sum_{i=1}^n \left(\sigma_i^2 - \frac{1}{n}\right)^2 \\
&= \sum_i \sigma_i^4 - 1/n,
\end{aligned} \tag{A.20}$$

where we used $\sum_{i=1}^{n} \sigma_i^2 = 1$ (see Eq. equation 2.1). Now, we compute $V(H_\ell)$ using the singular values.

$$W_\ell H_\ell = \frac{1}{\|\Sigma_\ell^{1/2}\|_F} V_\ell' \Sigma_\ell^{1/2} V_\ell$$

Hence, the following holds

$$\begin{aligned} V(W_\ell H_\ell) &= \sum_{i=1}^{n} \left(\frac{\sigma_i}{\sum_j \sigma_j} - \frac{1}{n} \right)^2 \\ &= \frac{1}{(\sum_{i=1}^{n} \sigma_i)^2} - 1/n \end{aligned}$$

Combining with Eq. equation A.20, we get

$$V(H_\ell) - V(W_\ell H_\ell) = \sum_{i=1}^{n} \sigma_i^4 - \frac{1}{(\sum_{i=1}^{n} \sigma_i)^2}.$$

To show that the right side of the above equation is positive, we need to prove

$$\sum_i \sigma_i^4 \left(\sum_{i=1}^{n} \sigma_i \right)^2 > 1 = \left(\sum_i \sigma_i^2 \right)^4$$

holds. Using Cauchy-Schwarz inequality, we get

$$\begin{aligned} \left(\sum_i \sigma_i^2 \right)^4 &= \left(\sum_i \sigma_i^{3/2} \sigma_i^{1/2} \right)^4 \\ &\leq \left(\sum_i \sigma_i \right)^2 \left(\sum_i \sigma_i^2 \sigma_i \right)^2 \\ &\leq \left(\sum_i \sigma_i \right)^2 \left(\sum_i \sigma_i^4 \right), \end{aligned}$$

where the equality in the above inequality is met only for $\sigma_i^2 = \frac{1}{n}$ (so $V(H_\ell) = 0$) under $\mathcal{A}_1$.

A.5 Comparisons with a BN-replacement

Here, we compare iterative orthogonalization with two baselines: (i) initialization with random orthogonal weights [SMG13a], and (ii) adaptive gradient clipping [Bro+21].

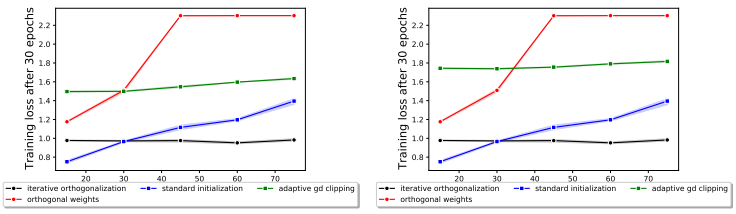
(i). Orthogonal weights achieve orthogonal representations in deep *linear* networks. Consider a MLP whose weight are initialized by Xavier's scheme. Let the weight matrix at layer ℓ admit the SVD decomposition $W_\ell = U_\ell \mathrm{diag}(\sigma_\ell) V_\ell$. Then, we replace the weight matrix by the orthogonal matrix $(\mathrm{mean}(\sigma_\ell)) U_\ell V_\ell^\top$. Since ReLU networks with orthogonal weights are prone to the alignment of representations in deep layers, we observed that such an initialization does not help with the slow down of training with depth (see Fig. A.1a).

(ii). Recently, [Bro+21] propose an effective replacement for BN —- based on gradient clipping. Let G_ℓ denotes the gradient of training loss with respect to W_ℓ. Given a clipping parameter λ, adaptive gradient clipping adjusts the norm of G_ℓ as

$$\widehat{G}_\ell = \begin{cases} \left(\lambda \frac{\|W_\ell\|_F^*}{\|G_\ell\|_F}\right) G_\ell & \frac{\|G_\ell\|_F}{\|W_\ell\|_F^*} \geq \lambda \\ G_\ell & \text{otherwise} \end{cases} \tag{A.21}$$

where $\|W_\ell\|_F^* = \max\{\|W_\ell\|_F, 10^{-2}\}$.

Fig. A.1a, and A.1b presents results for two different choice of clipping parameters $\lambda = 0.1$ and $\lambda = 1$, respectively. These plots demonstrate adaptive gradient clipping effectively alleviates the training slow down with depth. Yet, we observe iterative orthogonalization achieves a better training loss after 30 epochs. It is not known how the clipping enhance the training. While, iterative orthogonalization is inspired by orthogonalization of representations with BN layers, which is theoretically established in this chapter.



(a) $\lambda = 0.1$ (b) $\lambda = 1$

Figure A.1: Iterative orthogonalization vs adaptive gradient clipping. Horizontal axis: the network depth. Vertical axis: training loss for each network after 30 SGD epochs. For more details, see Sec. 2.4. Mean and 95% confidence interval of 4 independent runs.

B

Proof of Chapter 3

B.1 A concentration bound for the empirical covariance matrix

The following analysis pertains to the deviation between the sample covariance matrix, normalized by the true covariance, and the identity matrix. For a collection of d independent identically distributed i.i.d. samples in $\mathbb{R}^d$, represented as $x_1, x_2, ..., x_d \in \mathbb{R}^n$, the sample covariance matrix C_d is given by:

$$C_d = \frac{1}{d}\sum_{i=1}^{d} x_i x_i^T. \tag{B.1}$$

The true covariance matrix C is defined as the expected outer product of the samples, or:

$$C = \mathbb{E}[x_i x_i^T]. \tag{B.2}$$

We are interested in bounding the deviation of C_d from the covariance matrix C in terms of their Frobenius norm (denoted as $\|.\|_F$), as outlined in the lemma below. Note that if activation ϕ is uniformly bounded $\phi(x)^2 \leq Bx^2$, and BN is the batch-norm operator, then $\|\phi(\text{BN}(x))^2\| \leq B\|\text{BN}(x)\|^2 \leq Bn$. Thus, activations applied after normalization layers obey

the condition of Lemma B.1. With this point in mind, we will prove the concentration result for vectors that are uniformly bounded by the same quantity.

Lemma B.1. *Let $x_1, \dots, x_d \in \mathbb{R}^n$ be i.i.d. random vectors with covariance $\mathbb{E}x_i x_i^\top = C$ and sample covariance $C_d := \frac{1}{d}\sum_{i=1}^d x_i x_i^\top$. If the vector norms are universally bounded such that $\|x_i\|^2 \leq nB$ holds almost surely, then for $t \lesssim \sqrt{d}$, the following is true:*

$$P\left(\|C_d - C\|_F \gtrsim t\varepsilon\right) \leq \exp(-t^2), \qquad \varepsilon := \frac{Bn}{\sqrt{d}}. \tag{B.3}$$

Here, the probability is taken over the random vectors $x_1, \dots, x_d$.

We will use the last lemma to prove Theorem 3.2.

Lemma B.2. *Under the same conditions as Lemma B.1, if the covariance matrix C is not degenerate, i.e., it does not possess zero eigenvalues, for $t \lesssim \sqrt{d}$ it holds*

$$P\left(\|C^{-1}C_d - I_n\|_F \gtrsim t\varepsilon\right) \leq \exp(-t^2), \qquad \varepsilon := \frac{B\|C^{-1}\|n}{\sqrt{d}}. \tag{B.4}$$

Proof of Lemma B.1. Recall that Bernstein's inequality provides an upper bound on the probability that the sum exceeds a certain threshold t. Given i.i.d. variables $X_1, \dots, X_d$, it states that are uniformly bounded $|X_i| \leq B$ for all i, we have

$$\mathbb{P}\left(\frac{1}{d}\sum_{i=1}^d X_i \geq t\right) \leq 2\exp\left(-\frac{dt^2/2}{K^2 + Kt/3}\right), \tag{B.5}$$

where $t > 0$ and σ^2 is the variance of $\sum_{i=1}^d E[X_i^2] \leq dB^2$. Define $X_i := \|x_i x_i^\top - C\|_F^2$. We have

$$\begin{aligned}
\|x_i x_i^\top - C\|_F^2 &\leq \|x_i x_i\|_F + \|C\|_F && \text{(B.6)}\\
&\leq Bn + \|Ex_i x_i^\top\|_F && \text{(B.7)}\\
&\leq Bn + E\|x_i x_i\|_F && \text{(B.8)}\\
&\leq 2Bn. && \text{(B.9)}
\end{aligned}$$

Thus, we can plug $K := 2Bn$ into the Bernstein's inequality to get

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{d}\sum_{i=1}^d \|x_i x_i^\top - C\|_F \geq t\right) \leq && \text{(B.10)}\\
\exp\left(-\frac{dt^2/2}{4n^2B^2 + 2Bnt/3}\right). && \text{(B.11)}
\end{aligned}$$

Since $\|\cdot\|_F$ is convex, Jensen's inequality which implies that moving the averaging inside can only decrease its value, which in turn implies

$$\mathbb{P}\left(\|\frac{1}{d}\sum_{i=1}^{d} x_i x_i^\top - C\|_F \geq t\right) \tag{B.12}$$

$$\leq \exp\left(-\frac{dt^2}{8n^2B^2(1+\frac{t}{6nB})}\right). \tag{B.13}$$

We can now rename $t\sqrt{d}/(\sqrt{8}Bn)$ as t and use definition of sample covariance to conclude

$$\mathbb{P}\left(\|C_d - C\|_F \geq t\frac{\sqrt{8}Bn}{\sqrt{d}}\right) \leq \exp\left(-\frac{t^2}{(1+\frac{t}{3\sqrt{2d}})}\right), \tag{B.14}$$

which can be restated as

$$\mathbb{P}\left(\|C_d - C\|_F \gtrsim t\varepsilon\right) \leq \exp(-t^2), \qquad \varepsilon := \frac{Bn}{\sqrt{d}}, \tag{B.15}$$

which holds if $t \lesssim \sqrt{d}$. □

Proof of Lemma B.2. Consider transformed vectors $z_i := C^{-1/2}x_i$. Note that we have $\mathbb{E}z_i z_i^\top = C^{-1}C = I_n$. Thus, we can apply Lemma B.1 on deviations of sample covariance of z_i's from I_n. Furthermore, we have $\|z_i\|^2 \leq \|C^{-1}\|\|x_i\|^2 \leq \|C^{-1}\|Bn$. So we can invoke Lemma B.1 by setting B to $\|C^{-1}\|B$. □

B.2 Analyzing Gram Dynamics Around Fixed Points

Equipped with the results established so far, we now turn our attention to the dynamics of Gram matrices in relation to the total variation of the Multi-Layer Perceptron (MLP) Markov chain. In particular, we demonstrate a specific construction based on fixed-point G_*, and show that after one layer update the total variation distance is bounded.

Lemma B.3 (Restated Lemma 3.3)**.** *Let $\hat{h} \in \mathbb{R}^{d\times n}$ be constructed by drawing its rows i.i.d. from $\mathcal{N}(0, G_*)$. Let $\hat{\mu}$ denote the distribution of $\hat{h}$. Given that fixed-point Gram matrix G_* is non-degenerate and the activation is uniformly bounded $\phi(x)^2 \leq Bx^2$, then*

$$\|\hat{\mu} - T(\hat{\mu})\|_{tv} \lesssim \varepsilon^2 \textsc{Ln}(1/\varepsilon), \qquad \varepsilon := \frac{n\|G_*^{-1}\|B}{\sqrt{d}}, \tag{B.16}$$

holds if B and $\|G_*^{-1}\|$ are non-zero.

Under the assumption of geometric contraction, irrespective of the initial distribution, the total variation distance to the stationary distribution contracts by $1-\alpha$, for some $\alpha > 0$, after one transition T. This result, together with Lemma 3.3, provides a tool to approximate the stationary distribution by a matrix constructed from the fixed-point Gram matrix G_*.

Lemma B.4. *Let $\hat{\mu}$ denote the distribution of a random matrix in $\mathbb{R}^{d\times n}$, whose rows are drawn i.i.d. from $\mathcal{N}(0, G_*)$. Assuming the rapid mixing condition 3.1 holds with constant $\alpha > 0$, then*

$$\|\hat{\mu} - \mu_*\|_{tv} \lesssim \alpha^{-1}\varepsilon^2 \textsc{Ln}(1/\varepsilon), \qquad \varepsilon := \frac{nB\|G_*^{-1}\|}{\sqrt{d}}. \tag{B.17}$$

We can finally the results about total variation into the context of Gram matrix dynamics through depth.

Lemma B.5. *Let μ_ℓ denote the hidden representation of a BN-MLP, and $\hat{\mu}$ denote the distribution of a matrix whose rows are drawn from $\mathcal{N}(0, G_*)$. If hidden representations obey the rapidly mixing assumption with rate $1-\alpha$, for $\alpha > 0$, then*

$$\|\mu_\ell - \hat{\mu}\|_{tv} \lesssim (1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \textsc{Ln}(1/\varepsilon), \qquad \varepsilon := \frac{nB\|G_*^{-1}\|}{\sqrt{d}}. \tag{B.18}$$

With the necessary lemmas in place, we are now ready to present our main theorem.

Theorem B.6 (Restated Theorem 3.2)**.** *For an MLP chain G_ℓ that originates from a non-degenerate input G_0, and that has a non-degenerate fixed point $G_,$ and that obeys the rapidly mixing assumption with $\alpha > 0$, we have the following:*

$$\begin{aligned}\mathbb{P}(\|G_* - G\|_F \geq t) \lesssim \\ t^{-2}(\|G_*\|^2(1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \textsc{Ln}(1/\varepsilon)),\end{aligned} \tag{B.19}$$

with $\varepsilon := nB\kappa(G_*)/\sqrt{d}$.

The proof of the theorem relies on the following lemma bounds Gram matrix deviations by total variation.

Lemma B.7. *Conditioned on Gram matrices $G_*, G \in \mathbb{R}^{n\times n}$, construct $h, \hat{h} \in \mathbb{R}^{d\times n}$ by drawing their rows i.i.d. from $\mathcal{N}(0, G)$ and $\mathcal{N}(0, G_*)$. If G_* is non-degenerate, the following hols for total variation between h and $\hat{h}$:*

$$tv(h, \hat{h}) \geq \frac{t}{100}\mathbb{P}(\|G_*^{-1}G - I_n\|_F^2 \geq t), \tag{B.20}$$

where the probability is defined over G.

The proof of this theorem follows directly from the lemmas we have established.

Proof of Theorem B.6. We apply the total variation bound established in Lemma B.5 and combine this with the lower bound stated in Lemma B.7

$$\frac{t}{100}\mathbb{P}(\|G_*^{-1}G - I_n\|_F^2 \geq t) \tag{B.21}$$

$$\leq tv(h_\ell, \hat{h}) \lesssim (1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \textsc{Ln}(1/\varepsilon). \tag{B.22}$$

Omitting constants we have

$$\mathbb{P}(\|G_*^{-1}G - I_n\|_F \geq \sqrt{t}) \tag{B.23}$$

$$\lesssim t^{-1}((1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \textsc{Ln}(1/\varepsilon)). \tag{B.24}$$

By a change of variables we get

$$\mathbb{P}\left(\|G_*^{-1}G - I_n\|_F \geq t\right) \tag{B.25}$$

$$\lesssim t^{-2}((1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \textsc{Ln}(1/\varepsilon)). \tag{B.26}$$

Note that $\|G^* - G\|_F = \|G^*(G_*^{-1}G - I_n)\|_F$, which is bounded by $\|G_*\|\|G_*^{-1}G - I_n\|_F$. Thus

$$\mathbb{P}\left(\|G_* - G\|_F \geq t\right) \tag{B.27}$$

$$\leq \mathbb{P}\left(\|G_*^{-1}G - I_n\|_F \geq t/\|G_*\|\right) \tag{B.28}$$

$$\leq t^{-2}\|G_*\|^2((1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \textsc{Ln}(1/\varepsilon)), \tag{B.29}$$

where the last equation is due to equation B.22. The above inequality obtains

$$\varepsilon := \frac{Bn}{\sqrt{d}}\|G_*\|\|G_*^{-1}\| \tag{B.30}$$

$$\mathbb{P}(\|G_* - G\|_F \geq t) \lesssim \tag{B.31}$$

$$t^{-2}\left(\|G_*\|^2(1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \textsc{Ln}(1/\varepsilon))\right). \tag{B.32}$$

Recall that $\|G_*\|\|G_*^{-1}\|$ encodes the ratio of largest to smallest eigenvalue of G_*, which is its condition number $\kappa(G_*)$. This finalizes the proof. □

Proof of Lemma B.5. First, note that geometric contraction assumption implies

$$\|\mu_\ell - \mu_*\|_{tv} \leq (1-\alpha)\|\mu_{\ell-1} - \mu_*\|_{tv} \leq (1-\alpha)^\ell, \tag{B.33}$$

which by numerical inequality $1 - x \leq \exp(-x)$ can be bounded by $\exp(-\alpha\ell)$. We can invoke Lemma B.4 and triangle inequality for total variation to conclude the proof. □

Proof of Lemma B.4. Recall that the rapidly mixing assumption implies that $\|T(\hat{\mu}) - \mu_*\|_{tv} \leq (1-\alpha)\|\hat{\mu} - \mu_*\|_{tv}$. Furthermore, invoking Lemma B.3, we have

$$\|T(\hat{\mu}) - \hat{\mu}\|_{tv} \lesssim \varepsilon^2 \textsc{Ln}(1/\varepsilon), \tag{B.34}$$

$$\implies \|\hat{\mu} - \mu_*\|_{tv} \lesssim \alpha^{-1}\varepsilon^2 \textsc{Ln}(1/\varepsilon), \tag{B.35}$$

where the last line is implied by the triangle inequality for total variation. □

Proof of Lemma B.3. Let us explicitly construct $T(\hat{\mu})$. Recall that $\hat{\mu}$ describes distribution of $\hat{h}$ whose rows are drawn i.i.d. from $\mathcal{N}(0, G_*)$. Define $h := W\phi(\textsc{Bn}(\hat{h}))$, where W is a Gaussian with i.i.d. elements $\mathcal{N}(0, 1/d)$. Thus, by construction, distribution of h follows $T(\hat{\mu})$. Our main proof strategy of upper bounding total variation between $\hat{h}$ and h is to bound it conditioned on the proximity of G to G_*.

