Proceedings of Machine Learning Research vol 145:1–39, 2022 3rd Annual Conference on Mathematical and Scientific Machine Learning

Freeze and Chaos: NTK views on DNN Normalization, Checkerboard and Boundary Artifacts

Arthur Jacot ARTHUR.JACOT@EPFL.CH
Franck Gabriel FRANCK.GABRIEL@EPFL.CH
François Ged FRANCOIS.GED@EPFL.CH
Clément Hongler CLEMENT.HONGLER@EPFL.CH

Editors: Bin Dong, Qianxiao Li, Lei Wang, Zhi-Qin John Xu

Abstract

We analyze architectural features of Deep Neural Networks (DNNs) using the so-called Neural Tangent Kernel (NTK), which describes the training and generalization of DNNs in the infinite-width setting. In this setting, we show that for fully-connected DNNs, as the depth grows, two regimes appear: *freeze* (or *order*), where the (scaled) NTK converges to a constant, and *chaos*, where it converges to a Kronecker delta. Extreme freeze slows down training while extreme chaos hinders generalization. Using the scaled ReLU as a nonlinearity, we end up in the frozen regime. In contrast, Layer Normalization brings the network into the chaotic regime. We observe a similar effect for Batch Normalization (BN) applied after the last nonlinearity. We uncover the same freeze and chaos modes in Deep Deconvolutional Networks (DC-NNs). Our analysis explains the appearance of so-called checkerboard patterns and border artifacts. Moving the network into the chaotic regime prevents checkerboard patterns; we propose a graph-based parametrization which eliminates border artifacts; finally, we introduce a new layer-dependent learning rate to improve the convergence of DC-NNs. We illustrate our findings on DCGANs: the frozen regime leads to a collapse of the generator to a checkerboard mode, which can be avoided by tuning the nonlinearity to reach the chaotic regime. As a result, we are able to obtain good quality samples for DCGANs without BN.

Keywords: NTK, Freeze, Order, Chaos, Checkerboard patterns, GANs

1. Introduction

The training of Deep Neural Networks (DNN) involves a great variety of architecture choices. It is therefore crucial to find tools to understand their effects and to compare them. For example, Batch Normalization (BN) Ioffe and Szegedy (2015) has proven to be crucial in the training of DNNs but remains ill-understood. While BN was initially introduced to solve the problem of “covariate shift”, recent results Santurkar et al. (2018) suggest an effect on the smoothness of the loss surface. Some alternatives to BN have been proposed Lei Ba et al. (2016); Salimans and Kingma (2016); Klambauer et al. (2017), yet it remains difficult to compare them theoretically. Recent theoretical results Yang et al. (2019) suggest some relation to the transition from “order” (freeze) to “chaos” observed as the depth of the NN goes to infinity Poole et al. (2016); Daniely et al. (2016); Yang and Schoenholz (2017); Schoenholz et al. (2017); Hayou et al. (2019a).

The impact of architecture is very apparent in GANs Goodfellow et al. (2014): their results are heavily affected by the architecture of the generator and discriminator Radford et al. (2015); Zhang et al. (2018); Brock et al. (2018); Karras et al. (2018) and the training may fail without BN Arpit et al. (2016); Xiang and Li (2017).

Recently, there has been important advances Jacot et al. (2018); Du et al. (2019); Allen-Zhu et al. (2018); Chizat and Bach (2018b); Lee et al. (2019) in the understanding of the training of DNNs when the number of neurons in each hidden layer is very large. These results give new tools to study the asymptotic effect of BN. In particular, the Neural Tangent Kernel (NTK) Jacot et al. (2018) illustrates the effect of architecture on the training of DNNs and also describes their loss surface Karakida et al. (2018); Jacot et al. (2020). The NTK can easily be extended to Convolutional Neural Networks (CNNs) and other architectures Yang (2019); Arora et al. (2019), hence allowing comparison. Since the first apparition of this work on arxiv, the freeze/chaos regimes for the NTK has been further observed or studied in Hayou et al. (2019c,b); Xiao et al. (2020); Huang et al. (2020); Buchanan et al. (2021); Wang et al. (2021). To stay consistent with the literature, we will henceforth use the term *order* in place of *freeze*.

1.1. Our Contributions

In Section 3, we study fully-connected deep neural networks of infinite width as the depth L increases. Using a characteristic value $r_{\sigma,\beta}$ (for the non-linearity σ and the amount of bias β), we identify two regimes:

- In the **Ordered regime** (when $r_{\sigma,\beta} < 1$) the NTK approaches a constant kernel, leading to an ill-conditioned kernel Gram matrix and a very narrow valley around the global minimum, hence hurting convergence of the network.

- In the **Chaotic regime** (when $r_{\sigma,\beta} > 1$) the NTK approaches a Kronecker delta kernel, leading to an identity kernel Gram matrix and wide valley around the global minimum, leading to fast convergence but conversely hurting generalization.

For very large depths only critical networks ($r_{\sigma,\beta} = 1$) can be trained successfully Hayou et al. (2019c,b); Xiao et al. (2020). Outside of this large depth regime, the characteristic value plays a similar role to the lengthscale parameters in traditional kernel methods, depending on the application different values of $r_{\sigma,\beta}$ may be optimal. Therefore we discuss in Section 4 how $r_{\sigma,\beta}$ can be changed. A network can be pushed towards the ordered regime by increasing the amount of bias β. Unfortunately even for $\beta = 0$ the network can remain in the ordered regime: to move to the chaotic regime, we show that one can use normalization. We study three types of normalizations and show their 'chaotic' properties:

- We introduce **Nonlinearity Normalization**, which modifies the non-linearity $\sigma(x) \mapsto \frac{\sigma(x)-b}{v}$ to normalize it over random Gaussian inputs. With a normalized nonlinearity, the characteristic value $r_{\sigma,\beta}$ can always reach the chaotic region for small enough β.

- We show that in the infinite width limit, **Layer Normalization** has no effect on training when applied before the nonlinearity and is equivalent to Nonlinearity Normalization when applied after the nonlinearity: in the latter case, the network can therefore reach the chaotic regime.

- We show that **Batch Normalization** at the last layer of the network controls the intensity of the constant mode of the kernel Gram matrix which otherwise dominates in the ordered regime, hence avoiding the slow convergence related to the ordered phase.

Finally in Section 5.2, we conduct a similar analysis on deconvolutional networks, to understand problems of mode collapse in Generative Adversarial Networks (GANs). Mode collapse occurs when a GAN only generates the same image for all inputs. Typically the generated image features checkerboard patterns (high values on regularly spaced pixels) and border artifacts (low intensity pixels close to the border). We show that these problems can be mitigated by modifying the generator:

- To avoid **border artifacts**, we propose a Graph-based parameterization of deconvolutional networks which ensures that the intensity of the NTK is constant over the whole image, preventing the dip in intensity on the border with the traditional parametrization.
- To circumvent the collapse and the **checkerboard patterns** we show that one needs to avoid the ordered regime, where the dominating eigenvectors of the NTK Gram matrix are constant over the inputs of the generator and feature checkerboard patterns. This may explain why normalization is so crucial in practice for the training of GANs, to avoid the ordered regime in the generator.