Bounding deviations $\|G_*^{-1}G - I\|_F$. Recall that based on the fixed-point property of G_* we have

$$\mathbb{E}_{w\sim N(0,G_*)}\phi(\textsc{Bn}(w))^{\otimes 2} = G_*. \tag{B.36}$$

Define sampled Gram of activations $G := \frac{1}{d}\phi(\textsc{Bn}(\hat{h}))^\top\phi(\textsc{Bn}(\hat{h}))$, which is equal in expectation to $\mathbb{E}G = G_*$. By construction of batch norm operator which maps every row to $\sqrt{n}$-sphere, and the uniform bound $\phi(x)^2 \leq Bx^2$, we can conclude that rows of $\phi(\textsc{Bn}(\hat{h}))$ are always bounded by

$$\|\textsc{Bn}(x)\| \leq \sqrt{n} \qquad \forall x \in \mathbb{R}^n \tag{B.37}$$

$$\implies \|\phi(\textsc{Bn}(x))\| \leq \sqrt{Bn}, \qquad \forall x \in \mathbb{R}^n \tag{B.38}$$

$$\implies \|\mathrm{row}_k(\phi(\textsc{Bn}(\hat{h})))\|^2 \leq Bn, \qquad \forall k. \tag{B.39}$$

This allows us to invoke Lemma B.2 to conclude:

$$\mathbb{P}(\|G_*^{-1}G - I_n\|_F \geq t\varepsilon) \leq \exp(-t^2), \tag{B.40}$$

where $\varepsilon = B\|G_*^{-1}\|n/\sqrt{d}$.

Bounding total variation $tv(h, \hat{h})$. Define set of matrices $N_t := \{M \in \mathbb{R}^{n\times n} : \|{G_*}^{-1}M - I_n\|_F^2 \leq \varepsilon^2 t^2\}$. Observe that conditioned on G, h is equal in distribution to a matrix rows are drawn i.i.d. from $\mathcal{N}(0, G)$. Thus, we can decompose the total variation based on depending on G belonging to neighborhood of G_* or not

$$tv(h, \hat{h}) \leq \mathbb{P}\left\{\|{G_*}^{-1}G - I_n\|_F^2 \geq t^2\varepsilon^2\right\} \tag{B.41}$$

$$+ \sup_{G\in N_t} tv(\mathcal{N}(0, G), \mathcal{N}(0, G_*)) \tag{B.42}$$

$$\lesssim \mathbb{P}\left\{\|{G_*}^{-1}G - I_n\|_F \geq t\varepsilon\right\} + \frac{3}{2}t^2\varepsilon^2, \tag{B.43}$$

where in the last line we use the upper bound on total variation between two Gaussian matrices from [DMR18]. Plugging our result for deviation of G and G_* we have

$$tv(h, \hat{h}) \leq \frac{3}{2} t^2 \varepsilon^2 + \exp(-t^2), \tag{B.44}$$

which holds for all $t \lesssim \sqrt{d}$. In particular, we can set $t^2 := \textsc{Ln}(2/3\varepsilon^2)$ which implies

$$tv(h, \hat{h}) \lesssim \frac{3\varepsilon^2}{2}(1 + \textsc{Ln}(2/2\varepsilon^2)), \tag{B.45}$$

which omitting constants can be restated as

$$tv(h, \hat{h}) \lesssim \varepsilon^2 \textsc{Ln}(1/\varepsilon^2). \tag{B.46}$$

To finish the proof, observe that condition $t \lesssim \sqrt{d}$ translates to $\textsc{Ln}(1/\varepsilon^2) = O(d)$ which in turn requires $\varepsilon \gtrsim \exp(-d/2)$. Plugging the definition of ε we have $nB\|G_*^{-1}\| \gtrsim \sqrt{d}\exp(-d/2)$. Since the right-hand side is $o(1)$, and $n \geq 1$, this condition will always hold if the boundedness and conditioning are non-zero $B, \|G_*^{-1}\| > 0$. □

Proof of Lemma B.7. Define set of matrices $N_t := \{M \in \mathbb{R}^{n \times n} : \|{G_*}^{-1} M - I_n\|_F^2 \leq t\}$. We have

$$\begin{aligned}
tv(h, \hat{h}) &= \int_G \mathbb{P}(G) tv(\mathcal{N}(0, G), \mathcal{N}(0, G_*)) && \text{(B.47)}\\
&= \int_{G \in N_t} \mathbb{P}(G) tv(\mathcal{N}(0, G), \mathcal{N}(0, G_*)) && \text{(B.48)}\\
&\quad + \int_{G \notin N_t} \mathbb{P}(G) tv(\mathcal{N}(0, G), \mathcal{N}(0, G_*)) && \text{(B.49)}\\
&\geq \int_{G \notin N_t} \mathbb{P}(G) tv(\mathcal{N}(0, G), \mathcal{N}(0, G_*)) && \text{(B.50)}\\
&\geq \mathbb{P}(G \notin N_t) \inf_{G \notin N_t} tv(\mathcal{N}(0, G), \mathcal{N}(0, G_*)) && \text{(B.51)}\\
&\geq \frac{t}{100} \mathbb{P}\left(\|G_*^{-1} G - I_n\|_F^2 \geq t\right), && \text{(B.52)}
\end{aligned}$$

where in the last line we have used the lower bound for total variation of multivariate Gaussians from [DMR18]. □

C

Supplemental proofs and experiments for Chapter 6

C.1 Conditional orthogonalization

Isometry after rotation. Our analysis is based on [JDB23b, Corollary 3], which we restate in the following Lemma:

Lemma C.1. *For all non-degenerate matrices $X \in \mathbb{R}^{d\times d}$, we have:*

$$\mathcal{I}(\textsc{Bn}(X)) \geq \left(1 + \frac{\mathit{variance}\{\|X_{j\cdot}\|\}_{j=1}^{d}}{(\mathit{mean}\{\|X_{j\cdot}\|\}_{j=1}^{d})^2}\right)\mathcal{I}(X)\,. \tag{C.1}$$

Lemma C.1 proves isometry bias of BN. The next lemma proves that isometry does not change under rotation

Lemma C.2 (Isometry after rotation)**.** *Let $X \in \mathbb{R}^{d\times d}$ and $W \sim \mathbb{O}_d$ be a random orthogonal matrix and $X' = WX$. Then,*

$$\mathcal{I}(\textsc{Bn}(X')) \geq \left(1 + \frac{\mathit{variance}\{\|X'_{j\cdot}\|\}_{j=1}^{d}}{(\mathit{mean}\{\|X'_{j\cdot}\|\}_{j=1}^{d})^2}\right)\mathcal{I}(X)\,. \tag{C.2}$$

Proof. Using properties of the determinant, we have

$$\det(X'X'^{\top}) = \det(W)^2\det(XX^{\top}) = \det(XX^{\top}), \tag{C.3}$$

where the last equation holds since W is an orthogonal matrix. Furthermore,

$$\begin{aligned} \operatorname{tr}(X'X'^{\top}) &= \operatorname{tr}(WXX^{\top}W^{\top}) && \text{(C.4)} \\ &= \operatorname{tr}(XX^{\top}\underbrace{W^{\top}W}_{=I}) && \text{(C.5)} \\ &= \operatorname{tr}(XX^{\top})\,. && \text{(C.6)} \end{aligned}$$

Combining the last two equations with Lemma C.1 concludes the proof. □

Increasing isometry with rotations and BN. The last lemma proves the isometry bias does not decrease with rotation and BN. However, this does not prove a strict decrease in isometry with BN and rotation. The next lemma proves there exists an orthogonal matrix for which the isometry is strictly increasing.

Corollary C.3 (Increasing isometry)**.** *Let $X \in \mathbb{R}^{d\times d}$ and denote its singular value decomposition $X = U\operatorname{diag}\left(\{\sigma_i\}_{i=1}^{d}\right)V^{\top}$, where U and V are orthogonal matrices. Then, we have:*

$$\mathcal{I}(\textsc{Bn}(U^{\top}H)) = 1\,. \tag{C.7}$$

Proof. Let $S = \operatorname{diag}\left(\{\sigma_i\}_{i=1}^{d}\right)$ be the diagonal matrix containing the singular values of X. Then, we have:

$$\begin{aligned} \textsc{Bn}(U^{\top}X) &= \textsc{Bn}(U^{\top}USV^{\top}) && \text{(C.8)} \\ &= \textsc{Bn}(SV^{\top}) && \text{(C.9)} \\ &= \operatorname{diag}(SV^{\top}VS)^{-\frac{1}{2}}SV^{\top} && \text{(C.10)} \\ &= S^{-1}SV^{\top} && \text{(C.11)} \\ &= V^{\top}, && \text{(C.12)} \end{aligned}$$

which has maximum isometry 1 since it's an orthonormal matrix. □

Thus, there exists a rotation that increases isometry with BN for each non-orthogonal matrix. The proof of the last corollary is based on a straightforward application of Lemma 6.2.

Orthogonalization with randomness. The isometric is non-decreasing in Lemma 6.2 and provably increases for a certain rotation matrix (as stated in the last corollary). Hence, it is possible to increase isometry with random orthogonal matrices.

Theorem C.4. *Suppose $W \sim \mathbb{O}_d$ is a matrix drawn from $\mathbb{O}_d$ such that $W \overset{d}{=} WU$ for all orthogonal matrices U. Let $\{\lambda_i\}_{i=1}^d$ be the eigenvalues of $XX^\top$. Then,*

$$\mathbb{E}_W\left[\mathcal{I}(\textsc{Bn}(WX))|\, X\right] \geq \left(\frac{1}{1-\frac{\sum_{k=1}^d(\lambda_k-1)^2}{2d^2(d+2)}}\right)\mathcal{I}(X) \tag{C.13}$$

holds for all $X = \textsc{Bn}(\cdot)$, with equality for orthogonal matrices.

Remark C.5. Note that the assumption on $X = \textsc{Bn}(\cdot)$, can be viewed as an induction hypothesis, in that we can recursively apply this theorem to arrive at a quantitative rate at depth.

Notably, $\sum_{i=1}^d \lambda_i = d$ if $X = \textsc{Bn}(\cdot)$. Hence, one can expect that $\sum_{i=1}^d \lambda_i^2 < d^2$ for all full random matrices X in form of $X = \textsc{Bn}(\cdot)$.

Proof. We need to compute the variance/mean ratio in Lemma 6.2. Let $X \in \mathbb{R}^{d\times d}$ have SVD decomposition $X = U\mathrm{diag}\{\sigma_i\}_{i=1}^d V^\top$ where U and V are orthogonal matrix and $\sigma_i^2 = \lambda_i$. Since the distribution of W is invariant to transformations with orthogonal matrices, the distribution of W equates those of $X' = W\mathrm{diag}\{\sigma_i\}V^\top$. It easy to check that

$$\|X'_{j\cdot}\| = \sqrt{\sum_{i=1}^d \sigma_i^2 W_{ji}^2} = \sqrt{\sum_{i=1}^d \lambda_i W_{ji}^2}\,. \tag{C.14}$$

Thus,

$$\sum_{j=1}^d \|X'_{j\cdot}\|^2 = \sum_{i=1}^d \lambda_i = d, \tag{C.15}$$

where the last equality holds due to the batch normalization. Thus, we have

$$\mathbb{E}\left[\frac{d\sum_{j=1}^d \|X'_{j\cdot}\|^2}{(\sum_{j=1}^d \|X'_{j\cdot}\|)^2}\right] = \mathbb{E}\left[\frac{d^2}{(\sum_j \|X'_{j\cdot}\|)^2}\right] \geq \frac{d^2}{\mathbb{E}\left[\left(\sum_{j=1}^d \|X'_{j\cdot}\|\right)^2\right]}\,. \tag{C.16}$$

We need to estimate

$$\mathbb{E}\left[\|X'_i\|\|X'_j\|\right] = \mathbb{E}\left[\left(\sum_{k=1}^d \lambda_k W_{ik}^2\right)^{\frac{1}{2}}\left(\sum_{k=1}^d \lambda_k W_{jk}^2\right)^{\frac{1}{2}}\right]\,. \tag{C.17}$$

Since square root function is concave, we have $\sqrt{x} \leq 1 + \frac{1}{2}(x-1)$. Thus

$$\mathbb{E}\left[\|X_i'\|\|X_j'\|\right] \leq \mathbb{E}\left[\left(1 + 0.5\sum_{k=1}^{d}(\lambda_k - 1)W_{ik}^2\right)\left(1 + 0.5\sum_{k=1}^{d}(\lambda_k - 1)W_{jk}^2\right)\right] \tag{C.18}$$

$$= 1 + \frac{1}{4}\sum_{k,q}(\lambda_k - 1)(\lambda_q - 1)\mathbb{E}\left[W_{ik}^2 W_{jq}^2\right] \tag{C.19}$$

$$= 1 + \frac{1}{4}\left[\sum_{k\neq q}(\lambda_k - 1)(\lambda_q - 1)\underbrace{\mathbb{E}\left[W_{ik}^2 W_{jq}^2\right]}_{E_1} + \sum_{k=q}(\lambda_k - 1)(\lambda_k - 1)\underbrace{\mathbb{E}\left[W_{ik}^2 W_{jk}^2\right]}_{E_2}\right], \tag{C.20}$$

where in the first equality we have used the fact that the cross terms reduce, where the expectations are applications of Weingarten calculus:

$$\mathbb{E}\left[0.5\sum_{k=1}^{d}(\lambda_k - 1)W_{ik}^2 + 0.5\sum_{k=1}^{d}(\lambda_k - 1)W_{jk}^2\right] = \mathbb{E}\left[0.5\sum_{k=1}^{d}\lambda_k W_{ik}^2 + 0.5\sum_{k=1}^{d}\lambda_k W_{jk}^2 - 1\right] \tag{C.21}$$

$$= 0.5\sum_{k=1}^{d}\lambda_k\mathbb{E}\left[W_{ik}^2\right] + 0.5\sum_{k=1}^{d}\lambda_k\mathbb{E}\left[W_{jk}^2\right] - 1 \tag{C.22}$$

$$= \frac{0.5}{d}\sum_{k=1}^{d}\lambda_k + \frac{0.5}{d}\sum_{k=1}^{d}\lambda_k - 1 \tag{C.23}$$

$$= 0\,. \tag{C.24}$$

The main quantity we must compute is an expectation of polynomials taken over the Haar measure of the Orthogonal group $O(n)$. To carry out the computation, we make use of Weingarten calculus [BCS11; CŚ06; Wei78]. More specifically, we make use of the of the Weingarten formula, studied by [CŚ06; CMN22]:

$$\int_{O(n)} r_{i_1 j_1} r_{i_2 j_2} \ldots r_{i_{2d} j_{2d}} d\mu(O(n)) = \sum_{\sigma\in\mathcal{M}_{2d}}\sum_{\sigma\in\mathcal{M}_{2n}} \Delta_\sigma(\boldsymbol{i})\Delta_\sigma(\boldsymbol{j})\mathrm{Wg}^O(\sigma^{-1}\tau), \tag{C.25}$$

where $\mu(O(n))$ is the Haar measure of Orthogonal group. For an in depth explanation of each quantity in the Weingarten formula, we refer the reader to Collins, Matsumoto, and Novak [CMN22, Section 5.2].

The quantity we focus on is $\mathbb{E}_W\left[W_{ik}W_{ik}W_{jq}W_{jq}\right]$. We will do the computation on multiple cases, based on the equalities of k, q. Notice that $i \neq j$ in all cases. It suffices if we focus on the two distinct cases: $E_1 = \mathbb{E}_W\left[W_{ik}^2 W_{jq}^2\right]$ $(k \neq q)$ and $E_2 = \mathbb{E}_W\left[W_{ik}^2 W_{jk}^2\right]$ $(k = q)$.

We first compute E_1.

Following the procedure from Collins, Matsumoto, and Novak [CMN22, Section 5.2], we take the index sequences to be $\boldsymbol{i} = (i, i, j, j)$ and $\boldsymbol{j} = (k, k, q, q)$. Similarly, we get $\Delta_\sigma(\boldsymbol{i}) = \Delta_\tau(\boldsymbol{j}) = 1$ only if $\sigma = \{\{1, 2\}, \{3, 4\}\}$ and $\tau = \{\{1, 2\}, \{3, 4\}\}$.

Consdering σ, τ as permutations, we get:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}, \tag{C.26}$$

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}, \tag{C.27}$$

$$\sigma^{-1}\tau = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}, \tag{C.28}$$

where $\sigma^{-1}\tau$ has coset-type $[1, 1]$. Finally, we plug the results back into the formula and we obtain:

$$E_1 = \mathbb{E}_W \left[W_{ik}^2 W_{jq}^2\right] = \mathrm{Wg}^O([1, 1]) = \frac{d+1}{d(d+2)(d-1)},$$

where the last equality is based on the results in Section 7 of [CM09].

We compute E_2.

Similar to the previous expression, we take the index sequences to be to be $\boldsymbol{i} = (i, i, j, j)$ and $\boldsymbol{j} = (k, k, k, k)$. Thus, we obtain $\Delta_\sigma(\boldsymbol{i}) = \Delta_\tau(\boldsymbol{j}) = 1$ only if $\sigma = \{\{1, 2\}, \{3, 4\}\}$ and $\tau_1 = \{\{1, 2\}, \{3, 4\}\}$, $\tau_2 = \{\{1, 3\}, \{2, 4\}\}$, $\tau_1 = \{\{1, 4\}, \{2, 3\}\}$. Similarly, we compute $\sigma^{-1}\tau_i$ for $i \in \{1, 2, 3\}$. Notice that σ is the identity permutation, thus yielding $\sigma^{-1}\tau_i = \tau_i$, with the coset-types $[1, 1], [2], [2]$ respectively, for each $i \in \{1, 2, 3\}$.

Plugging back into the original equation, we obtain:

$$E_2 = \mathbb{E}_W \left[W_{ik}^2 W_{jk}^2\right] = \mathrm{Wg}^O([1, 1]) + 2\mathrm{Wg}^O([2]) \tag{C.29}$$

$$= \frac{d+1}{d(d+2)(d-1)} + 2\frac{-1}{d(d+2)(d-1)} \tag{C.30}$$

$$= \frac{d-1}{d(d+2)(d-1)}\,. \tag{C.31}$$

Thus, plugging back into the original inequality, we obtain:

$$\mathbb{E}\|X_i'\|\|X_j'\| \tag{C.32}$$

$$\leq 1 + \frac{1}{4}\left[\sum_{k\neq q}(\lambda_k - 1)(\lambda_q - 1)\frac{d+1}{d(d+2)(d-1)} + \sum_{k=q}(\lambda_k - 1)(\lambda_k - 1)\frac{d-1}{d(d+2)(d-1)}\right] \tag{C.33}$$

$$= 1 + \frac{1}{4d(d+2)(d-1)}\left[\underbrace{\sum_{k\neq q}(\lambda_k - 1)(\lambda_q - 1)}_{S_{\neq}} - \underbrace{\sum_{k=q}(\lambda_k - 1)(\lambda_k - 1)}_{S_{=}}\right] \tag{C.34}$$

$$= 1 - \frac{\sum_k(\lambda_k - 1)^2}{2d(d+2)(d-1)}, \tag{C.35}$$

where we have used that $S_{\neq} + S_{=} = \sum_{k,q}(\lambda_k - 1)(\lambda_q - 1) = (\sum_{k=1}^{d}(\lambda_k - 1))^2 = 0$ in the last equality.

Thus, we obtain:

$$\mathbb{E}\left[\left(\sum_j \|X_{j\cdot}'\|\right)^2\right] = \mathbb{E}\left[\sum_j \|X_{j\cdot}'\|^2 + 2\sum_{i<j}\|X_{i\cdot}'\|\|X_{j\cdot}'\|\right] \tag{C.36}$$

$$= d + 2\sum_{i<j}\mathbb{E}\left[\|X_{i\cdot}'\|\|X_{j\cdot}'\|\right] \tag{C.37}$$

$$\leq d + (d^2 - d)\left(1 - \frac{\sum_k(\lambda_k - 1)^2}{2d(d+2)(d-1)}\right) \tag{C.38}$$

$$= d^2 - \frac{\sum_k(\lambda_k - 1)^2}{2(d+2)}\,. \tag{C.39}$$

□

Corollary C.6 (Isometry gap bound)**.** *Suppose the same setup as in Theorem 6.3. Then, we have:*

$$\mathbb{E}_W[\psi(X')|X] \leq \psi(X) + \log\left(1 - \frac{\sum_k(\lambda_k - 1)^2}{2d^2(d+2)}\right). \tag{C.40}$$

Remark C.7. Notice that the term $\frac{\sum_{k=1}^{d}(\lambda_k-1)^2}{2d^2(d+2)} = \mathcal{O}(\frac{1}{d})$, yielding $\log\left[1 - \frac{\sum_{k=1}^{d}(\lambda_k-1)^2}{2d^2(d+2)}\right] \leq 0$.

Proof. From Lemma 6.2, we know that:

$$-\log \mathcal{I}(\text{BN}(X')) \leq -\log \mathcal{I}(X) - \log \frac{d^2}{\left(\sum_{j=1}^{d} \|X'_{j\cdot}\|\right)^2} \tag{C.41}$$

$$\implies -\log \mathcal{I}(\text{BN}(X')) \leq -\log \mathcal{I}(X) - \log d^2 + \log \left(\sum_{j=1}^{d} \|X'_{j\cdot}\|\right)^2 \tag{C.42}$$

$$\implies \mathbb{E}_W[-\log \mathcal{I}(\text{BN}(X'))|X] \leq -\log \mathcal{I}(X) - \log d^2 + \mathbb{E}_W\left[\log \left(\sum_{j=1}^{d} \|X'_{j\cdot}\|\right)^2 \middle| X\right] \tag{C.43}$$

$$\leq -\log \mathcal{I}(X) - \log d^2 + \log \mathbb{E}_W\left[\left(\sum_{j=1}^{d} \|X'_{j\cdot}\|\right)^2 \middle| X\right] \tag{C.44}$$

$$\leq -\log \mathcal{I}(X) - \log d^2 + \log \left(d^2 - \frac{\sum_k (\lambda_k - 1)^2}{2(d+2)}\right) \tag{C.45}$$

$$\leq -\log \mathcal{I}(X) + \log \left(1 - \frac{\sum_k (\lambda_k - 1)^2}{2d^2(d+2)}\right), \tag{C.46}$$

where in inequality C.44 we have used the fact that $\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$ and in inequality C.45 we have used the bound obtained in proof Theorem 6.3, equation C.39. Thus, we obtain:

$$\mathbb{E}_W[\psi(X')|X] \leq \psi(X) + \log \left(1 - \frac{\sum_k (\lambda_k - 1)^2}{2d^2(d+2)}\right). \tag{C.47}$$

□

C.2 Isometry gap decay rate

Before we start with the main part of our analysis, let us establish a simple result on the relation between isometry gap and orthogonality:

Lemma C.8 (Isometry gap and orthogonality)**.** *If* $\psi(X) \leq \frac{c}{16d}$*, then eigenvalues of* $X^\top X$ *are within* $[1-c, 1+c]$.

Note that, in order to simplify the calculations, we use the fact that $\frac{1}{d(d+2)} \approx \frac{1}{d^2}$ in the following proofs.

Based on the conditional expectation in Corollary C.6, we have:

$$\mathbb{E}\left[\psi(X^{\ell+1})|X^\ell\right] \leq \psi(X^\ell) - \frac{\sigma_\lambda(X^\ell)}{2d^2}. \tag{C.48}$$

Now, we prove a lemma that is conditioned on the previous layer isometry gap being smaller or larger than $\frac{1}{16d}$.

Lemma C.9 (Isometry gap conditional bound)**.** *For X^ℓ being the representations of an MLP under our setting, we have:*

$$\mathbb{E}\left[\psi(X^{\ell+1})\middle| X_\ell, \psi(X^\ell) \leq \frac{1}{16d}\right] \leq \psi(X^\ell)\left(1 - \frac{1}{2d^2}\right), \tag{C.49}$$

$$\mathbb{E}\left[\psi(X^{\ell+1})\middle| X_\ell, \psi(X^\ell) > \frac{1}{16d}\right] \leq \psi(X^\ell) - \frac{1}{32d^3}\,. \tag{C.50}$$

Proof of Lemma C.9. Let $\lambda_k = 1 + \epsilon_k$, and assume without loss of generality that $\sum_{k=1}^d \epsilon_k = 0$. Then, using the numerical inequality $\log(1+x) \geq x - x^2$, when $|x| \leq \frac{1}{2}$ we have:

$$\sigma_\lambda(X) = \frac{1}{d}\sum_{k=1}^d \epsilon_k^2 \tag{C.51}$$

$$\max_k |\epsilon_k| \leq \frac{1}{2} \implies \psi(X) = -\frac{1}{d}\sum_{k=1}^d \log(1+\epsilon_k) \leq -\frac{1}{d}\sum_{k=1}^d (\epsilon_k - \epsilon_k^2) = \sigma_\lambda(X)\,. \tag{C.52}$$

Altogether, we have

$$\max_i |\epsilon_i| \leq \frac{1}{2} \implies \psi(X) \leq \sigma_\lambda(X)\,. \tag{C.53}$$

Now, we can restate the condition in terms of an inequality on the isometry gap. Thus, we can write:

$$d\psi(X) = -\sum_{i=1}^d \log\lambda_i = -\sum_{i=1}^d \log(1+\epsilon_i) \geq -\sum_{i=1}^d \left(\epsilon_i - \frac{3\epsilon_i^2}{6+4\epsilon_i}\right) = \sum_{i=1}^d \frac{3\epsilon_i^2}{6+4\epsilon_i}, \tag{C.54}$$

where we used the fact that $\sum_i \lambda_i = d$ implying $\sum_i \epsilon_i = 0$ and also used the inequality $\log(1+x) \leq x - \frac{6+x}{6+4x}$ when $x \geq -1$ for ϵ_i's. Note that because $|\varepsilon_i| \leq \frac{1}{2}$, we get $6 + 4\epsilon_i \geq 4$, all terms on the right-hand side are positive, implying that each term is bounded by the upper bound:

$$\frac{3\epsilon_i^2}{6+4\epsilon_i} \leq d\psi(X) \quad \forall i \in \{1, 2, \ldots, d\}\,. \tag{C.55}$$

By construction, we have $\epsilon_i \geq -1$ and $\frac{3\epsilon_i^2}{6+4\epsilon_i} \leq \frac{1}{16}$. Since $6 + 4\epsilon_i \geq 2$, we can multiply both sides by $6 + 4\epsilon_i$ and conclude $3\epsilon_i^2 - \frac{6+4\epsilon_i}{16} \leq 0$. We can now solve the quadratic equation and obtain $\frac{1-\sqrt{74}}{24} \leq \epsilon_i \leq \frac{1+\sqrt{74}}{24}$ which numerically becomes $-0.35 \leq \epsilon_i \leq 0.4$, implying $|\epsilon_i| < 0.5$.

By solving the quadratic equation above we can guarantee that

$$\psi(X) \leq \frac{1}{16d} \implies \max_i |\epsilon_i| \leq \frac{1}{2} \implies \psi(X) \leq \sigma_\lambda(X)\,. \tag{C.56}$$

Furthermore, we can restate the condition on maximum using:

$$\max_k |\epsilon_k| = \sqrt{\max_k \epsilon_k^2} \leq \sqrt{\sum_k \epsilon_k^2} = \sqrt{d\sigma_\lambda(X)} \tag{C.57}$$

and conclude that

$$\sigma_\lambda(X) \leq \frac{1}{4d} \implies \max_i |\varepsilon_i| \leq \frac{1}{2} \implies \psi(X) \leq \sigma_\lambda(X)\,. \tag{C.58}$$

Using this statement, we have

$$\sigma_\lambda(X) \leq \frac{1}{16d} \implies \psi(X) \leq \sigma_\lambda(X) \implies \psi(X) \leq \frac{1}{16d}\,. \tag{C.59}$$

If we negate and flip the two sides we arrive at

$$\psi(X) > \frac{1}{16d} \implies \sigma_\lambda(X) > \frac{1}{16d}\,. \tag{C.60}$$

Thus, we can simplify the recurrence

$$\mathbb{E}[\psi(X^{\ell+1})|X^\ell] \leq \psi(X^\ell) - \frac{\sigma_\lambda(X^\ell)}{2d^2} \tag{C.61}$$

as follows

$$\mathbb{E}\left[\psi(X^{\ell+1})\middle| X^\ell, \psi(X^\ell) \leq \frac{1}{16d}\right] \leq \psi(X^\ell)\left(1 - \frac{1}{2d^2}\right), \tag{C.62}$$

$$\mathbb{E}\left[\psi(X^{\ell+1})\middle| X^\ell, \psi(X^\ell) > \frac{1}{16d}\right] \leq \psi(X^\ell) - \frac{1}{32d^3}, \tag{C.63}$$

where we used equation C.56 in the first one and equation C.60 in the second one. □

Proof of Theorem 6.1. From Lemma C.1, we know that $\psi(X^0) \geq \psi(X^1) \geq \cdots \geq \psi(X^L) \geq 0$, for any layer $0 \leq \ell \leq L$. Thus, we get using equation C.50:

$$\mathbb{E}\left[\psi(X^{\ell+1})\middle| X^\ell, \psi(X^\ell) > \frac{1}{16d}\right] \leq \psi(X^\ell) - \frac{1}{32d^3} \tag{C.64}$$

$$= \left(1 - \frac{1}{32d^3\psi(X^\ell)}\right)\psi(X^\ell) \tag{C.65}$$

$$\leq \left(1 - \frac{1}{32d^3\psi(X^0)}\right)\psi(X^\ell), \tag{C.66}$$

where in the last step we have used the fact that $\psi(X^\ell) \leq \psi(X^0)$.