The traditional technique to avoid Mode Collapse is to use Batch Normalization. Based on our results, we are able to train a simple DC-GAN without Batch Normalization, using a Graph-based parameterization and Nonlinearity Normalization.

1.2. Related Works

The order/chaos transition was first observed for the covariance of the activations in neural networks at initialization Poole et al. (2016); Daniely et al. (2016); Yang and Schoenholz (2017); Schoenholz et al. (2017); Hayou et al. (2019a). The frontier between the two regimes is the same as for the NTK, however the NTK analysis allows one to describe the behavior of the network during training.

Since and simultaneously with the original release of this paper on arxiv, there has been numerous works studying the order/chaos transition for the NTK: the edge of chaos ($r_{\sigma,\beta} = 1$) is studied in more details for both fully-connected and convolutional networks in Hayou et al. (2019c,b); Xiao et al. (2020) and the effect of resnet architecture in Hayou et al. (2019c,b); Huang et al. (2020). To our knowledge, only our paper shows the chaotic effect of normalization and the order/chaos transition in deconvolutional networks leading to checkerboard patterns. Furthermore, while the aforementioned works conclude that only the edge of chaos is viable for training of very deep networks, we show that for reasonable depths the characteristic value plays a similar role to the lengthscale parameters in traditional kernel methods, and we show that for GANs it is advantageous to have a generator in the chaotic regime.

Our work (as well as the aforementioned order/chaos literature) studies infinitely wide DNNs in the linear or lazy regime, characterized by the NTK staying constant during training, by changing the initialization and/or parametrization of DNNs, one can instead reach the so-called mean-field regime where the NTK evolves in time Rotskoff and Vanden-Eijnden (2018); Chizat and Bach (2018a); Mei et al. (2019); Yang and Hu (2020). To our knowledge, the order/chaos transition in the mean-field regime has not yet been studied.

Finally note that as described in Hanin (2018); Hanin and Nica (2019), the limiting behavior of the NTK can be very different in the limit when both width and depth go to infinity simultaneously than in the finite depth, infinite width limit of Jacot et al. (2018); Du et al. (2019); Allen-Zhu et al. (2018); Lee et al. (2019). This work (and other order/chaos literature) gives finite depth bounds for the infinite width limit, roughly speaking, our work applies to large depths and widths but with a width significantly larger than the depth, while in Hanin (2018); Hanin and Nica (2019) the depth and width are of the same order.

2. Fully-Connected Neural Networks

The first type of architecture we consider are deep Fully-Connected Neural Networks (FC-NNs). An FC-NN $\mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ with nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$ consists of $L + 1$ layers ($L - 1$ hidden layers), respectively containing $n_0, n_1, \dots, n_L$ neurons. The parameters are the connection weight matrices $W^{(\ell)} \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$ and bias vectors $b^{(\ell)} \in \mathbb{R}^{n_{\ell+1}}$ for $\ell = 0, 1, \dots, L - 1$. Following Jacot et al. (2018), the network parameters are aggregated into a single vector $\theta \in \mathbb{R}^P$ and initialized using iid standard Gaussians $\mathcal{N}(0, 1)$. For $\theta \in \mathbb{R}^P$, the DNN network function $f_\theta : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ is defined as $f_\theta(x) = \tilde{\alpha}^{(L)}(x)$, where the activations and preactivations $\alpha^{(\ell)}, \tilde{\alpha}^{(\ell)}$ are recursively constructed using the NTK parametrization: we set $\alpha^{(0)}(x) = x$ and, for $\ell = 0, \dots, L - 1$,

$$\begin{aligned}\tilde{\alpha}^{(\ell+1)}(x) &= \frac{\sqrt{1-\beta^2}}{\sqrt{n_\ell}} W^{(\ell)} \alpha^{(\ell)}(x) + \beta b^{(\ell)} \\ \alpha^{(\ell+1)}(x) &= \sigma\left(\tilde{\alpha}^{(\ell+1)}(x)\right),\end{aligned}$$

where σ is applied entry-wise and $\beta \geq 0$.

Remark 1

The hyperparameter β allows one to balance the relative contributions of the connection weights and of the biases during training; in our numerical experiments, we set $\beta = 0.1$. Note that the variance of the normalized bias $\beta b^{(\ell)}$ at initialization can be tuned by β.

2.1. Neural Tangent Kernel

The NTK Jacot et al. (2018) describes the evolution of $(f_{\theta_t})_{t\geq 0}$ in function space during training. In the FC-NN case, the NTK $\Theta_\theta^{(L)} : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_L \times n_L}$ is defined by

$$\Theta_{\theta,kk'}^{(L)}(x, x') = \sum_{p=1}^{P} \partial_{\theta_p} f_{\theta,k}(x)\, \partial_{\theta_p} f_{\theta,k'}(x').$$

For a dataset $x_1, \dots, x_N \in \mathbb{R}^{n_0}$, we define the *output* vector $Y_\theta = (f_{\theta,k}(x_i))_{ik} \in \mathbb{R}^{Nn_L}$. The DNN is trained by optimizing a cost $C : \mathbb{R}^{n_L N} \to \mathbb{R}$ through gradient descent, defining a flow $\partial_t \theta_t = -\nabla_\theta C(Y_\theta)\big|_{\theta_t}$. The evolution of the output vector Y_θ can be expressed in terms of the NTK Gram Matrix $\tilde{\Theta}_\theta^{(L)} = \left(\Theta_{\theta,km}^{(L)}(x_i, x_j)\right)_{ik,jm} \in \mathbb{R}^{n_L N \times n_L N}$ and gradient $\nabla_Y C(Y_{\theta_t}) \in \mathbb{R}^{n_L N}$:

$$\partial_t Y_{\theta_t} = -\tilde{\Theta}_{\theta_t}^{(L)} \nabla_Y C(Y_{\theta_t}).$$

2.2. Infinite-Width Limit

Following Neal (1996); Cho and Saul (2009); Lee et al. (2018), in the overparametrized regime at initialization, the preactivations $\left(\tilde{\alpha}_i^{(\ell)}\right)_{i=1,\dots,n_\ell}$ are described by iid centered Gaussian processes with covariance kernels $\Sigma^{(\ell)}$ constructed as follows. For a kernel K, set

$$\mathbb{L}_K^g\left(z_0, z_1\right) = \mathbb{E}_{(y_0,y_1)\sim\mathcal{N}\left(0,\left(K(z_i,z_j)\right)_{i,j=0,1}\right)}\left[g\left(y_0\right)g\left(y_1\right)\right].$$