Thus, we can combine equation C.49 and equation C.50 and obtain:

$$\mathbb{E}[\psi(X^{\ell+1})|X^\ell] = \mathbb{E}\left[\psi(X^{\ell+1})\middle| X^\ell, \psi(X^\ell) \leq \frac{1}{16d}\right] \mathbf{1}_{\psi(X^\ell)\leq\frac{1}{16d}} \tag{C.67}$$

$$+ \mathbb{E}\left[\psi(X^{\ell+1})\middle| X^\ell, \psi(X^\ell) > \frac{1}{16d}\right] \mathbf{1}_{\psi(X^\ell)>\frac{1}{16d}} \tag{C.68}$$

$$\leq \max\left(\mathbb{E}\left[\psi(X^{\ell+1})\middle| X^\ell, \psi(X^\ell) \leq \frac{1}{16d}\right], \mathbb{E}\left[\psi(X^{\ell+1})\middle| X^\ell, \psi(X^\ell) > \frac{1}{16d}\right]\right) \tag{C.69}$$

$$\leq \max\left(1 - \frac{1}{2d^2}, 1 - \frac{1}{32d^3\psi(X^0)}\right)\psi(X^\ell) \tag{C.70}$$

$$= \left(1 - \min\left(\frac{1}{2d^2}, \frac{1}{32d^3\psi(X^0)}\right)\right)\psi(X^\ell) \tag{C.71}$$

$$= \left(1 - \frac{1}{\max(2d^2, 32d^3\psi(X^0))}\right)\psi(X^\ell) \tag{C.72}$$

$$\leq \exp\left[-\underbrace{\frac{1}{\max(2d^2, 32d^3\psi(X^0))}}_{k}\right]\psi(X^\ell) \tag{C.73}$$

$$= \exp\left(-\frac{1}{k}\right)\psi(X^\ell)\,. \tag{C.74}$$

By iterated expectations over X^ℓ we get:

$$\mathbb{E}[\psi(X^{\ell+1})] \leq \exp\left(-\frac{1}{k}\right)\mathbb{E}[\psi(X^\ell)] \leq \exp\left(-\frac{\ell}{k}\right)\psi(X^0)\,. \tag{C.75}$$

Note that since $\max(2d^2, 32d^3\psi(X^0)) \leq 2d^2(1 + 16d\psi(X^0))$, we can conclude the proof. □

C.3 Gradient norm bound

In the following section, we denote by $H^\ell = W^\ell X^\ell$ the pre-normalization values. Moreover, we define as $\mathcal{F}_L : \mathbb{R}^{d\times d} \to \mathbb{R}^{C\times d}$, where C is the number of output classes, as the functional composition of an L layers MLP, following the update rule defined in equation 6.3, i.e.:

$$\mathcal{F}_L(X_L) = \textsc{Bn}(W^L\mathcal{F}_{L-1}(X_{L-1}))\,. \tag{C.76}$$

Let us restate the theorem for completeness:

Theorem C.10 (Restated Thm. 6.5)**.** *For any $\mathcal{O}(1)$-Lipschitz loss function L and non-degenerate input X_0, we have:*

$$\log\|\frac{\partial \mathcal{L}}{\partial W^\ell}\| \lesssim d^5 \left(\psi(X_0)^3 + 1 - e^{-\frac{L}{32d^4}}\right) \tag{C.77}$$

holds for all $\ell \leq L$, where possibly $L \to \infty$.

In particular, the following lemma guarantees that the Lipschitz conditions are met for practical loss functions:

Lemma C.11. *In a classification setting, cross entropy and mean squared error losses are $\mathcal{O}(1)$-Lipschitz.*

The main idea for the feasibility of this theorem is the presence of perfectly isometric weight matrices that are orthonormal, and the linear activation that does not lead to vanishing or exploding gradients. The only remaining layers to be analyzed are the batch normalization layers. Thus, our main goal is to show that the sum of log-norm gradient of BN layers remains bounded even if the network has infinite depth. To do so, we shall relate the norm of the gradient of those layers to the isometry gap of representations, and use the bounds from the previous section to establish that the log-norm sum is bounded.

Proof of Thm. C.10. Now, considering an L layer deep model, where L can possibly be $L \to \infty$, we can finalize the proof of Theorem 6.5. Consider an MLP model as defined in equation 6.3. Let $H^L = W^L X^L$ be the logits of the model, where $H^L \in \mathbb{R}^{C\times d}$, $W^L \in \mathbb{R}^{C\times d}$, W^L is an orthogonal matrix and C is the number of output classes. Denote as $\mathcal{L}(H^L, y)$ the loss of model for an input matrix, with ground truth y. Then, applying the chain rule, we have:

$$\frac{\partial \mathcal{L}}{\partial W^\ell} = \frac{\partial \mathcal{L}}{\partial H^L}\frac{\partial H^L}{\partial X^L}\frac{\partial X^L}{\partial X^{L-1}}\cdots\frac{\partial X^{\ell+2}}{\partial X^{\ell+1}}\frac{\partial X^{\ell+1}}{\partial H^\ell}\frac{\partial H^\ell}{\partial W^\ell}\,. \tag{C.78}$$

By taking the logarithm of the norm of each factor and applying Lemma C.21, we get:

$$\log\|\frac{\partial \mathcal{L}}{\partial W^\ell}\| \leq \log\|\frac{\partial \mathcal{L}}{\partial H^L}\| + \log\underbrace{\|\frac{\partial H^L}{\partial X^L}\|}_{\|W^L\|} + \sum_{k=\ell+1}^{L}\log\|\frac{\partial X^{k+1}}{\partial X^k}\| + \log\underbrace{\|\frac{\partial X^{\ell+1}}{\partial H^\ell}\|}_{\|J_{\text{BN}}(H^\ell)\|} + \log\|\frac{\partial H^\ell}{\partial W^\ell}\| \tag{C.79}$$

$$\leq \log\|\frac{\partial \mathcal{L}}{\partial H^L}\| + \sum_{k=\ell}^{L}\log\|J_{\text{BN}}(H^k)\| + \log\|\frac{\partial H^\ell}{\partial W^\ell}\|\,. \tag{C.80}$$

where $\log \underbrace{\|\frac{\partial H^L}{\partial X^L}\|}_{\|W^L\|} = \log \|W^L\| = 0$, since the orthogonal matrix W^L has operator norm 1.

Since $\frac{\partial H^\ell}{\partial W^\ell} = X^\ell$ and $X^\ell = \textsc{Bn}(H^{\ell-1})$ is batch normalized, this means that $\|\frac{\partial H^\ell}{\partial W^\ell}\| \leq d$. Thus, the main part is to bound the Jacobian log-norms, which is provided by the following lemma:

Lemma C.12. *We have*

$$\sum_{k=1}^{L} \log\|J_{\textsc{Bn}}(H^k)\| \lesssim d^5(\psi(X_0)^3 + 1 - e^{-\frac{L}{32d^4}}) \,. \tag{C.81}$$

Finally, we can plug the bound from Lemma C.12 in equation C.80 and obtain the conclusion:

$$\log\|\frac{\partial \mathcal{L}}{\partial W^\ell}\| \lesssim \log\|\frac{\partial \mathcal{L}}{\partial H^L}\| + \log d + d^5\left(\psi(X_0)^3 + 1 - e^{-\frac{L}{32d^4}}\right) \,. \tag{C.82}$$

Note that, for $L \to \infty$, we get:

$$\log\|\frac{\partial \mathcal{L}}{\partial W^\ell}\| \lesssim \log\|\frac{\partial \mathcal{L}}{\partial H^L}\| + \log d + d^5(\psi(X_0)^3 + 1) \,. \tag{C.83}$$

In order to conclude the bound, it suffices to show that the norm of the gradient of the loss with respect to the logits is bounded, which is the objective of Lemma C.11. □

Proof of Lemma C.12. The proof of this lemma is chiefly relying on the following bound on the Jacobian of batch normalization layers, which we will state and prove beforehand.

Lemma C.13 (Log-norm bound)**.** *If $X \in \mathbb{R}^{d\times d}$ is the input to a* Bn *layer, its Jacobian operator norm is bounded by*

$$\log \|J_{\textsc{Bn}}(X)\|_{op} \leq d\psi(X) + 1 \,. \tag{C.84}$$

Furthermore, if $\psi(X) \leq \frac{1}{16d}$, then we have

$$\log \|J_{\textsc{Bn}}(X)\|_{op} \leq 2\sqrt{d\psi(X)} \,. \tag{C.85}$$

Based on the lemma above, we shall define S as the hitting time, corresponding to the first layer in our case, that the isometry gap drops below the critical value of $\frac{1}{16d}$:

$$S = \min\left\{\ell : \psi(X^\ell) \leq \frac{1}{16d}\right\} \,. \tag{C.86}$$

So, we first bound the total log-grad norm for layers 1 up to S, and subsequently $S+1$ up to L:

$$\log \|J_{\mathcal{F}_L}(X)\|_{op} \leq \sum_{\ell=1}^{L} \log \|J_{\text{BN}}(X^\ell)\|_{op} \tag{C.87}$$

$$\leq \sum_{\ell=1}^{S} (d\psi(X^\ell) + 1) + 2 \sum_{\ell=S+1}^{L} \sqrt{d\psi(X^\ell)} \tag{C.88}$$

$$\leq \sum_{\ell=1}^{S} (d\psi(X^0) + 1) + 2 \sum_{\ell=S+1}^{L} \sqrt{d\psi(X^\ell)} \tag{C.89}$$

$$= S(d\psi(X^0) + 1) + 2 \sum_{\ell=S+1}^{L} \sqrt{d\psi(X^\ell)}, \tag{C.90}$$

where we have used that $\psi(X^0)$ as an upper bound on $\psi(X^\ell)$.

Thus, taking expectation we get:

$$\mathbb{E} \log \|J_{\mathcal{F}_L}(X)\|_{op} \leq (d\psi(X^0) + 1)\mathbb{E}[S] + 2 \sum_{\ell=S+1}^{L} \mathbb{E}\sqrt{d\psi(X^\ell)}\,. \tag{C.91}$$

Note that S is a random variable, which is why the expectation over the number of layers appears at the last line. Thus, we can bound the log-norm by bounding $\mathbb{E}[S]$ and the summation separately.

Lemma C.14 (stopping time bound)**.** *We have* $\mathbb{E}[S] \lesssim 512d^4\psi(X^0)^2$ *if* $\psi(X_0) > \frac{1}{16d}$*, and* $\mathbb{E}[S] = 0$ *if* $\psi(X_0) \leq \frac{1}{16d}$.

Lemma C.15 (second phase bound)**.** *We have*

$$\sum_{\ell=S+1}^{L} \mathbb{E}\sqrt{d\psi(X^\ell)} \leq 32d^{4.5}\psi(X_0)^{0.5}\left(1 - e^{-\frac{L}{32d^4}}\right)\,.$$

Thus, we have the following 2 cases, based on whether $\psi(X_0)$ is below or over the $\frac{1}{16d}$ threshold. If we plug the bounds in equation C.91 we get the following.

If $\psi(X_0) \leq \frac{1}{16d}$, then:

$$\mathbb{E} \log \|J_{\mathcal{F}_L}(X)\|_{op} \leq 32d^{4.5}\psi(X_0)^{0.5}\left(1 - e^{-\frac{L}{32d^4}}\right) \tag{C.92}$$

$$\lesssim d^{4.5}\psi(X_0)^{0.5}\left(1 - e^{-\frac{L}{32d^4}}\right), \tag{C.93}$$

and if $\psi(X_0) > \frac{1}{16d}$ then:

$$\mathbb{E} \log \|J_{\mathcal{F}_L}(X)\|_{op} \leq 512d^4\psi(X_0)^2(1 + d\psi(X_0)) + 32d^4\left(1 - e^{-\frac{L}{32d^4}}\right) \tag{C.94}$$

$$\lesssim d^5\psi(X_0)^3 + d^4\left(1 - e^{-\frac{L}{32d^4}}\right) \tag{C.95}$$

$$\lesssim d^4\left(d\psi(X_0)^3 + 1 - e^{-\frac{L}{32d^4}}\right)\,. \tag{C.96}$$

In fact, the maximum of the two bounds is

$$\mathbb{E}\log\|J_{\mathcal{F}_L}(X)\|_{op} \lesssim d^5\left(\psi(X_0)^3 + 1 - e^{-\frac{L}{32d^4}}\right) . \tag{C.97}$$

□

Proof of Lemma C.15. By the bound in Lemma C.9, we have

$$\mathbb{E}\left[\psi(X^{\ell+1})\middle|\psi(X^\ell) \le \frac{1}{16d}\right] \le \psi(X^\ell)\left(1 - \frac{1}{2d^2}\right) . \tag{C.98}$$

Since we assumed $\ell \ge S$, the conditional inequality always holds and thus we have the Markov bound

$$\ell \ge S \implies q := \Pr\left\{\psi(X^{\ell+1}) \ge \left(1 - \frac{1}{4d^2}\right)\psi(X^\ell)\right\} \le \frac{1 - \frac{1}{2d^2}}{1 - \frac{1}{4d^2}} \le 1 - \frac{1}{4d^2} . \tag{C.99}$$

We define as failure the event $\bar{A} = \{\psi(X^{\ell+1}) \ge (1 - \frac{1}{4d^2})\psi(X^\ell)\}$ with probability q, and conversely as success the event A with probability $1 - q$. In other words, the probability that $\psi(X^{\ell+1})$ does not decrease by at least a factor of $1 - \frac{1}{4d^2}$ is bounded by the failure probability $1 - \frac{1}{4d^2}$.

Since $\psi(X^{\ell+1}) \le \psi(X^\ell)$, then under the assumption that $\ell \ge S$ we can upper bound $\sqrt{\psi(X^{\ell+1})}$ with $\sqrt{\psi(X^\ell)}$ in case of failure with probability q, and with $\sqrt{(1 - \frac{1}{4d^2})\psi(X^\ell)}$ in case of success

with probability $1-q$:

$$\mathbb{E}\left[\sqrt{\psi(X^{\ell+1})}|\psi(X^{\ell})\right] \tag{C.100}$$

$$= \mathbb{E}\left[\sqrt{\psi(X^{\ell+1})}|\psi(X^{\ell}), \bar{A}\right] q + \mathbb{E}\left[\sqrt{\psi(X^{\ell+1})}|\psi(X^{\ell}), A\right](1-q) \tag{C.101}$$

$$\leq \sqrt{\psi(X^{\ell})}q + \sqrt{\psi(X^{\ell})\left(1-\frac{1}{4d^2}\right)}(1-q) \tag{C.102}$$

$$= \sqrt{\psi(X^{\ell})}\left(q + \sqrt{1-\frac{1}{4d^2}}(1-q)\right) \tag{C.103}$$

$$= \sqrt{\psi(X^{\ell})}\left(\sqrt{1-\frac{1}{4d^2}} + \left(1-\sqrt{1-\frac{1}{4d^2}}\right)q\right) \qquad \text{monotonic in } q \tag{C.104}$$

$$\leq \sqrt{\psi(X^{\ell})}\left(\sqrt{1-\frac{1}{4d^2}} + \left(1-\sqrt{1-\frac{1}{4d^2}}\right)\left(1-\frac{1}{4d^2}\right)\right) \qquad \text{plug } q \leq 1-\frac{1}{4d^2} \tag{C.105}$$

$$= \sqrt{\psi(X^{\ell})}\left(1-\frac{1}{4d^2} + \sqrt{1-\frac{1}{4d^2}}\frac{1}{4d^2}\right) \qquad \text{rearranging terms} \tag{C.106}$$

$$\leq \sqrt{\psi(X^{\ell})}\left(1-\frac{1}{4d^2} + \left(1-\frac{1}{8d^2}\right)\frac{1}{4d^2}\right) \qquad \sqrt{1-x} \leq 1-\frac{x}{2} \text{ for } x \geq 0 \tag{C.107}$$

$$= \sqrt{\psi(X^{\ell})}\left(1-\frac{1}{32d^4}\right) . \tag{C.108}$$

Thus, for $\ell \geq S$, we have

$$\mathbb{E}\sqrt{\psi(X^{\ell+1})} = \mathbb{E}_{X^{\ell}}\mathbb{E}\left[\sqrt{\psi(X^{\ell+1})}|\psi(X^{\ell})\right] \leq \mathbb{E}\sqrt{\psi(X^{\ell})}\left(1-\frac{1}{32d^4}\right) . \tag{C.109}$$

The summation starts from below $\sqrt{d\psi(X_S)}$, and will decay by rate $1-\frac{1}{32d^4}$, which is upper bounded by the geometric series:

$$\sqrt{d\psi(X_S)}\sum_{k=0}^{L}\left(1-\frac{1}{32d^4}\right)^k \leq \sqrt{d\psi(X_0)}32d^4\left(1-\left(1-\frac{1}{32d^4}\right)^{L+1}\right) \tag{C.110}$$

$$\leq 32d^{4.5}\psi(X_0)^{0.5}\left(1-e^{-\frac{L}{32d^4}}\right) . \tag{C.111}$$

□

Proof of Lemma C.14. By Lemma C.9 we have

$$\Pr\left\{\psi(X^\ell) \geq \frac{1}{16d}\right\} \leq \exp\left(-\frac{\ell}{\max(2d^2, 32d^3\psi(X^0))}\right) 16d\psi(X^0)\,. \tag{C.112}$$

Thus, we have

$$\Pr\{S \geq \ell\} \leq \exp\left(-\frac{\ell}{\max(2d^2, 32d^3\psi(X^0))}\right) 16d\psi(X^0)\,. \tag{C.113}$$

Since S is a non-negative integer valued random variable, we can thus bound $\mathbb{E}[S]$ as:

$$\begin{aligned}
\mathbb{E}[S] &= \sum_{\ell=1}^{\infty} P\{S \geq \ell\} && \text{(C.114)}\\
&\leq 16d\psi(X^0) \sum_{\ell=1}^{\infty} \exp\left(\frac{-\ell}{k}\right) && \text{(C.115)}\\
&= 16d\psi(X^0) \frac{1}{\exp(\frac{1}{k}) - 1} && \text{(C.116)}\\
&\leq 16d\psi(X^0)k && \text{(C.117)}\\
&= 16d\psi(X^0)\max(2d^2, 32d^3\psi(X^0)) && \text{(C.118)}\\
&\lesssim 512d^4\psi(X^0)^2. && \text{(C.119)}
\end{aligned}$$

□

Proof of Lemma C.13: Bounding BN grad-norm with isometry gap

The proof of the Lemma relies on two main observations that are crystallized in the following lemmas that first establish a bound on Jacobian operator norm based on the inverse of smallest eigenvalue, and then establish a lower bound for the smallest eigenvalue using the isometry gap.

Lemma C.16. *Let $X \in \mathbb{R}^{d\times d}$ and let $\{\lambda_i\}_{i=1}^d$ be the eigenvalues of $XX^\top$. Then, we have that:*

$$\|J_{\text{BN}}(X)\|_{op}^2 \leq \frac{1}{\lambda_d}, \tag{C.120}$$

where J_{BN} is the Jacobian of the $\text{BN}(\cdot)$ *operator.*

Using the above lemma we have $\log \|J_{\text{BN}}(X)\|_{op} \leq -\log \lambda_d$. The following lemma upper bounds this quantity using isometry gap:

Lemma C.17. *The minimum eigenvalue of a Gram matrix that is trace-normalized is lower-bounded by the isometry gap as* $-\log \lambda_d \leq d\psi(X) + 1$. *Furthermore, if* $\psi(X) \leq \frac{1}{16d}$, *then* $-\log \lambda_d \leq 2\sqrt{d\psi(X)}$.

Plugging these two values we have the bounds

$$\log \|J_{\text{BN}}(X)\|_{op} \leq d\psi(X) + 1, \tag{C.121}$$

$$\psi(X) \leq \frac{1}{16d} \implies \log \|J_{\text{BN}}(X)\|_{op} \leq 2\sqrt{d\psi(X)}\,. \tag{C.122}$$

Now we can turn our attention to the proof of the Lemmas used in the proof. The proof of relationship between minimum eigenvalue and isometry gap is obtained by merely a few numerical inequalities:

Proof of Lemma C.17. Let $\{\lambda_i\}_{i=1}^d$ be the eigenvalues of $X^\top X$. Since the matrix is trace-normalized, we have $\sum_{k=1}^d \lambda_k = d$.

The arithmetic mean of the top $d-1$ values can be written as

$$\frac{1}{d-1}\sum_{k=1}^{d-1} \lambda_k = 1 + \frac{1-\lambda_d}{d-1}\,. \tag{C.123}$$

Thus, we have that their geometric mean is bounded by the same value. Therefore, we have the following bound:

$$\prod_{k=1}^{d} \lambda_k \leq \lambda_d \left(1 + \frac{1-\lambda_d}{d-1}\right)^{d-1} \tag{C.124}$$

$$\implies d \log \mathcal{I}(X) \leq \log(\lambda_d) + (d-1)\log\left(1 + \frac{1-\lambda_d}{d-1}\right), \tag{C.125}$$

where in the second inequality we have taken logarithm of both sides. Now, we can apply the numerical inequality $\log(1+x) \leq x$ to conclude:

$$-d\psi(X) \leq \log \lambda_d + 1 - \lambda_d\,. \tag{C.126}$$

Since λ_d is non-negative, this clearly implies the first inequality: $-\log \lambda_d \leq d\psi(X) + 1$.

For the second inequality first we use the numerical inequality $\log(x) + 1 - x \leq -\frac{(x-1)^2}{2}, \forall x \in [0,1]$ to conclude

$$\frac{(1-\lambda_d)^2}{2} \leq d\psi(X) \tag{C.127}$$

$$\implies \lambda_d \geq 1 - \sqrt{2d\psi(X)} \tag{C.128}$$

$$\implies -\log \lambda_d \leq -\log(1 - \sqrt{2d\psi(X)})\,. \tag{C.129}$$

We can now use the inequality $-\log(1-x) \leq \sqrt{2}x$ for $0 \leq x \leq \frac{1}{2}$ to conclude that

$$-\log \lambda_d \leq 2\sqrt{d\psi(X)} \tag{C.130}$$

when $2\sqrt{d\psi(X)} \leq \frac{1}{2}$, which is equivalent to $\psi(X) \leq \frac{1}{16d}$.

□

For proving Lemma C.13, we first analyze BN operator on a row, and then invoke this bound and the special structure J_{BN} to derive the main proof.

Lemma C.18. *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ defined as $f(x) = \frac{x}{\|x\|}$ be the elementwise normalization of the x. Then:*

$$J_f(x) = \frac{1}{\|x\|} I_{d^2} - \frac{1}{\|x\|^3} x \otimes x, \tag{C.131}$$

where $\otimes$ is the outer product.

Proof. To begin, notice that for $x \in \mathbb{R}^d$ we have $\frac{\partial \|x\|}{\partial x} = \frac{x}{\|x\|}$. Denote by $y_i := [f(x)]_i = \frac{x_i}{\|x\|}$. Then the Jacobian entries become:

$$\frac{\partial y_i}{\partial x_i} = \frac{\|x\| - \frac{x_i}{\|x\|} x_i}{\|x\|^2} = \frac{1}{\|x\|} - \frac{1}{\|x\|^3} x_i x_i, \tag{C.132}$$

$$\frac{\partial y_i}{\partial x_j} = \frac{-\frac{x_j}{\|x\|} x_i}{\|x\|^2} = -\frac{1}{\|x\|^3} x_i x_j \ . \tag{C.133}$$

Assembling the equations into matrix form, we obtain:

$$J_f(x) = \frac{1}{\|x\|} I_{d^2} - \frac{1}{\|x\|^3} x \otimes x \ . \tag{C.134}$$

□

Corollary C.19. *The $J_f(x)$ has the eigenvalue $\frac{1}{\|x\|}$ with multiplicity $d^2 - 1$ and 0 with multiplicity 1.*

Lemma C.20. *Let $XX^\top = \sum_{i=1}^d \lambda_i u_i u_i^\top$, where $XX^\top = U \Lambda U^\top$ is the eigendecomposition of $XX^\top$. Then, we have that $\min_j \|X_{j\cdot}\|^2 \geq \min_k \lambda_k$.*

Proof.

$$\|X_{j\cdot}\|^2 = (XX^\top)_{jj} = \sum_{i=1}^{d} \lambda_i u_{ji}^2 \geq \min_k \lambda_k \sum_{i=1}^{d} u_{ji}^2 = \min_k \lambda_k, \tag{C.135}$$

where in the last equality we have used the fact that U orthogonal. Since this is true for all rows j, we get $\min_j \|X_{j\cdot}\|^2 \geq \min_k \lambda_k$. □

Having established the above properties, we are equipped to prove that the Jacobian of a BN layer is bounded by the inverse of the minimum eigenvalue of its input Gram matrix.

Proof of Lemma C.16. To begin, notice that since $\text{BN} : \mathbb{R}^{d^2} \to \mathbb{R}^{d^2}$, implies that $J_{\text{BN}}(X) \in \mathbb{R}^{d^2 \times d^2}$. Denote $X' = \text{BN}(X)$. Since the normalization happens on each row independently of the other rows, the only non-zero derivatives in the Jacobian correspond to changes in output row i with regards to the same input row i, i.e.:

$$\frac{\partial X'_{ik}}{\partial X_{jl}} = 0, \quad \forall i \neq j, \forall k, l\,. \tag{C.136}$$

This creates a block-diagonal structure in $J_{\text{BN}}(X)$, with d blocks on the main diagonal, where each block has size $d \times d$ and is equal to $J_f(X_{i\cdot})$, where f is as defined in Lemma C.18. Therefore, due to the block-diagonal structure, we know that:

$$\lambda\left[J_{\text{BN}}(X)\right] = \bigcup_{i=1}^{d} \lambda\left[J_f(X_{i\cdot})\right], \tag{C.137}$$

where $\lambda[\cdot]$ denotes the eigenvalue spectrum. From Corollary C.19, we know that

$$\lambda\left[J_{\text{BN}}(X)\right] = \left\{\frac{1}{\|X_{i\cdot}\|}\right\}_{i=1}^{d} \cup \{0\} \tag{C.138}$$

with their respective multiplicites. Finally, using Lemma C.20, this implies:

$$\|J_{\text{BN}}(X)\|_2^2 = \left[\max_i \frac{1}{\|X_{i\cdot}\|}\right]^2 = \left[\frac{1}{\min_i \|X_{i\cdot}\|}\right]^2 \leq \frac{1}{\min_k \lambda_k}\,. \tag{C.139}$$

□

Proof of Lemma C.11. Assuming the the logits are passed through a softmax layer, we analyze the case of Cross Entropy Loss for one sample i in a C-classes classification problem. Denoting $z_i = H_{i\cdot}^L$, we have:

$$\mathcal{L}(z_i, y_i) = -\sum_{i=1}^{C} y_i \log p_i, \tag{C.140}$$

where $p_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$ is the probability vector for sample i after passing through the softmax function.

Computing the partial derivatives, we obtain:

$$\frac{\partial p_i}{\partial z_i} = p_i(1 - p_k), \quad i = k \tag{C.141}$$

$$\frac{\partial p_i}{\partial z_k} = -p_i p_k, \quad i \neq k \, . \tag{C.142}$$

Finally, we can compute the gradient of the loss with respect to the logits:

$$\nabla_{z_k} \mathcal{L} = \sum_{i=1}^{C} \left(-y_i \frac{\partial \log(p_i)}{\partial z_k} \right) \tag{C.143}$$

$$= \sum_{i=1}^{C} \left(-y_i \frac{1}{p_i} \frac{\partial p_i}{\partial z_k} \right) \tag{C.144}$$

$$= \sum_{i \neq k} (y_i p_k) + (-y_k(1 - p_k)) \tag{C.145}$$

$$= p_k \sum_{i \neq k} y_i - y_k(1 - p_k) \tag{C.146}$$

$$\implies \|\nabla_{z_k} \mathcal{L}\| \leq \left\| p_k \sum_{i \neq k} y_k \right\| + \|y_k(1 - p_k)\| \leq 2 \, . \tag{C.147}$$

Since the gradient of the loss with regard to each sample is bounded, we can conclude that the operator norm of the Jacobian of the loss with regards to the logits matrix H^L is also bounded.

In a similar analysis, we now shift our attention towards the Mean Squared Error (MSE) loss:

$$\mathcal{L}(z_i, y_i) = \frac{1}{C} \sum_{i=1}^{C} (y_i - p_i)^2 \, .$$

We want to compute the gradient of the loss with respect to each logit z_k:

$$\|\nabla_{z_k} \mathcal{L}\| = \frac{1}{C} \sum_{i=1}^{C} 2(y_i - p_i) \frac{\partial(-p_i)}{\partial z_k} \tag{C.148}$$

$$\implies \|\nabla_{z_k} \mathcal{L}\| \leq \frac{2}{C} \sum_{i=1}^{C} \left| (y_i - p_i) \frac{\partial p_i}{\partial z_k} \right| \leq \frac{2}{C} \sum_{i=1}^{C} |y_i - p_i| \cdot \left| \frac{\partial p_i}{\partial z_k} \right| \leq 2 \, . \tag{C.149}$$

By substituting these derivatives into the gradient equation, we can derive the gradient for each logit with respect to the MSE loss.

□

Lemma C.21. *Let $X^\ell \in \mathbb{R}^{d\times d}$ be the hidden representations of layer $\ell > 0$ as defined in equation 6.3. Then, we have that:*

$$\log\|\frac{\partial X^{\ell+1}}{\partial X^\ell}\| \leq \log\|J_{\text{BN}}(H^\ell)\| \,. \tag{C.150}$$

Proof of Lemma C.21. By definition, we have that:

$$H^\ell = W^\ell X^\ell, \tag{C.151}$$

$$X^{\ell+1} = \text{BN}(H^\ell) \,. \tag{C.152}$$

Therefore, applying the chain rule, we get:

$$\frac{\partial X^{\ell+1}}{\partial X^\ell} = \frac{\partial X^{\ell+1}}{\partial H^\ell}\frac{\partial H^\ell}{\partial X^\ell} = J_{\text{BN}}(H^\ell)W^\ell \,. \tag{C.153}$$

Taking the logarithm of the norm of this quantity, we reach the conclusion:

$$\log\|\frac{\partial X^{\ell+1}}{\partial X^\ell}\| \leq \log\|J_{\text{BN}}(H^\ell)\| + \log\|W^\ell\| = \log\|J_{\text{BN}}(H^\ell)\|, \tag{C.154}$$

where we have used the fact that the spectrum of the orthogonal matrix W^ℓ contains only the singular value 1 with multiplicity d. □

C.4 Linear independence in common datasets

In this section, we empirically verify the assumption that popular datasets do not suffer from rank collapse in most practical settings.

We provide empirical evidence for CIFAR10, MNIST, FashionMNIST and CIFAR100. We test this assumption by randomly sampling 100 input batches of sizes $n = 16, 32, 64, 128, 256, 512$ from each of these datasets and then measuring the rank of the Gram matrix of these randomly sampled batches using the `matrix_rank()` function provided in PyTorch. We stop at size 512 since we approach the dimensionality of some datasets, i.e. FashionMNIST, MNIST. We show in Table C.1 the average rank with the standard deviation for each n, over 100 randomly sampled batches.

We would like to remark that these datasets are fairly simple in terms of dimensionality and semantics, which can lead to correlated samples. Furthermore, the rank degeneracy can be alleviated even in the larger batch sizes through various data augmentations techniques. Note that these datasets have a high degree of correlation between samples. Most notably, the average cosine similarity between samples in a 512 size batch is $0.81, 0.40, 0.58, 0.81$ for CIFAR10, MNIST, FashionMNIST and CIFAR100 respectively.

Table C.1: Average rank of Gram matrix of input batches of size n from different datasets. Mean and standard deviation are computed over 100 randomly selected input batches, where the samples are chosen without replacement.

Dataset	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$
CIFAR10	16.0 ± 0.0	32.0 ± 0.0	64.0 ± 0.0	127.99 ± 0.09	221.06 ± 2.89	203.70 ± 3.58
MNIST	16.0 ± 0.0	32.0 ± 0.0	64.0 ± 0.0	128.00 ± 0.00	250.48 ± 1.53	318.04 ± 2.82
FashionMNIST	16.0 ± 0.0	32.0 ± 0.0	64.0 ± 0.0	128.00 ± 0.00	238.19 ± 3.27	275.37 ± 4.20
CIFAR100	16.0 ± 0.0	32.0 ± 0.0	64.0 ± 0.0	127.92 ± 0.27	218.11 ± 3.69	201.51 ± 3.95

C.5 Activation shaping

In this section, we explain the full procedure for shaping the activation, as well as expand on the heuristic we use to choose the pre-activation gain. Under the functional structure of the MLP in equation 6.11, let α_ℓ be the pre-activation gain.

More formally, since the gradient norm has an exponential growth in depth, as shown in Figure C.9, we can compute the linear growth rate of log-norm of gradients in depth. We define the rate of explosion for a model of depth L and gain α at layer ℓ as the slope of the log norm of the gradients:

$$R(\ell, \alpha_\ell) = \frac{\log \|\nabla_{W_\ell} \mathcal{L}\| - \log \|\nabla_{W_{\ell-10}} \mathcal{L}\|}{10} . \tag{C.155}$$

Since the rate function is not perfectly linear and has noisy peaks, we measure the slope with a 10 layer gap in order to capture the true behaviour instead of the noise.

Our goal is to choose α such that the sum of the rates across the layers in depth is bounded by a constant that does not depend on the depth of the model, i.e. $R(\ell, \alpha_\ell) \leq \beta$, where β is independent of L. One choice to achieve this is to pick a gain such that the sum of the rates behaves like a decaying harmonic sum in depth.

To this end, we measure the rate of explosion at multiple layers in a 1000 layer deep model, for various gains α which are constant across the layers in Figure C.1 and notice that it behaves as $R(\ell) = c_1 \alpha_2^c$. In order to have the sum of rates across layers behave like a bounded harmonic series in depth, we must choose the gain such that it decays roughly as $\alpha^{c_2} = \ell^{-k}$ where $k > 1$ results in convergence. Therefore, we can obtain a heuristic for picking a gain such that the gradients remain bounded in depth as $\alpha_\ell = \ell^{-k/c_2}$, where we refer to k/c_2 as the gain exponent.