The *activation kernels* $\Sigma^{(\ell)}$ are defined recursively by

$$\Sigma^{(0)}\left(z_0, z_1\right) = \beta^2 + \frac{\left(1-\beta^2\right)}{n_0} z_0^T z_1$$
$$\Sigma^{(\ell+1)}\left(z_0, z_1\right) = \beta^2 + \left(1-\beta^2\right)\mathbb{L}_{\Sigma^{(\ell)}}^{\sigma}\left(z_0, z_1\right).$$

While random at initialization, in the infinite-width-limit, the NTK converges to a deterministic limit, which is moreover constant during training:

Theorem 2 *As $n_1, \ldots, n_{L-1} \to \infty$, for any $z_0, z_1 \in \mathbb{R}^{n_0}$ and any $t \geq 0$, the kernel $\Theta_{\theta_t}^{(L)}\left(z_0, z_1\right)$ converges to $\Theta_\infty^{(L)}\left(z_0, z_1\right) \otimes \mathrm{Id}_{n_L}$, where*

$$\Theta_\infty^{(L)}\left(z_0, z_1\right) = \sum_{\ell=1}^{L}\Sigma^{(\ell)}\left(z_0, z\right)\prod_{l=\ell+1}^{L}\dot{\Sigma}^{(l)}\left(z_0, z_1\right)$$

and $\dot{\Sigma}^{(l)} = (1-\beta^2)\mathbb{L}_{\Sigma^{(l-1)}}^{\dot{\sigma}}$ with $\dot{\sigma}$ denoting the derivative of σ.

We refer to Jacot et al. (2018) for a proof for the sequential limit $n_1 \to \infty, \ldots, n_{L-1} \to \infty$ and Yang (2019); Arora et al. (2019) for the simultaneous limit $\min\left(n_1, \ldots, n_{L-1}\right) \to \infty$. As a consequence, in the infinite-width limit, the dynamics of the labels $Y_{\theta_t,k} \in \mathbb{R}^N$ for each outputs k acquires a simple form in terms of the limiting NTK Gram matrix $\tilde{\Theta}_\infty^{(L)} \in \mathbb{R}^{N\times N}$

$$\partial_t Y_{\theta_t,k} = -\tilde{\Theta}_\infty^{(L)}\nabla_{Y_k}C(Y_{\theta_t}),$$

where the Gram matrix is now fixed.

3. Order and Chaos in FC-NNs

We now investigate the large L behavior of the NTK (in the infinite-width limit), revealing a transition between two phases: “order” and “chaos”. To ensure that the variance of the neurons is constant for all depths ($\Sigma^{(\ell)}(x,x) = 1$) we consider *standardized* nonlinearity, i.e. such that

$$\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma^2\left(x\right)\right] = 1$$

and inputs on the *standard $\sqrt{n_0}$-sphere*[1]

$$\mathbb{S}_{n_0} = \left\{x \in \mathbb{R}^{n_0} : \|x\| = \sqrt{n_0}\right\}.$$

For a standardized σ, the large-depth behavior of the *normalized NTK*

$$\vartheta^{(L)}\left(x, y\right) := \frac{\Theta_\infty^{(L)}\left(x, y\right)}{\sqrt{\Theta_\infty^{(L)}\left(x, x\right)\Theta_\infty^{(L)}\left(y, y\right)}}$$

[1] Note that high dimensional datasets tend to concentrate on hyperspheres: for example in GANs Goodfellow et al. (2014) the inputs of a generator are vectors of iid $\mathcal{N}(0,1)$ entries which concentrate around $\mathbb{S}_{n_0}$ for large dimensions.

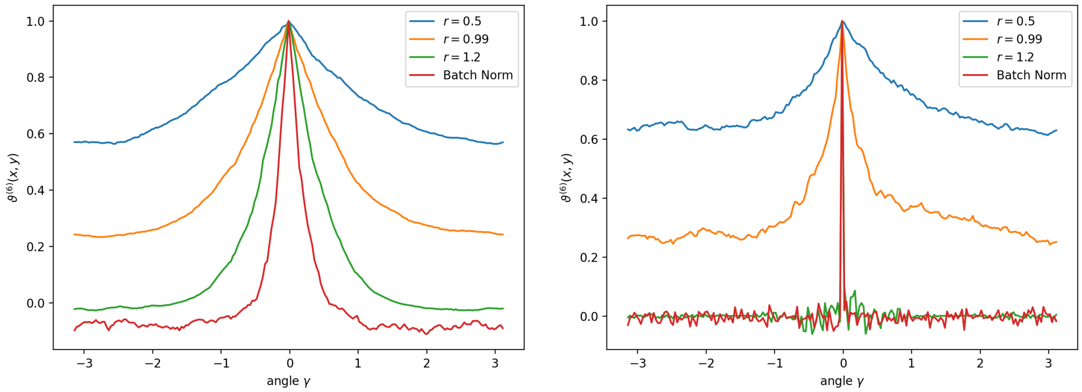

Figure 1: The NTK on the unit circle for four architectures with depth $L = 5$ (top) and $L = 25$ (bottom) are plotted: vanilla ReLU network with $\beta = 1.0$ (blue) and $\beta = 0.1$ (orange), with a normalized ReLU / Layer norm. (green) and with Batch Norm (red). Both networks have width 3000, but the deeper network is further from convergence, leading to more noise.

is determined by the *characteristic value*

$$r_{\sigma,\beta} = (1 - \beta^2)\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\dot{\sigma}^2\left(x\right)\right]. \tag{1}$$

Theorem 3 *Suppose that σ is twice differentiable and standardized.*

Order: *If $r_{\sigma,\beta} < 1$, there exists $C_1 > 0$ such that for $x, y \in \mathbb{S}_{n_0}$,*

$$1 - C_1 L r_{\sigma,\beta}^L \leq \vartheta^{(L)}\left(x, y\right) \leq 1.$$

Chaos: *If $r_{\sigma,\beta} > 1$, for $x \neq \pm y$ in $\mathbb{S}_{n_0}$, there exist $h < 1$ and $C_2 > 0$, such that*

$$\left|\vartheta^{(L)}\left(x, y\right)\right| \leq C_2 h^L.$$

Theorem 3 shows that in the ordered regime, the normalized NTK $\vartheta^{(L)}$ converges to a constant as $L \to \infty$, whereas in the chaotic regime, it converges to a Kronecker δ (taking value 1 on the diagonal, 0 elsewhere). This suggests that the training of deep FC-NN is heavily influenced by the characteristic value: when $r_{\sigma,\beta} < 1$, $\Theta^{(L)}$ becomes constant, thus slowing down the training, whereas when $r_{\sigma,\beta} > 1$, $\Theta^{(L)}$ is concentrates on the diagonal, ensuring fast training, but limiting generalization. To train very deep FC-NNs, it is necessary to lie “on the edge of chaos” $r_{\sigma,\beta} = 1$ Poole et al. (2016); Yang and Schoenholz (2017).