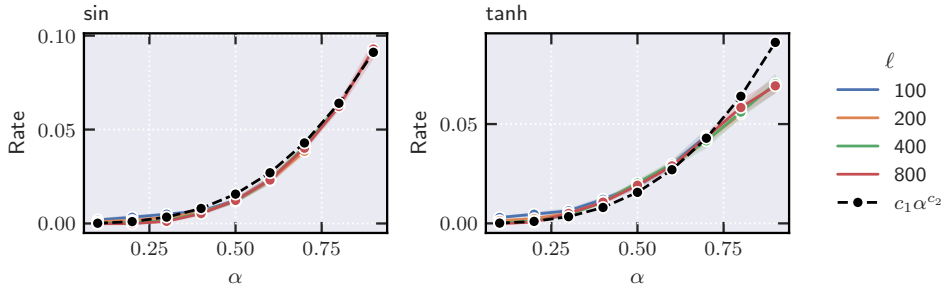


Figure C.1: Explosion rate of the log norm of the gradients at initialization for an MLP model with orthogonal weights and batch normalization, for sin and tanh nonlinearities measured for a 1000 layer deep model at layers ℓ as a function of gain α. The black trace shows the fitted function $c_1\alpha_2^c$. Traces are averaged over 10 independent runs, with the shades showing the 95% confidence interval.

This reduces the problem to picking the exponent such that the sum stays bounded. We show how the behaviour of the explosion rate at the early layers, for various models, is impacted by the exponent in Figure C.2. Note that for several exponent values, we able to reduce the exponential explosion rate and obtain trainable models, which we show in Section C.7.

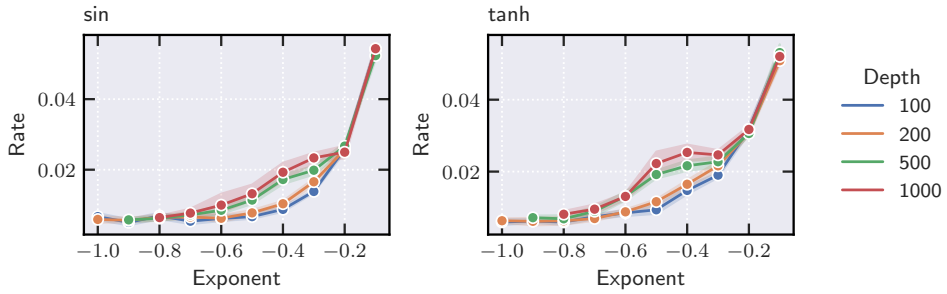


Figure C.2: Explosion rate of the log norm of the gradients at initialization for an MLP model with orthogonal weights and batch normalization, for sin and tanh nonlinearities at depths $100, 200, 500, 1000$ as a function of the gain exponent. Traces are averaged over 10 independent runs, where the shade shows the 95% confidence interval. Rate is measured at $\ell = 10$ to avoid the any transient effects of the function

C.6 Implicit orthogonality during training

In this section, we provide empirical evidence that our architecture during training maintains orthogonality across depths, while maintaining bounded gradients. Figure C.3 shows the evolution of the isometry gap of the weight matrices W_ℓ during training, for models at different depths and different nonlinearities. In order to show that these weights are updated gradient descent, we also show the evolution of the norm of the loss gradients with regards to matrices W_ℓ in Figure C.4.

These experiments are performed on an MLP with orthogonal weight matrices and batch normalization, with sin and tanh activations. The width is set to 100, batch size 100 and learning rate 0.001. The gain exponent is set to a fixed value for all experiments. The measurements are performed on a single batch of size 100 from CIFAR10, after each epoch of training on the same dataset.

C.7 Other experiments

In this section we provide the train and test accuracies of deep MLPs on 4 popular image datasets, namely MNIST, FashionMNIST, CIFAR10, CIFAR100. Hyperparameters and measurements procedure are described in Section 6.3.

Supplementary train and test results on MNIST, FashionMNIST, CIFAR10, CIFAR100

Supplemental figures

We present empirical results in Figure 6.2 showing that degenerate input batches are a hard constraint for orthogonalization without gradient explosion. For MLPs with different depths, we show that by repeating samples in a batch of size 10 we get an exponential gradient explosion, which is unavoidable theoretically.

Furthermore, we show how non-linearities affect the gradient explosion rate in Figure C.9. Using standard batch normalization and fully connected layers from PyTorch we show that non-linearities maintain a large isometry gap. This is a critical issue for our theoretical framework, since we take advantage of the fact that the identity activation achieves perfect orthogonality in order to prove that the gradients remain bounded in depth.

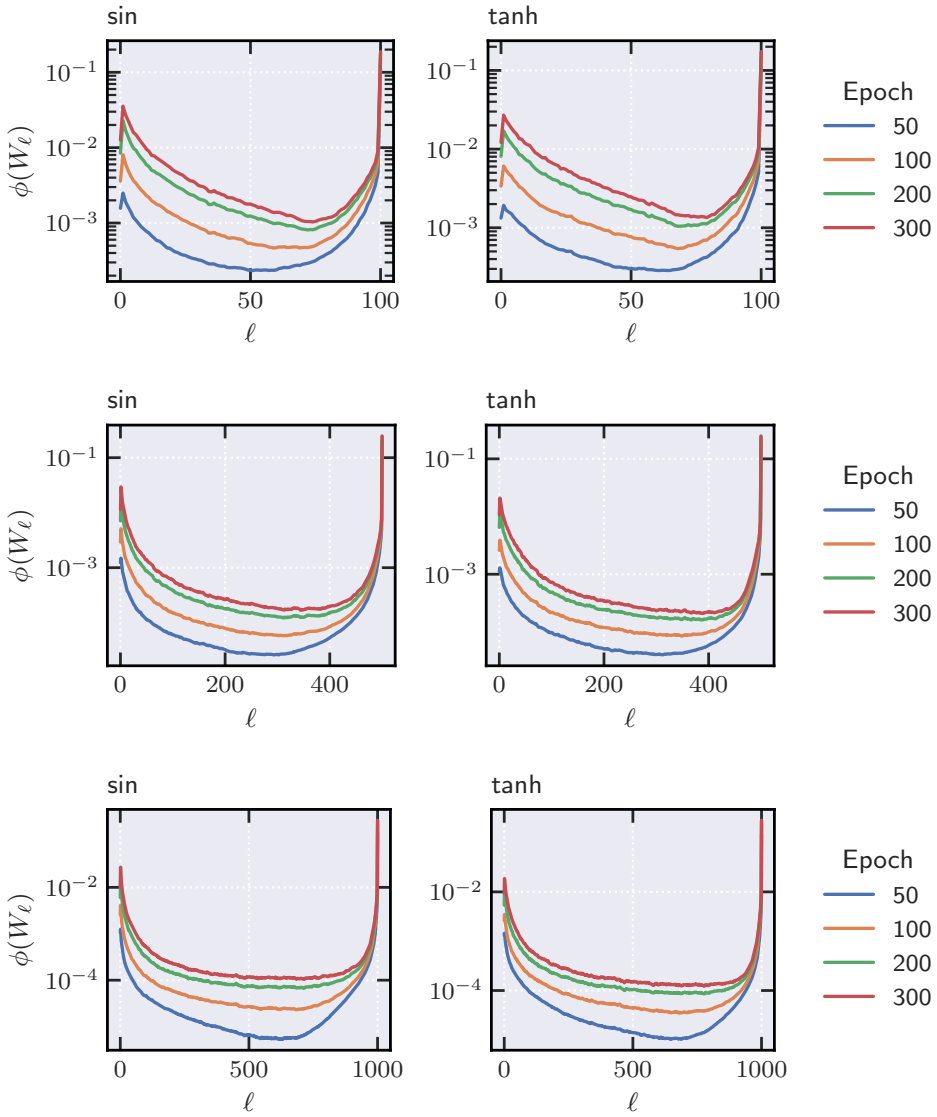


Figure C.3: Contrasting the isometry gap of weight matrices during training for MLPs of depth 100 (top), 500 (middle), 1000 (bottom). The middle layers become increasingly more orthogonal with depth, while maintaining a small isometry gap. During training, the isometry gap also remains low, suggesting the matrices remain close to being orthogonal.

Influence of mean reduction on the gradient bound

In this section, we compare whether adding mean reduction and the additional factor of $\frac{1}{n}$ in the denominator of the batch normalization module influences our gradient bound. As expected, we show in Figure C.10 that in both cases, for the identity activation, the result remains similar, with

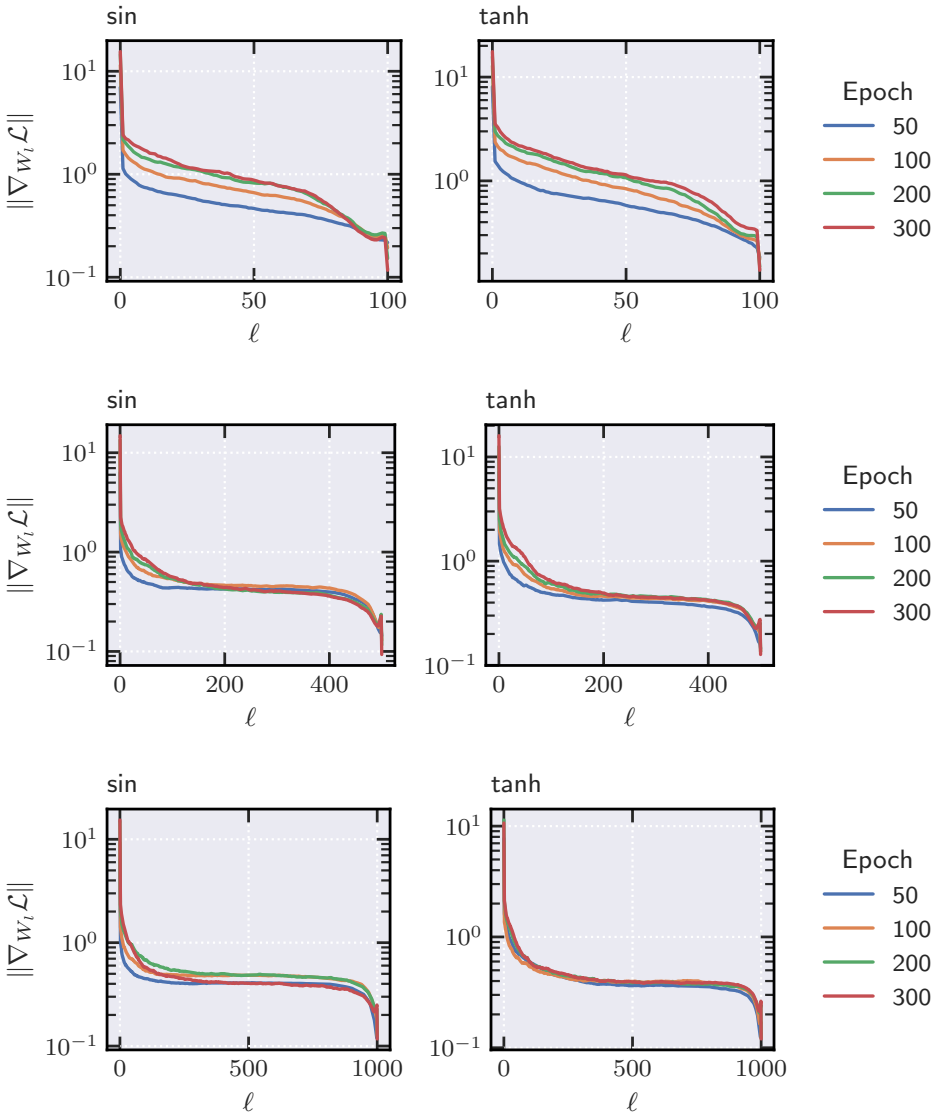


Figure C.4: Contrasting the Frobenius norm of the gradients of the loss with respect to the weights during training for MLPs of depth 100 (top), 500 (middle), 1000 (bottom). The gradients do not vanish during training and across different depths for all layers, suggesting that the orthogonality evidenced in Figure C.3 is not due to the weights not being updated during SGD.

the gradients remaining bounded in depth.

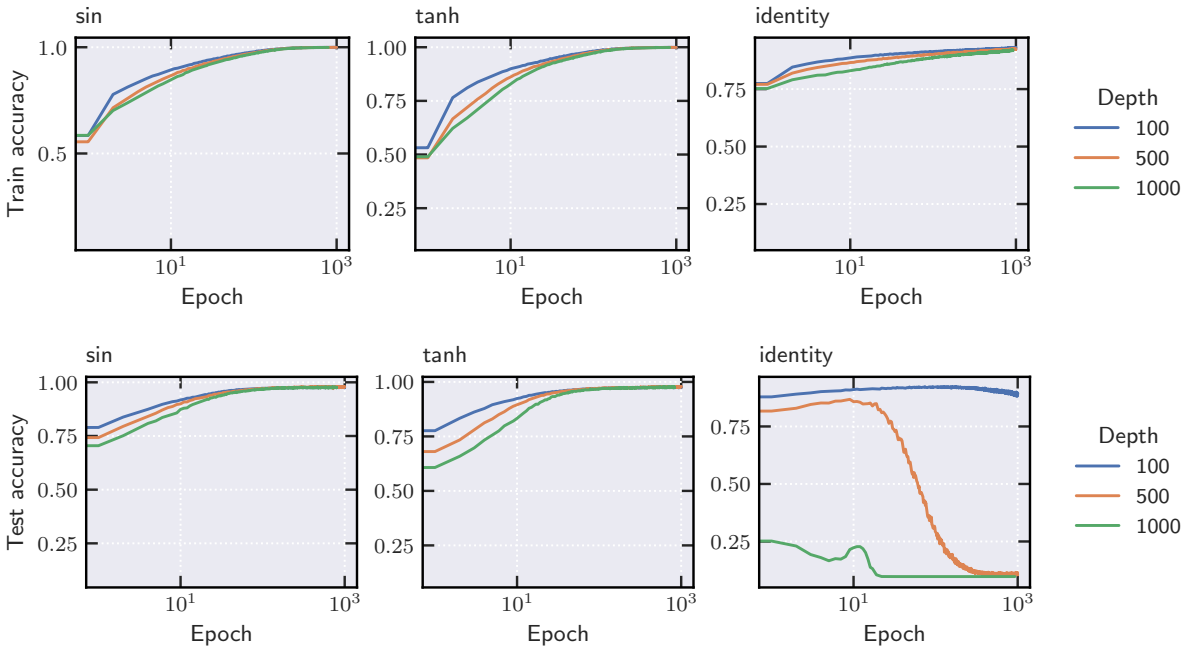


Figure C.5: Contrasting the train and test accuracy of MLPs with gained sin, tanh and identity activations on MNIST. The identity activation performs much worse than the nonlinearities, indicating the fact that the sin and tanh networks are not operating in the linear regime. The network is trained with vanilla SGD and the hyperparameters are width 100, batch size 100, learning rate 0.001.

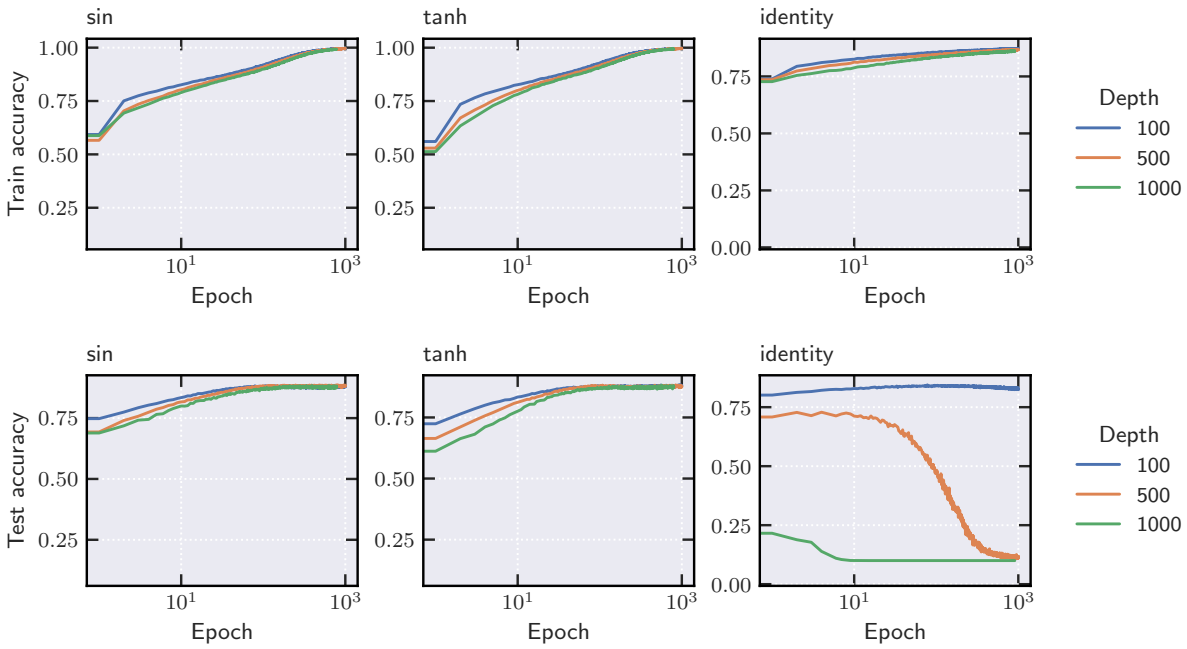


Figure C.6: Contrasting the train and test accuracy of MLPs with gained sin, tanh and identity activations on FashionMNIST. The identity activation performs much worse than the nonlinearities, indicating the fact that the sin and tanh networks are not operating in the linear regime. The networks are trained with vanilla SGD and the hyperparameters are width 100, batch size 100, learning rate 0.001.