The order/chaos transition can also be related to the “roughness” of the loss around a global minimum. As observed in Jacot et al. (2020) the eigenvalues of the Hessian at convergence are the same as those of the NTK Gram matrix. In the chaotic regime all eigenvalues are close to each other, leading to a “wide valley” around the minimum, on the other hand in the ordered regime, the dominating eigenvalue (corresponding to the constant mode) is much larger than the other eigenvalues, leading to a very “narrow valley”.

3.1. Order and Chaos for ReLU networks

Theorem 3 does not apply directly to the standardized ReLU $\sigma(x) = \sqrt{2}\max(x, 0)$, because it is not differentiable in 0. The characteristic value for the standardized ReLU is $r_{\sigma,\beta} = 1 - \beta^2$ which lies in the ordered regime for $\beta > 0$:

Theorem 4 *With the same notation as in Theorem 3, taking σ to be the standardized ReLU and $\beta > 0$, the NTK is in the ordered regime: there exists a constant C such that $1 - C r_{\sigma,\beta}^{L/2} \leq \vartheta^{(L)}(x, y) \leq 1$.*

We observe two interesting (and potentially beneficial) properties of the standardized ReLU:

1. Its characteristic value $r_{\sigma,\beta} = 1 - \beta^2$ is very close to the 'edge of chaos' for small β and typically with LeCun initialization the variance of the bias at initialization is $\frac{1}{w}$ for w the width, which roughly corresponds to a choice of $\beta = \frac{1}{\sqrt{w}}$.

2. The rate of convergence to the limiting kernel is smaller ($r_{\sigma,\beta}^{L/2}$) for the ReLU than for differentiable nonlinearities ($r_{\sigma,\beta}^{L}$)[2].

These observations suggest that an advantage of the ReLU is that the NTK of ReLU networks converges to its constant limit at a slower rate and may naturally offer a good tradeoff between generalization and training speed.

4. Chaotic effect of normalization

Figure 1 shows that even on the edge of chaos, the NTK may exhibit a strong constant component (i.e. $\vartheta(x, y) > 0.2$ for all x, y) which can lead to a bad conditioning of the Gram matrix governing the infinite-width training behavior. It may be helpful to slightly 'move' the network towards the chaotic regime to reduce this effect. In Figure 1, $r_{\sigma,\beta}$ plays a similar role to that of the lengthscale parameter in classical kernel methods: increasing $r_{\sigma,\beta}$ makes the NTK 'narrower', reducing the correlation length.

From the definition (1) of the characteristic value, we see that increasing the bias pushes the network towards the ordered regime, whereas $r_{\sigma,\beta}$ reaches its highest value $\mathbb{E}\left[\dot{\sigma}^2(x)\right]$ when the bias is 0, which may still be in the ordered regime (or on the edge with the ReLU). We are therefore interested in ways to push the network further towards the chaotic regime.

In this section, we show that Layer Normalization is asymptotically equivalent to Nonlinearity Normalization which entails $r_{\sigma,\beta} > 1$ for β small enough. While Batch normalization cannot be directly interpreted in terms of $r_{\sigma,\beta}$, it is easy to show that it directly controls the constant component of the NTK, which is characteristic of the ordered regime.

4.1. Nonlinearity Normalization

Intuitively, the dominating constant component in ReLU networks is partly a consequence of the ReLU being non-negative: after the first hidden layer, all negative correlations become positive (i.e. $\Sigma^{(1)}(x, y) \geq \beta$ for all x, y, even $x = -y$). One can address this issue thanks to the following. We

[2] Of course the rates of Theorems 3 and 4 may not be tight, but from the proofs in Appendix B.1 one can observe that the rate of $r_{\sigma,\beta}^{L/2}$ appears as a result of the non-differentiability of the ReLU.

shall write Z for a random variable with standard normal distribution. We say that σ is normalized if $\mathbb{E}[\sigma(Z)] = 0$ and $\mathbb{E}[\sigma(Z)^2] = 1$. In particular, if $\sigma \neq \mathrm{id}$, then

$$\overline{\sigma}(\cdot) := \frac{\sigma(\cdot) - \mathbb{E}[\sigma(Z)]}{\sqrt{\mathbb{E}[(\sigma(Z) - \mathbb{E}[\sigma(Z)])^2]}}$$

is normalized. By Poincaré Inequality, after nonlinearity normalization, one can always reach the chaotic regime:

Proposition 5 *If $\sigma \neq \mathrm{id}$ is normalized, then $\mathbb{E}\left[\dot{\sigma}^2\left(Z\right)\right] > 1$ and $r_{\sigma,\beta} > 1$ for $\beta > 0$ small enough.*

4.2. Layer Normalization

Nonlinearity Normalization is closely related to Layer Normalization (LN). We define a normalization layer on any vector $v \in \mathbb{R}^d$ as

$$\mathrm{LN}(v) = \sqrt{d}\frac{v - \bar{v}}{\|v - \bar{v}\|}.$$

for $\bar{v} = \frac{1}{d}\sum_i v_i$. We consider two types of Layer normalization depending on whether we apply the normalization layer before or after the nonlinearity: pre-nonlinearity LN where the activations are changed to $\alpha^{(\ell)}(x) = \sigma(\mathrm{LN}(\tilde{\alpha}^{(\ell)}(x)))$ and post-nonlinearity LN where they are changed to $\alpha^{(\ell)}(x) = \mathrm{LN}(\sigma(\tilde{\alpha}^{(\ell)}(x)))$. Depending on whether Layer Normalization is applied before or after the nonlinearity it has either no effect or is equivalent to Nonlinearity Normalization:

Proposition 6 *Suppose that the inputs belong to $\mathbb{S}_{n_0}$ and that σ is standardized. In the infinite width limit, the network function is the same at initialization and during training:*

- *with or without pre-nonlinearity LN,*
- *with Post-nonlinearity LN or with Nonlinearity Normalization.*

Proof (sketch) At initialization, the normalization parameters $\bar{v}$ and $\|v - \bar{v}\|/\sqrt{d}$ respectively converge to 0 and 1 for pre-nonlinearity LN, and to $\mathbb{E}[\sigma(Z)]$ and $\sqrt{\mathbb{E}[(\sigma(Z) - \mathbb{E}[\sigma(Z)])^2]}$ for post-nonlinearity LN. These values stay asymptotically constant during training because the rate of change of the (pre-)activations is sufficiently small in the linear/lazy regime. ■

4.3. Batch Normalization

For any $N \times d$ matrix of features X leading to a $N \times N$ Gram matrix $K = \frac{1}{d}XX^T$, the Rayleigh quotient $\frac{1}{N}\mathbf{1}^T K\mathbf{1}$ of the constant vector $\mathbf{1}$ measures how big the constant component is. Applying Batch Normalization (BN) at a layer ℓ centers (and standardizes) the activations[3] $\alpha_j^{(\ell)}(x_i)$ over a batch $x_1, ..., x_N$, thus zeroing the constant Rayleigh quotient of the $N \times N$ features Gram matrices $\tilde{\Sigma}^{(\ell)}$ with entries $\tilde{\Sigma}_{ij}^{(\ell)} = \frac{1}{n_\ell}\sum_{k=1}^{n_\ell} \alpha_k^{(\ell)}(x_i)\alpha_k^{(\ell)}(x_j)$. Adding a single BN layer after the last hidden layer controls the constant Rayleigh quotient of the NTK Gram matrix $\tilde{\Theta}^{(L)}$:

Lemma 7 *Consider FC-NN with L layers, with a post-nonlinearity-BN after the last nonlinearity. Then $\frac{1}{N}\mathbf{1}^T\tilde{\Theta}^{(L)}\mathbf{1} = \beta^2$.*

[3] We consider here *post-nonlinearity* BN, it is common to normalize the pre-activations $\tilde{\alpha}^{(\ell)}$ instead.

In contrast, for a network in the extreme ordered regime, i.e. such that $\Theta^{(L)}(x,y) \approx c$ for some constant $c > 0$, the constant Rayleigh quotient scales as $\frac{1}{N}\mathbf{1}^T\tilde{\Theta}^{(L)}\mathbf{1} \approx cN$. The analysis of BN presented in Karakida et al. (2019) is also closely related to this phenomenon.

The chaotic effect of Batch Normalization can also be observed in Figure 1 where the NTK with Nonlinearity and Batch Normalization have a similar behavior.

5. Graph-based Neural Networks and Generative Adversarial Networks

As for FC-NNs, we will show in this section that deconvolutional networks (defined below) exhibit a similar order/chaos transition. Thanks to this analysis, we will see how border artifacts and checkerboard patterns can be avoided by adapting the parametrization and the learning rates of generative adversarial networks (GANs). We first introduce a slightly more general formalism in Section 5.1 and then state our results in Section 5.2.

5.1. Graph-based Neural Networks

In this section, we introduce Graph-based neural networks (GB-NNs) and write deconvolutional networks as a special case. We then describe the infinite width limit of the NTK of GB-NNs.

5.1.1. DEFINITION

In GB-NNs, as in a convolutional neural network, each neuron is indexed by its layer ℓ, its channel $i \in \{1,...,n_\ell\}$ and its location (e.g. the pixel on the image). The position p of a neuron determines its connections with the neurons of the previous and subsequent layers. Furthermore certain connections are shared, i.e. they evolve together. We abstract these concepts in the following manner:

For each layer $\ell = 0,...,L$, the neurons are indexed by a position $p \in I_\ell$ and a channel $i = 1,...,n_\ell$. The sets of positions I_ℓ can be any set, in particular any subset of $\mathbb{Z}^D$. Each position $p \in I_{\ell+1}$ has a set of parents $P(p) \subset I_\ell$ which are neurons of the previous layer connected to p. The connections from the parent (q,ℓ) to the position $(p,\ell+1)$ are encoded in an $n_\ell \times n_{\ell+1}$ weight matrix $W^{(\ell,q\to p)}$. Finally two connections $q \to p$ and $q' \to p'$ can be shared, setting the corresponding matrices to be equal $W^{(\ell,q\to p)} = W^{(\ell,q'\to p')}$.

The inputs of the network x are vectors in $(\mathbb{R}^{n_0})^{I_0}$, for example for colour images of width w and height h, we have $n_0 = 3$ and $I_0 = \{1,...,w\} \times \{1,...,h\} \subset \mathbb{Z}^2$. The activations and preactivations $\alpha^{(\ell)}, \tilde{\alpha}^{(\ell)} \in (\mathbb{R}^{n_\ell})^{I_\ell}$ are constructed recursively using the *graph-based parametrization* that we now introduce: we set $\alpha^{(0,p)}(x) = x^{(p)}$ and for $\ell = 0,\dots,L-1$ and any position $p \in I_{\ell+1}$,

$$\tilde{\alpha}^{(\ell+1,p)}(x) = \beta b^{(\ell)} + \frac{\sqrt{1-\beta^2}}{\sqrt{|P(p)|\,n_\ell}} \sum_{q\in P(p)} W^{(\ell,q\to p)}\alpha^{(\ell,q)}(x) \tag{2}$$

$$\alpha^{(\ell+1,p)}(x) = \sigma\left(\tilde{\alpha}^{(\ell+1,p)}(x)\right)$$

where σ is applied entry-wise, $\beta \geq 0$ and $|P(p)|$ is the cardinality of $P(p)$.

Remark. Note that normalizing according to the number of parents is similar to a common normalization in the context of *graph neural networks* Hua; Li et al. (2020); Sim; Sabanayagam et al. (2022). Graph neural networks deal with graph-structured data by working on a fixed graph, iteratively updating the values of its nodes (and edges) without changing its shape to make predictions.

One such update can be an averaging over the neighbors' values, hence normalizing by the number of neighbors of a node. A GB-NN makes computations from one layer to the other in a directed fashion, the sizes of the layers and connections between them need not be the same, akin to a convolutional neural network. This is in the same vein as the computation skeleton of Section 4 in Daniely et al. (2016). Even though for some very specific instances a graph neural network can be written as a GB-NN, they are not equivalent in general.

Assumption. Henceforth, we will only consider GB-NNs that enjoy the following property: shared weights do not lead to the same neuron, that is, at any layer $\ell+1$, for all neuron $p \in I_{\ell+1}$ and all $q, q' \in P(p)$, the weight matrices $W^{(\ell,q\to p)}, W^{(\ell,q'\to p)}$ are not shared.

If this assumption is not fulfilled, this may alter some of the forthcoming results. However, it is satisfied for typical architectures in the literature, and in particular it holds for deconvolutional networks described below.