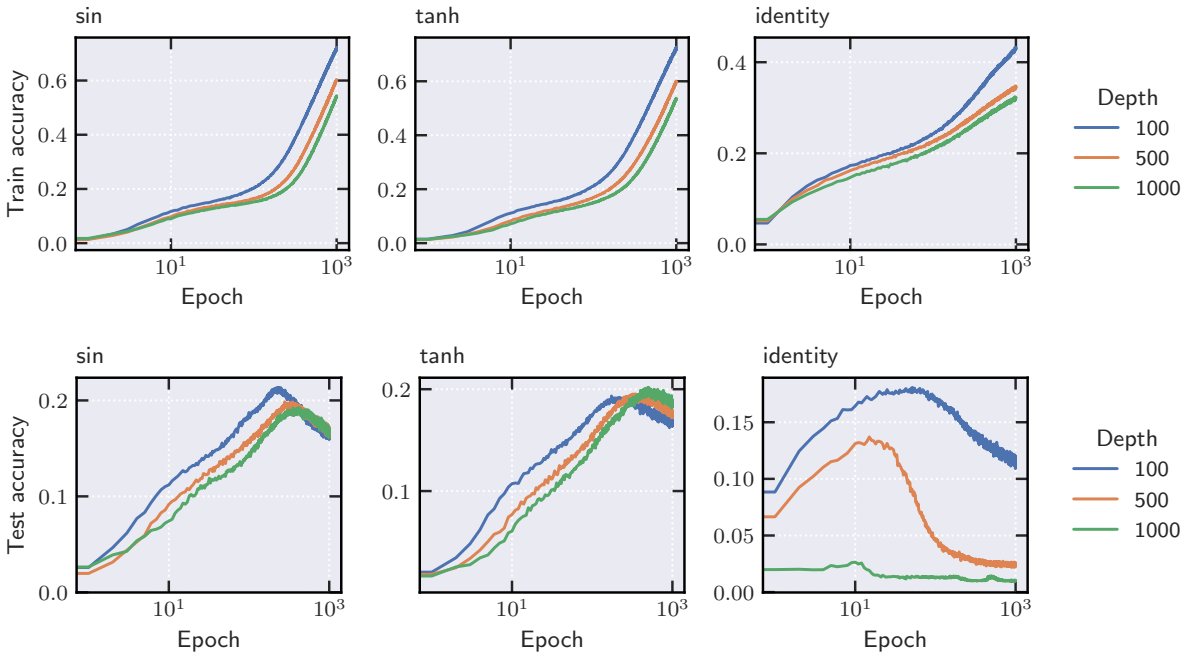


Figure C.7: Contrasting the train and test accuracy of MLPs with gained sin, tanh and identity activations on CIFAR100. The identity activation performs much worse than the nonlinearities, indicating the fact that the sin and tanh networks are not operating in the linear regime. The networks are trained with vanilla SGD and the hyperparameters are width 100, batch size 100, learning rate 0.001.

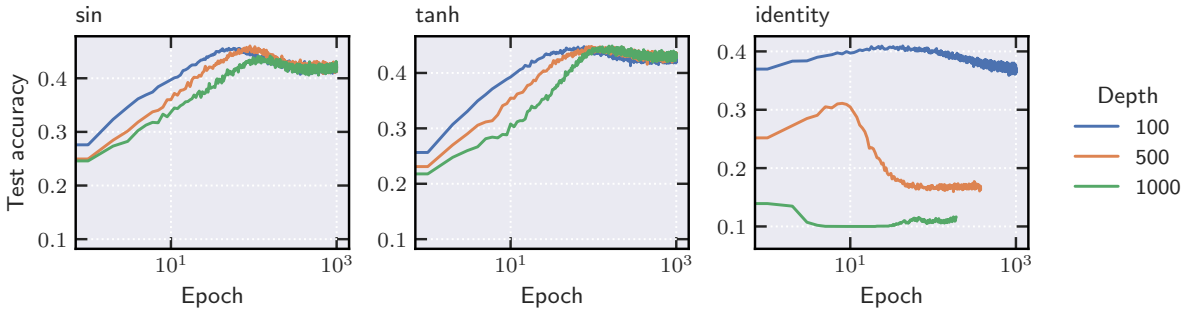


Figure C.8: Contrasting the train test accuracy of MLPs with gained sin, tanh and identity activations on CIFAR10. The identity activation performs much worse than the nonlinearities, indicating the fact that the sin and tanh networks are not operating in the linear regime. The networks are trained with vanilla SGD and the hyperparameters are width 100, batch size 100, learning rate 0.001.

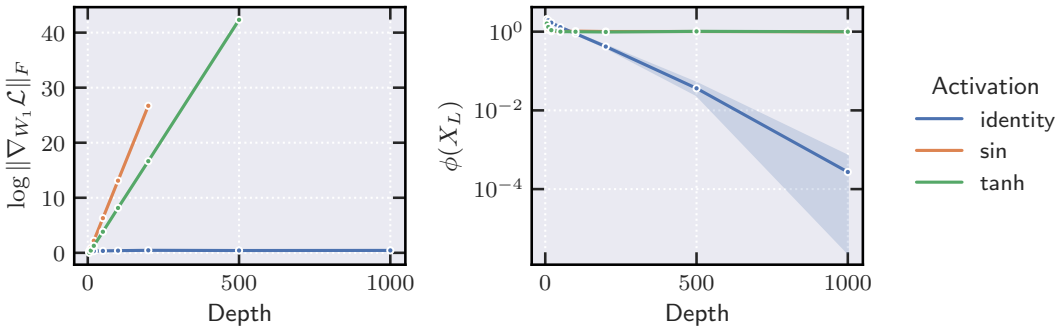


Figure C.9: Left: average log-norm of gradients (log-scale y-axis) at the first layer for networks with different depths, evaluted on CIFAR10. Right: Isometry gap (log-scale y-axis) at the last layer for networks with different depths, evaluted on CIFAR10. The MLP is initialized with orthogonal weights and batch normalization, with standard modules, with $\sin, \tanh,$ identity non-linearities. After stabilizing the isometry gap, the non-linearities have an exponential gradient explosion.

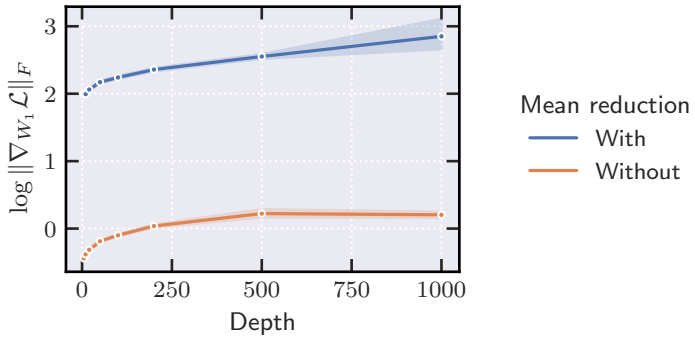


Figure C.10: Comparing the gradient explosion rate in networks with standard batch normalization (blue) and networks with the simplified batch normalization operator from our theoretical framework (orange). Notice that the 2 traces are similar in terms of gradient explosion. Traces are averaged over 10 runs with the shaded regions showing the 95% confidence interval. Samples are from CIFAR10.

List of Tables

List of Figures

List of Algorithms

Bibliography

[AAK21] Naman Agarwal, Pranjal Awasthi, and Satyen Kale. "A deep conditioning treatment of neural networks". In: *Algorithmic Learning Theory*. PMLR. 2021 (cit. on p. 15).

[AB16] Guillaume Alain and Yoshua Bengio. "Understanding intermediate layers using linear classifier probes". In: *arXiv preprint arXiv:1610.01644* (2016) (cit. on p. 1).

[ALL19] Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. "Theoretical analysis of auto rate-tuning by batch normalization". In: *International Conference on Learning Representations* (2019) (cit. on p. 9).

[Aro+19a] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. "A convergence analysis of gradient descent for deep linear neural networks". In: *International Conference on Learning Representations* (2019) (cit. on pp. 11, 68).

[Aro+19b] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. "On exact computation with an infinitely wide neural net". In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 41).

[ASB16] Martin Arjovsky, Amar Shah, and Yoshua Bengio. "Unitary evolution recurrent neural networks". In: *International conference on machine learning*. PMLR. 2016, pp. 1120–1128 (cit. on p. 68).

[Ba+19] Jimmy Ba, Murat Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang. "Generalization of two-layer neural networks: An asymptotic viewpoint". In: *International Conference on Learning Representations*. 2019 (cit. on p. 23).

[Bah+20] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. "Statistical mechanics of deep learning". In: *Annual Review of Condensed Matter Physics* (2020) (cit. on p. 11).

[BCS11] Teodor Banica, Benoit Collins, and Jean-Marc Schlenker. “On polynomial integrals over the orthogonal group”. In: *Journal of Combinatorial Theory, Series A* 118.3 (2011), pp. 778–795 (cit. on pp. 69, 72, 110).

[BD19] Rebekka Burkholz and Alina Dubatovka. “Initialization of relus for dynamical isometry”. In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 67).

[BDS21] Andrew Brock, Soham De, and Samuel L Smith. “Characterizing signal propagation to close the performance gap in unnormalized resnets”. In: *arXiv preprint arXiv:2101.08692* (2021) (cit. on p. 67).

[BGS20] Yaniv Blumenfeld, Dar Gilboa, and Daniel Soudry. “Beyond signal propagation: is feature diversity necessary in deep neural network initialization?” In: *International Conference on Machine Learning*. PMLR. 2020, pp. 960–969 (cit. on p. 67).

[BH89] Pierre Baldi and Kurt Hornik. “Neural networks and principal component analysis: Learning from examples without local minima”. In: *Neural networks* 2.1 (1989), pp. 53–58 (cit. on p. 68).

[BH95] Pierre F Baldi and Kurt Hornik. “Learning in linear neural networks: A survey”. In: *IEEE Transactions on neural networks* 6.4 (1995), pp. 837–858 (cit. on p. 68).

[BHL19] Peter L. Bartlett, David P. Helmbold, and Philip M. Long. “Gradient descent with identity initialization efficiently learns positive-definite linear transformations by deep residual networks”. In: *Neural computation* (2019) (cit. on p. 11).

[Bjo+18] Nils Bjorck, Carla P. Gomes, Bart Selman, and Kilian Q. Weinberger. “Understanding batch normalization”. In: *Advances in Neural Information Processing Systems* 31 (2018) (cit. on pp. 9, 11, 13, 15, 71).

[BKH16] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. “Layer normalization”. In: *arXiv preprint arXiv:1607.06450* (2016) (cit. on pp. 2, 5, 21, 33, 35, 42).

[BL07] Yoshua Bengio and Yann LeCun. “Scaling learning algorithms toward AI”. In: *Large-scale kernel machines* (2007), pp. 321–360 (cit. on pp. 3, 5).

[BM19] Alberto Bietti and Julien Mairal. “On the inductive bias of neural tangent kernels”. In: *Advances in Neural Information Processing Systems* (2019) (cit. on p. 10).

[Bou+12] Philippe Bougerol et al. *Products of random matrices with applications to Schrödinger operators*. Springer Science & Business Media, 2012 (cit. on p. 15).

[Bro+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901 (cit. on p. 65).

[Bro+21] Andrew Brock, Soham De, Samuel L Smith, and Karen Simonyan. "High-performance large-scale image recognition without normalization". In: *arXiv preprint arXiv:2102.06171* (2021) (cit. on pp. 20, 97, 98).

[BSF94] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult". In: *IEEE transactions on neural networks* 5.2 (1994), pp. 157–166 (cit. on pp. 1, 3).

[CB18] 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* (2018) (cit. on pp. 23, 33).

[CB20] Lenaic Chizat and Francis Bach. "Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss". In: *Conference on Learning Theory*. 2020, pp. 1305–1338 (cit. on p. 23).

[CM09] Benoît Collins and Sho Matsumoto. "On some properties of orthogonal Weingarten functions". In: *Journal of Mathematical Physics* 50.11 (2009) (cit. on p. 111).

[CMN22] Benoit Collins, Sho Matsumoto, and Jonathan Novak. "The Weingarten calculus". In: *arXiv preprint arXiv:2109.14890* (2022) (cit. on pp. 69, 72, 110, 111).

[Col03] Benoît Collins. "Moments and cumulants of polynomial random variables on unitary groups, the Itzykson-Zuber integral, and free probability". In: *International Mathematics Research Notices* 2003.17 (2003), pp. 953–982 (cit. on p. 69).

[CŚ06] Benoît Collins and Piotr Śniady. "Integration with respect to the Haar measure on unitary, orthogonal and symplectic group". In: *Communications in Mathematical Physics* 264.3 (2006), pp. 773–795 (cit. on pp. 69, 70, 72, 110).

[CS09] Youngmin Cho and Lawrence Saul. "Kernel methods for deep learning". In: *Advances in neural information processing systems* 22 (2009) (cit. on p. 41).

[CUH15] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. "Fast and accurate deep network learning by exponential linear units (elus)". In: *arXiv preprint arXiv:1511.07289* (2015) (cit. on pp. 2, 4, 42).

[Dan+20] Hadi Daneshmand, Jonas Kohler, Francis Bach, Thomas Hofmann, and Aurelien Lucchi. "Batch normalization provably avoids ranks collapse for randomly initialised deep networks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 18387–18398 (cit. on pp. 2–5, 9, 11–17, 24, 25, 39, 65–67, 69, 71).

[DCL21] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. "Attention is not all you need: Pure attention loses rank doubly exponentially with depth". In: *International Conference on Machine Learning*. 2021, pp. 2793–2803 (cit. on pp. 4, 24, 65, 67).

[Dev+18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *NAACL* (2018) (cit. on pp. 1, 3, 35, 65).

[DFS16] Amit Daniely, Roy Frostig, and Yoram Singer. "Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity". In: *Advances in Neural Information Processing Systems* 29 (2016) (cit. on p. 41).

[DH19] Simon Du and Wei Hu. "Width provably matters in optimization for deep linear neural networks". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 1655–1664 (cit. on p. 68).

[DJB21] Hadi Daneshmand, Amir Joudaki, and Francis Bach. "Batch normalization orthogonalizes representations in deep random networks". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 4896–4906 (cit. on pp. 2, 4–6, 24, 25, 28, 39, 65, 66, 68–70, 72, 78).

[DMR18] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. "The total variation distance between high-dimensional Gaussians". In: *arXiv preprint arXiv:1810.08693* (2018) (cit. on p. 105).

[Doe38] Wolfgang Doeblin. "Sur deux problèmes de M. Kolmogoroff concernant les chaînes dénombrables". In: *Bulletin de la Société Mathématique de France* 66 (1938), pp. 210–220 (cit. on p. 28).

[DPKL19] Giacomo De Palma, Bobak Toussi Kiani, and Seth Lloyd. "Random deep neural networks are biased towards simple functions". In: *Advances in Neural Information Processing Systems* (2019) (cit. on pp. 10, 18).

[Ebe09] Andreas Eberle. “Markov processes”. In: *Lecture Notes at University of Bonn* (2009) (cit. on pp. 14, 28, 29).

[FC18] Jonathan Frankle and Michael Carbin. “The lottery ticket hypothesis: Finding sparse, trainable neural networks”. In: *arXiv preprint arXiv:1803.03635* (2018) (cit. on p. 3).

[Fen+22] Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. “Rank diminishing in deep neural networks”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 33054–33065 (cit. on pp. 23, 24, 32, 65).

[FSM21] Jonathan Frankle, David J Schwab, and Ari S Morcos. “Training batchnorm and only batchnorm: On the expressive power of random features in CNNs”. In: *ICLR* (2021) (cit. on pp. 9, 11).