5.1.2. DECONVOLUTIONAL NETWORKS

Deconvolutional networks (DC-NNs) in dimension D can be seen as a special case of GB-NNs. We first consider borderless DC-NNs, i.e. the set of positions are $I_\ell = \mathbb{Z}^D$ for all layers ℓ. Given window dimensions $(w_1, ..., w_D)$ and strides $(s_1, ..., s_D)$, the set of parents of $p \in I_{\ell+1}$ is the hyperrectangle $P(p) = \{\lfloor p_1/s_1 \rfloor + 1, ..., \lfloor p_1/s_1 \rfloor + w_1\} \times \cdots \times \{\lfloor p_D/s_D \rfloor + 1, ..., \lfloor p_D/s_D \rfloor + w_D\} \subset \mathbb{Z}^D$. Two connections $q \to p$ and $q' \to p'$ are shared if $s_d \mid p_d - p'_d$ (i.e. s_d is a divisor of $p_d - p'_d$) and $q_d - q'_d = \frac{p_d - p'_d}{s_d}$ for all $d = 1, ..., D$. This definition can easily be extended to any other choices of position sets $I_\ell \subset \mathbb{Z}^D$ (for example hyperrectangles) by considering $P(p) \cap I_\ell$ in place of $P(p)$ as parents of p.

5.1.3. NEURAL TANGENT KERNEL

As for FC-NNs , in the infinite width limit (when $n_1, ..., n_{L-1} \to \infty$) the preactivations $\tilde{\alpha}_i^{(\ell,p)}(x)$ converge to Gaussian processes with covariance

$$Cov\left(\tilde{\alpha}_i^{(\ell+1,p)}(x), \tilde{\alpha}_j^{(\ell+1,q)}(y)\right) = \delta_{ij}\Sigma^{(\ell,pq)}(x,y).$$

The behavior of the network during training is described by the NTK

$$\Theta_{ij}^{(\ell,pq)}(x,y) = \sum_{k=1}^{P} \partial_{\theta_k}\tilde{\alpha}_i^{(\ell+1,p)}(x)\partial_{\theta_k}\tilde{\alpha}_j^{(\ell+1,q)}(y).$$

In the Appendix E we prove the convergence $\Theta_{ij}^{(\ell,pq)}(x,y) \to \delta_{ij}\Theta_\infty^{(\ell,pq)}(x,y)$ of the NTK for the sequential limit $n_1, \cdots, n_{L-1} \to \infty$ and give formulas for the limiting kernels $\Sigma^{(\ell,pq)}(x,y)$ and $\Theta_\infty^{(\ell,pq)}(x,y)$. The simultaneous limit yields the same formulas.

5.2. Order/Chaos transition, Border Artifacts and Checkerboard Patterns

In the context of convolutional networks, in particular GANs, the order/chaos transition sheds light on some interesting phenomena: a common problem in GAN training is the so-called 'mode collapse', where the generator converges to a constant function, hence generating a single image instead of a variety of images. This problem is closely related to the fact that the constant mode of the NTK Gram matrix dominates, and indeed the problem of mode collapse is most prominent in the ordered regime (Figure 2), while normalization techniques (leading to a chaotic network) mitigate this problem.

Other typical problems related to GANs are the appearance of checkerboard patterns and border artifacts in the generated images. Checkerboard patterns occur when regularly spaced pixels are highly correlated, as in the top-right generated pictures in Figure 2. Our analysis of checkerboard patterns is a NTK-based theoretical explanation of Odena et al. (2016), in which checkerboard patterns in deconvolutional network are described. Border artifacts happen when pixels close to the border are visually distinct from those in the middle of the image, for example the border pixels will sometimes be darker (see top-right image in Figure 1).

Our goal is therefore to use the NTK to explain the appearance of border artifacts and checkerboard patterns in generated images. We show that the border artifacts issue can be solved by a change of parametrization and, after establishing the order/chaos transition for DC-NNs, that the checkerboard patterns occur in the ordered regime, and can hence be avoided by adding normalization and using layer-wise learning rates. With these changes we are able to train GANs on CelebA dataset without Batch Normalization.

5.2.1. Border Effects

A very important element of the graph-based parametrization proposed in Section 5.1.1 is the factors $1/\sqrt{|P(p)|\, n_\ell}$ in the definition of the preactivation (Equation 2): we scale the contribution of the previous layer according to the number of neurons $|P(p)|\, n_\ell$ (i.e. n_ℓ channels for each of the$|P(p)|$ positions) which are fed into the neuron. For inputs $x \in \mathbb{S}_{n_0}^{I_0}$ (i.e. such that $x^{(p)} \in \mathbb{S}_{n_0}$ for all p), these factors ensure that the limiting variance $\Sigma^{(\ell,pp)}\left(x,x\right)$ of $\tilde{\alpha}_i^{(\ell,p)}(x)$ at initialization is the same for all p:

Proposition 8 *For GB-NNs with the graph-based parametrization, $\Sigma^{(\ell,pp)}\left(x,x\right)$ and $\Theta_\infty^{(\ell,pp)}\left(x,x\right)$ do not depend neither on $p \in I_\ell$ nor on $x \in \mathbb{S}_{n_0}^{I_0}$.*

These factors are usually not present and to compensate, the variance of the weights at initialization is reduced. In convolutional networks with LeCun initialization, the standard deviation of the weights at initialization is set to $\frac{1}{\sqrt{whn_\ell}}$ for w and h the width and height of the window of convolution, which has roughly the effect of replacing the $\frac{1}{\sqrt{|P(p)|n_\ell}}$ factors by $\frac{1}{\sqrt{whn_\ell}}$. However whn_ℓ is the maximal number of parents that a neuron can have, it is typically attained at positions p in the middle of the image. Positions p on the border of the image have less parents hence leading to a smaller contribution of the previous layer. This leads both kernels $\Sigma^{(\ell,pp)}(x,x)$ and $\Theta^{(\ell,pp)}(x,x)$ to have lower intensity for $p \in I_\ell$ on the border (see Appendix G for an example when $I_\ell = \mathbb{N}$, i.e. when there is one border pixel), leading to border artifacts as seen in Figure 2.

5.2.2. Order, Chaos and Checkerboard Patterns

Large depths deconvolutional networks exhibit a similar Order/Chaos transition as that of FC-NNs, the values of the limiting kernel at different positions $\Theta^{(L,pq)}$ is especially interesting.

For GB-NNs, the value of an output neuron at a position $p \in I_L$ only depends on the inputs which are ancestors of p, i.e. all positions $q \in I_0$ such that there is a chain of connections from q

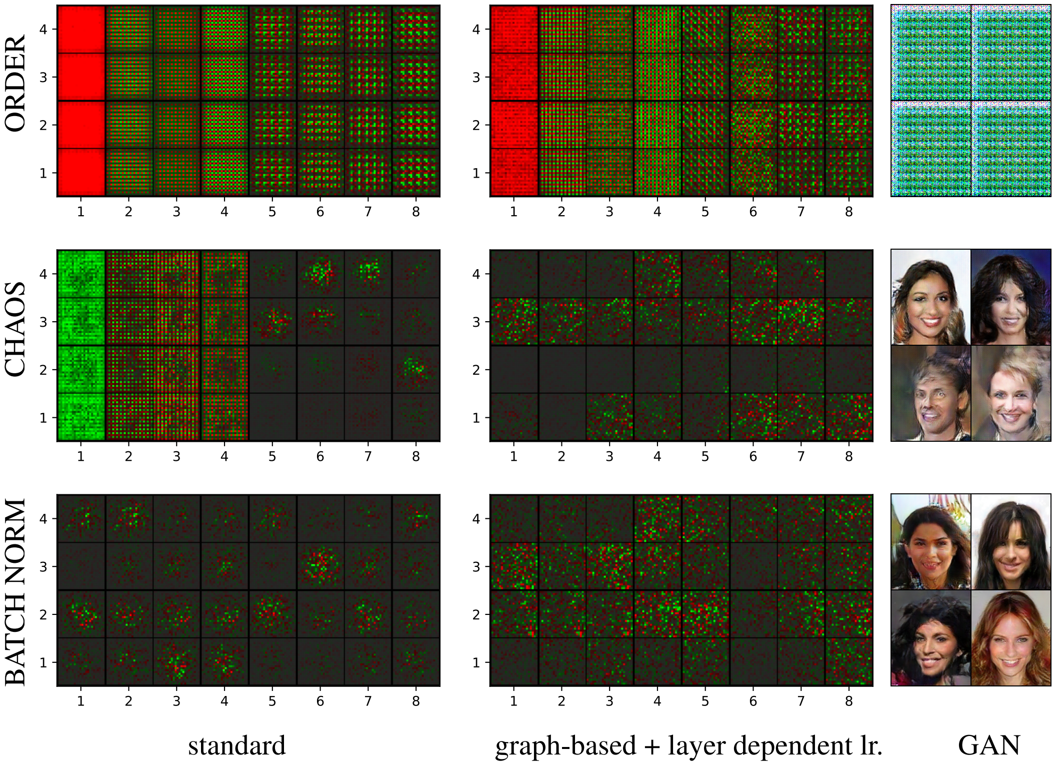

Figure 2: The left and middle columns represent the first 8 eigenvectors of the NTK Gram matrix of a DC-NN (L=3) on 4 inputs. (left) without the Graph-Based Parametrization (GBP) and the Layer-Dependent Learning Rate (LDLR); (middle) with GBP and LDLR. The right column represents the results of a GAN on CelebA with GBP and LDLR. Each line correspond to a choice of nonlinearity/normalization for the generator: (top) ReLU, (middle) normalized ReLU and (bottom) ReLU with Batch Normalization.

to p. For the same reason , the NTK $\Theta^{(L,pp')}(x,y)$ only depends on the values $x_q, y_{q'}$ for $q, q' \in I_0$ ancestors of p and p' respectively.

For a stride $s \in \{2,3,\dots\}^d$, we denote the s-valuation $v_s(n)$ of $n \in \mathbb{Z}^d$ as the largest $k \in \{0,1,2,\dots\}$ such that $s_i^k \mid n_i$ for all $i = 1,...,d$. The behaviour of the NTK $\Theta^{(L)}_{p,p'}(x,y)$ depends on the s-valuation of the difference of the two output positions. If $v_s(p'-p)$ is strictly smaller than L, the NTK $\Theta^{(L,pp')}(x,y)$ converges to a constant in the infinite-width limit for any $x, y \in \mathbb{S}^{I_0}_{n_0}$. Again the characteristic value $r_{\sigma,\beta}$ plays a central role in the behavior of the large-depth limit. In this context, we define the rescaled NTK as $\vartheta^{(L,pp')}(x,y) = \Theta^{(L,pp')}(x,y)/\sqrt{\Theta^{(L,pp)}(x,x)\Theta^{(L,p'p')}(y,y)}$ (note that the denominator actually does not depend on p, p', x nor y by Proposition 8)

Theorem 9 *Consider a borderless DC-NN with position sets $I_\ell = \mathbb{Z}^D$ for all layers ℓ, upsampling stride $s \in \{2,3,\dots\}^D$ and window sizes $w \in \{1,2,3,\dots\}^D$. For a standardized twice differentiable σ, there exist constants $C_1, C_2 > 0$, such that the following holds: for $x, y \in \mathbb{S}^{I_0}_{n_0}$, and any positions $p, p' \in I_L$, we have*

Order: *When* $r_{\sigma,\beta} < 1$*, taking* $v = \min\left(v_s\left(p - p'\right), L - 1\right)$*, we have*

$$\frac{1 - r_{\sigma,\beta}^{v+1}}{1 - r_{\sigma,\beta}^{L}} - C_1(v+1)r_{\sigma,\beta}^{v} \leq \vartheta^{(L,pp')}\left(x,y\right) \leq \frac{1 - r_{\sigma,\beta}^{v+1}}{1 - r_{\sigma,\beta}^{L}}.$$

Chaos: *When* $r_{\sigma,\beta} > 1$*, if either* $v_s\left(p - p'\right) < L$ *or if there exists* $c < 1$ *such that for all positions* $q \in I_0$ *which are ancestors of* p*,* $\left|x_q^T y_{q+\frac{p'-p}{s^L}}\right| < c$*, then there exists* $h < 1$ *such that*

$$\left|\vartheta^{(L,pp')}\left(x,y\right)\right| \leq C_2 h^L.$$

This theorem suggests that in the order regime, the correlations between differing positions p and p' increase with $v_s\left(p - p'\right)$, which is a strong feature of checkerboard patterns Odena et al. (2016). These artifacts typically appear in images generated by DC-NNs. The form of the NTK also suggests a strong affinity to these checkerboard patterns: they should dominate the NTK spectral decomposition. This is shown in Figure 2 where the eigenvectors of the NTK Gram matrix for a DC-NN are computed.

In the chaotic regime, the normalized NTK converges to a “scaled translation invariant” Kronecker delta. For two output positions p and $p' = p + ks^L$ we associate the two regions ω and $\omega' = \omega + k$ of the input space which are connected to p and p'. Then $\vartheta^{\left(L,p,p+ks^L\right)}\left(x,y\right)$ is one if the patch $y_{\omega'}$ is a k translation of x_ω and approximately zero otherwise.