[Fuk98] Kenji Fukumizu. “Effect of batch learning in multilayer neural networks”. In: *Gen* 1.04 (1998), 1E–03 (cit. on p. 68).

[GARA19] Adria Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. “Deep convolutional networks as shallow gaussian processes”. In: *International Conference on Learning Representations* (2019) (cit. on pp. 17, 18).

[GB10a] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256 (cit. on pp. 1–5, 19, 66–68).

[GB10b] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010, pp. 249–256 (cit. on p. 23).

[GBB11] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. “Deep sparse rectifier neural networks”. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323 (cit. on pp. 2, 42).

[GM+18] 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 (cit. on pp. 10, 18, 23, 24, 27).

[Goy+21] Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. "Larger-scale transformers for multilingual masked language modeling". In: *arXiv preprint arXiv:2105.00572* (2021) (cit. on p. 65).

[Hai10] Martin Hairer. "Convergence of Markov processes". In: *Lecture notes* (2010) (cit. on p. 15).

[Han18] Boris Hanin. "Which neural net architectures give rise to exploding and vanishing gradients?" In: *Advances in neural information processing systems* 31 (2018) (cit. on pp. 3, 4).

[Han22] Boris Hanin. "Correlation Functions in Random Fully Connected Neural Networks at Finite Width". In: *arXiv preprint arXiv:2204.01058* (2022) (cit. on pp. 24, 29).

[HDR19] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. "On the Impact of the Activation function on Deep Neural Networks Training". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 2672–2680 (cit. on p. 42).

[He+15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034 (cit. on pp. 2–5, 66, 67).

[He+16a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on pp. 1, 3–5, 9, 11, 35, 44).

[He+16b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on p. 67).

[He+23] Bobby He, James Martens, Guodong Zhang, Aleksandar Botev, Andrew Brock, Samuel L Smith, and Yee Whye Teh. "Deep transformers without shortcuts: Modifying self-attention for faithful signal propagation". In: *arXiv preprint arXiv:2302.10322* (2023) (cit. on p. 77).

[HG16] Dan Hendrycks and Kevin Gimpel. "Gaussian error linear units (GELUs)". In: *arXiv preprint arXiv:1606.08415* (2016) (cit. on p. 76).

[HJ15] Tamir Hazan and Tommi Jaakkola. "Steps toward deep kernel methods from infinite neural networks". In: *arXiv preprint arXiv:1508.05133* (2015) (cit. on p. 17).

[HN19] Boris Hanin and Mihai Nica. "Finite depth and width corrections to the neural tangent kernel". In: *arXiv preprint arXiv:1909.05989* (2019) (cit. on pp. 24, 27, 29).

[Hoc91] Sepp Hochreiter. "Untersuchungen zu dynamischen neuronalen Netzen". In: *Diploma, Technische Universität München* 91.1 (1991), p. 31 (cit. on p. 3).

[Hoc98] Sepp Hochreiter. "The vanishing gradient problem during learning recurrent neural nets and problem solutions". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6.02 (1998), pp. 107–116 (cit. on pp. 3, 67).

[HR18] Boris Hanin and David Rolnick. "How to start training: The effect of initialization and architecture". In: *Advances in neural information processing systems* 31 (2018) (cit. on pp. 4, 5).

[HSL16] Mikael Henaff, Arthur Szlam, and Yann LeCun. "Recurrent orthogonal networks and long-memory tasks". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 2034–2042 (cit. on p. 68).

[Hua+14] Po-Sen Huang, Haim Avron, Tara N Sainath, Vikas Sindhwani, and Bhuvana Ramabhadran. "Kernel methods match deep neural networks on timit". In: *ICASSP*. 2014 (cit. on p. 10).

[Hua+17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017 (cit. on pp. 5, 9).

[IS15] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International Conference on Machine Learning*. pmlr. 2015, pp. 448–456 (cit. on pp. 2, 5, 9, 10, 25, 35, 65, 67).

[JCF18] Cijo Jose, Moustapha Cissé, and Francois Fleuret. "Kronecker recurrent units". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2380–2389 (cit. on p. 79).

[JDB23a] Amir Joudaki, Hadi Daneshmand, and Francis Bach. "On Bridging the Gap between Mean Field and Finite Width in Deep Random Neural Networks with Batch Normalization". In: *International Conference on Machine Learning* (2023) (cit. on pp. 6, 66, 68, 69).

[JDB23b] Amir Joudaki, Hadi Daneshmand, and Francis Bach. "On the impact of activation and normalization in obtaining isometric embeddings at initialization". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 39855–39875 (cit. on pp. 6, 66, 70, 72, 107).

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in Neural Information Processing Systems* 31 (2018) (cit. on pp. 10, 23, 41).

[Jon01] Jones, Galin L. and Hobert, James P. "Honest exploration of intractable probability distributions via Markov chain Monte Carlo". In: *Statistical Science* (2001), pp. 312–334 (cit. on p. 28).

[KAA19] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. "The normalization method for alleviating pathological sharpness in wide neural networks". In: *Advances in Neural Information Processing Systems*. 2019 (cit. on p. 9).

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on p. 1).

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009) (cit. on p. 19).

[Kha11] Rafail Khasminskii. *Stochastic stability of differential equations*. Vol. 66. Springer Science & Business Media, 2011 (cit. on pp. 10, 14).

[Kla+17] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. "Self-normalizing neural networks". In: *Advances in Neural Information Processing Systems* 30 (2017) (cit. on pp. 2, 4, 10, 18, 23, 42, 44, 49, 67, 76).

[Koh+18] Jonas Kohler, Hadi Daneshmand, Aurelien Lucchi, Ming Zhou, Klaus Neymeyr, and Thomas Hofmann. "Exponential convergence rates for Batch Normalization: The power of length-direction decoupling in non-convex optimization". In: *arXiv preprint arXiv:1805.10694* (2018) (cit. on p. 9).

[KS76] John G Kemeny and J Laurie Snell. *Markov chains*. Springer-Verlag, New York, 1976 (cit. on p. 14).

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. Vol. 25. 2012, pp. 1097–1105 (cit. on p. 1).

[Kus67] Harold J Kushner. *Stochastic stability and control*. Tech. rep. 1967 (cit. on pp. 10, 14).

[KY03] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003 (cit. on pp. 10, 14).

[LCMR19] Mario Lezcano-Casado and David Martınez-Rubio. "Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3794–3803 (cit. on p. 79).

[LDT21] Ekdeep S Lubana, Robert Dick, and Hidenori Tanaka. "Beyond batchnorm: Towards a unified understanding of normalization in deep learning". In: *Advances in Neural Information Processing Systems* (2021) (cit. on pp. 21, 67).

[Lee+19a] 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* (2019) (cit. on pp. 10, 17).

[Lee+19b] 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* 32 (2019) (cit. on pp. 23, 41).

[Liu+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "RoBERTa: A robustly optimized BERT pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019) (cit. on p. 65).

[Liu+22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A ConvNet for the 2020s". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11976–11986 (cit. on p. 65).

[LJH15] Quoc V Le, Navdeep Jaitly, and Geoffrey E Hinton. "A simple way to initialize recurrent networks of rectified linear units". In: *arXiv preprint arXiv:1504.00941* (2015) (cit. on p. 68).

[LNR21] Mufan Li, Mihai Nica, and Dan Roy. "The future is log-Gaussian: ResNets and their infinite-depth-and-width limit at initialization". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 7852–7864 (cit. on p. 24).

[LNR22] Mufan Bill Li, Mihai Nica, and Daniel M. Roy. "The Neural Covariance SDE: Shaped Infinite Depth-and-Width Networks at Initialization". In: *Advances in Neural Information Processing Systems* (2022) (cit. on pp. 23, 24, 27, 29, 68, 69, 77).

[Luo+19] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. "Adaptive gradient methods with dynamic bound of learning rate". In: 2019 (cit. on p. 5).

[Mai+14] Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. "Convolutional kernel networks". In: *Advances in neural information processing systems* 27 (2014) (cit. on p. 41).

[Mar+21] James Martens, Andy Ballard, Guillaume Desjardins, Grzegorz Swirszcz, Valentin Dalibard, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. "Rapid training of deep neural networks without skip connections or normalization layers using deep kernel shaping". In: *arXiv preprint arXiv:2110.01765* (2021) (cit. on pp. 67, 77).

[Meh66] F Gustav Mehler. "Ueber die Entwicklung einer Function von beliebig vielen Variablen nach Laplaceschen Functionen höherer Ordnung." In: *Journal für die Reine und Angewandte Mathematik (in German)* (1866) (cit. on p. 46).

[Met+24] Alexandru Meterez, Amir Joudaki, Francesco Orabona, Alexander Immer, Gunnar Ratsch, and Hadi Daneshmand. "Towards Training Without Depth Limits: Batch Normalization Without Gradient Explosion". In: *The Twelfth International Conference on Learning Representations*. 2024 (cit. on p. 6).

[Mha+17] Zakaria Mhammedi, Andrew Hellicar, Ashfaqur Rahman, and James Bailey. "Efficient orthogonal parametrisation of recurrent neural networks using Householder reflections". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 2401–2409 (cit. on p. 80).

[MHN13] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. "Rectifier nonlinearities improve neural network acoustic models". In: *Proc. ICML*. Vol. 30. 1. 2013, p. 3 (cit. on pp. 2, 76).

[MM15] Dmytro Mishkin and Jiri Matas. "All you need is a good init". In: *arXiv preprint arXiv:1511.06422* (2015) (cit. on p. 68).

[Nea96] Radford M Neal. *Bayesian learning for neural networks*. Springer Science & Business Media, 1996 (cit. on pp. 10, 17, 23).

[New86] Charles M Newman. "The distribution of Lyapunov exponents: Exact results for random matrices". In: *Communications in mathematical physics* (1986) (cit. on p. 95).

[NH10] Vinod Nair and Geoffrey E Hinton. "Rectified linear units improve restricted boltzmann machines". In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, pp. 807–814 (cit. on pp. 2, 4, 42, 76).

[Noc+22a] Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. "Signal Propagation in Transformers: Theoretical Perspectives and the Role of Rank Collapse". In: *arXiv preprint arXiv:2206.03126* (2022) (cit. on p. 33).

[Noc+22b] Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. "Signal propagation in Transformers: Theoretical perspectives and the role of rank collapse". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27198–27211 (cit. on pp. 3, 4, 65).

[Noc+23] Lorenzo Noci, Chuning Li, Mufan Bill Li, Bobby He, Thomas Hofmann, Chris Maddison, and Daniel M Roy. "The Shaped Transformer: Attention Models in the Infinite Depth-and-Width Limit". In: *arXiv preprint arXiv:2306.17759* (2023) (cit. on p. 77).

[Pas+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems*. 2019 (cit. on p. 19).

[PMB13] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks". In: *arXiv preprint arXiv:1211.5063* (2013) (cit. on pp. 3–5).

[Poo+16] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. "Exponential expressivity in deep neural networks through transient chaos". In: *Advances in Neural Information Processing Systems* 29 (2016) (cit. on p. 42).

[PSG17] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. "Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice". In: *Advances in neural information processing systems* 30 (2017) (cit. on pp. 2–5, 68–70).

[PSG18] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. "The emergence of spectral universality in deep networks". In: *International Conference on Artificial Intelligence and Statistics*. 2018, pp. 1924–1932 (cit. on pp. 2, 5, 10, 18, 20, 23–26, 42).

[PW17] Jeffrey Pennington and Pratik Worah. "Nonlinear random matrix theory for deep learning". In: *Advances in Neural Information Processing Systems* 30 (2017) (cit. on pp. 23, 26, 68, 69).

[Raf+20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 5485–5551 (cit. on p. 65).

[Rag+17] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. "On the expressive power of deep neural networks". In: *international conference on machine learning*. PMLR. 2017, pp. 2847–2854 (cit. on p. 42).

[Ros95] Jeffrey S. Rosenthal. "Minorization conditions and convergence rates for Markov chain Monte Carlo". In: *Journal of the American Statistical Association* 90.430 (1995), pp. 558–566 (cit. on p. 28).

[RZL17] Prajit Ramachandran, Barret Zoph, and Quoc V Le. "Searching for activation functions". In: *arXiv preprint arXiv:1710.05941* (2017) (cit. on pp. 2, 23, 42).

[San+18] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. "How Does Batch Normalization Help Optimization?(No, It Is Not About Internal Covariate Shift)". In: *Advances in Neural Information Processing Systems* (2018) (cit. on pp. 5, 9).

[Saw72] Takamitsu Sawa. "Finite-sample properties of the k-class estimators". In: *Econometrica: Journal of the Econometric Society* (1972), pp. 653–680 (cit. on p. 91).

[Sch+17] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. "Deep information propagation". In: *International Conference on Learning Representations* (2017) (cit. on pp. 10, 11, 18, 42).

[SGS15] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. "Training very deep networks". In: *Advances in Neural Information Processing Systems* (2015) (cit. on p. 3).

[Sha19] Ohad Shamir. "Exponential convergence time of gradient descent for one-dimensional deep linear neural networks". In: *Conference on Learning Theory*. PMLR. 2019, pp. 2691–2713 (cit. on p. 68).

[Sil+17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. "Mastering the game of go without human knowledge". In: *nature* (2017) (cit. on p. 9).

[SK16] 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* 29 (2016) (cit. on p. 33).

[SMG13a] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks". In: *arXiv preprint arXiv:1312.6120* (2013) (cit. on pp. 2–5, 11, 13, 15, 20, 23, 24, 42, 65, 67, 68, 70, 71, 97).

[SMG13b] Andrew M Saxe, James L McClellans, and Surya Ganguli. "Learning hierarchical categories in deep neural networks". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 35. 2013 (cit. on p. 68).

[SMG19] Andrew M Saxe, James L McClelland, and Surya Ganguli. "A mathematical theory of semantic development in deep neural networks". In: *Proceedings of the National Academy of Sciences* 116.23 (2019), pp. 11537–11546 (cit. on p. 68).

[SS02] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002 (cit. on p. 41).

[SS04] Alex J Smola and Bernhard Schölkopf. "A tutorial on support vector regression". In: *Statistics and computing* 14 (2004), pp. 199–222 (cit. on p. 41).

[Vor+17] Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal. "On orthogonality and learning recurrent networks with long term dependencies". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 3570–3578 (cit. on p. 79).

[Wei78] Don Weingarten. "Asymptotic behavior of group integrals in the limit of infinite rank". In: *Journal of Mathematical Physics* 19.5 (1978), pp. 999–1001 (cit. on pp. 69, 72, 110).

[Woo+23] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. "ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 16133–16142 (cit. on p. 65).

[WR06] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. Vol. 2. 3. MIT press Cambridge, MA, 2006 (cit. on p. 42).

[Wu+21] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. "Cvt: Introducing convolutions to vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 22–31 (cit. on p. 65).

[Xia+18] Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. "Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks". In: *International Conference on Machine Learning*. 2018, pp. 5393–5402 (cit. on pp. 3, 4, 23, 68, 70).

[Yan19] Greg Yang. "Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation". In: *arXiv preprint arXiv:1902.04760* (2019) (cit. on p. 41).

[Yan+19] Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. "A Mean Field Theory of Batch Normalization". In: *International Conference on Learning Representations*. 2019 (cit. on pp. 4, 5, 10, 18, 23–28, 30, 32, 39, 42, 66–69, 72, 75, 78, 141).

[Yan20] Greg Yang. "Tensor programs ii: Neural tangent kernel for any architecture". In: *arXiv preprint arXiv:2006.14548* (2020) (cit. on p. 5).

[YSJ17] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. "Global optimality conditions for deep neural networks". In: *arXiv preprint arXiv:1707.02444* (2017) (cit. on p. 68).

[ZBM22] Guodong Zhang, Aleksandar Botev, and James Martens. "Deep learning without shortcuts: Shaping the kernel with tailored rectifiers". In: *arXiv preprint arXiv:2203.08120* (2022) (cit. on p. 77).

[ZDM19] Hongyi Zhang, Yann N Dauphin, and Tengyu Ma. "Fixup initialization: Residual learning without normalization". In: *arXiv preprint arXiv:1901.09321* (2019) (cit. on p. 5).

[ZF14] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer. 2014, pp. 818–833 (cit. on p. 5).

[Zha+17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization". In: *International Conference on Learning Representations*. 2017 (cit. on p. 41).

[Zha19] Jiawei Zhang. "Gradient descent based optimization algorithms for deep learning models training". In: *arXiv preprint arXiv:1903.03614* (2019) (cit. on p. 5).

[Zha+21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning (still) requires rethinking generalization". In: *Communications of the ACM* 64.3 (2021), pp. 107–115 (cit. on p. 41).