5.2.3. LAYER-DEPENDENT LEARNING RATE

The NTK is the sum $\Theta^{(L)} = \sum_\ell \Theta_{W^{(\ell)}}^{(L)} + \Theta_{b^{(\ell)}}^{(L)}$ over the contributions of the weights $\Theta_{W^{(\ell)}}^{(L,pq)}(x,y) = \sum_{ij} \partial_{W_{ij}^{(\ell)}} f_{\theta,p}(x) \partial_{W_{ij}^{(\ell)}} f_{\theta,q}(y)$ and biases $\Theta_{b^{(\ell)}}^{(L,pq)}(x,y) = \sum_j \partial_{b_j^{(\ell)}} f_{\theta,p}(x) \partial_{b_j^{(\ell)}} f_{\theta,q}(y)$. At the ℓ-th layer, the weights and biases can only contribute to checkerboard patterns of degree $v = L - \ell$ and $v = L - \ell - 1$, i.e. patterns with periods $s^{L-\ell}$ and $s^{L-\ell-1}$ respectively, in the following sense:

Proposition 10 *In a DC-NN with stride* $s \in \{2, 3, ...\}^d$*, the infinite width limiting NTK is such that* $\Theta_{W^{(\ell)}}^{(L,pp')}(x,y) = 0$ *if* $s^{L-\ell} \nmid p' - p$ *and* $\Theta_{b^{(\ell)}}^{(L,pp')}(x,y) = 0$ *if* $s^{L-\ell-1} \nmid p' - p$*.*

This suggests that the supports of $\Theta_{\infty,W^{(\ell)}}^{(L)}$ and $\Theta_{\infty,b^{(\ell)}}^{(L)}$ increase exponentially with ℓ, giving more importance to the last layers during training. In the classical parametrization, the balance is restored by letting the number of channels n_ℓ decrease with depth Radford et al. (2015). In the graph-based parametrization, the limiting NTK is not affected by the ratios $\frac{n_\ell}{n_k}$. To achieve the same effect, we divide the learning rate of the weights and bias of the ℓ-th layer by $S^{\frac{\ell}{2}}$ and $S^{\frac{(\ell+1)}{2}}$ respectively, where $S = \prod_i s_i$ is the product of the strides. Together with the graph-based parametrization and the normalization of the nonlinearity (in order to lie in the chaotic regime) this rescaling of the learning rate removes both border and checkerboard artifacts in Figure 2. Technical details are provided in Appendix F.2.

6. Conclusion

This article shows how the NTK can be used theoretically to understand the effect of architecture choices (such as decreasing the number of channels or batch normalization) on the training of DNNs.

In the context of GB-NNs we show that the "order" regime yields a strong affinity to constant modes and checkerboard artifacts: this slows down training and can contribute to a mode collapse of the DC-NN generator of GANs. We introduce simple modifications to solve these problems: the effectiveness of normalizing the nonlinearity, a graph-based parametrization and a layer-dependent learning rates is shown both theoretically and numerically.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A Convergence Theory for Deep Learning via Over-Parameterization. *CoRR*, abs/1811.03962, 2018. URL http://arxiv.org/abs/1811.03962.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019.

Devansh Arpit, Yingbo Zhou, Bhargava Kota, and Venu Govindaraju. Normalization propagation: A parametric technique for removing internal covariate shift in deep networks. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1168–1176, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/arpitb16.html.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Sam Buchanan, Dar Gilboa, and John Wright. Deep networks and the multiple manifold problem. *ArXiv*, abs/2008.11245, 2021.

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 31*, pages 3040–3050. Curran Associates, Inc., 2018a. URL http://papers.nips.cc/paper/7567-on-the-global-convergence-of-gradient-descent-for-over-parameterized-models-using-optimal-transport.pdf.

Lenaic Chizat and Francis Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018b.

Youngmin Cho and Lawrence K. Saul. Kernel Methods for Deep Learning. In *Advances in Neural Information Processing Systems 22*, pages 342–350. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3628-kernel-methods-for-deep-learning.pdf.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2253–2261. Curran Associates, Inc., 2016.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=S1eK3i09YQ`.

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*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL `http://proceedings.mlr.press/v9/glorot10a.html`.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, pages 2672–2680, jun 2014. URL `http://arxiv.org/abs/1406.2661`.

Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? *arXiv preprint arXiv:1801.03744*, 2018.

Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel, 2019.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In *International Conference on Machine Learning*, pages 2672–2680. PMLR, 2019a.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. Mean-field behaviour of neural tangent kernel for deep neural networks. *arXiv preprint arXiv:1905.13654*, 2019b.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. Training dynamics of deep networks using stochastic gradient descent via neural tangent kernel. *arXiv preprint arXiv:1905.13654*, 2019c.

Kaixuan Huang, Yuqing Wang, Molei Tao, and Tuo Zhao. Why do deep residual networks generalize better than deep feedforward networks?—a neural tangent kernel perspective. *Advances in Neural Information Processing Systems*, 33, 2020.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. URL `http://arxiv.org/abs/1502.03167`.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems 31*, pages 8580–8589. Curran Associates, Inc., 2018. URL `http://papers.nips.cc/paper/8076-neural-tangent-kernel-convergence-and-generalization-in-neural-networks.pdf`.

Arthur Jacot, Franck Gabriel, and Clement Hongler. The asymptotic spectrum of the hessian of dnn throughout training. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SkgscaNYPS`.

Ryo Karakida, Shotaro Akaho, and Shun-Ichi Amari. Universal Statistics of Fisher Information in Deep Neural Networks: Mean Field Approach. jun 2018. URL http://arxiv.org/abs/1806.01316.

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*, pages 6403–6413, 2019.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 971–980. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6698-self-normalizing-neural-networks.pdf.

Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

Jae Hoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep Neural Networks as Gaussian Processes. *ICLR*, 2018.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in neural information processing systems*, pages 8572–8583, 2019.

J. Lei Ba, J. R. Kiros, and G. E. Hinton. Layer Normalization. *arXiv e-prints*, July 2016.

Guohao Li, Chenxin Xiong, Ali K. Thabet, and Bernard Ghanem. Deepergcn: All you need to train deeper gcns. *ArXiv*, abs/2006.07739, 2020.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *arXiv preprint arXiv:1902.06015*, 2019.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996. ISBN 0387947248.

Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

Daniel S Park, Samuel L Smith, Jascha Sohl-dickstein, and Quoc V Le. Optimal SGD Hyperparameters for Fully Connected Networks. 2018.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3360–3368. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6322-exponential-expressivity-in-deep-neural-networks-through-transient-chaos.pdf.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems 31*, pages 7146–7155. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7945-parameters-as-interacting-particles-long-time-convergence-and-asymptotic-error-scaling-of-neural-networks.pdf.

Mahalakshmi Sabanayagam, Pascal Esser, and Debarghya Ghoshdastidar. New insights into graph convolutional networks using neural tangent kernels. 2022. URL https://arxiv.org/abs/2110.04060.

Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 901–909. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6114-weight-normalization-a-simple-reparameterization-to-accelerate-training-of-deep-neural-networks.pdf.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2483–2493. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7515-how-does-batch-normalization-help-optimization.pdf.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. 2017. URL https://openreview.net/pdf?id=H1W1UN9gg.

Tingran Wang, Sam Buchanan, Dar Gilboa, and John Wright. Deep networks provably classify data on curves. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=SFFFiGtdAt.

Sitao Xiang and Hao Li. On the effects of batch and weight normalization in generative adversarial networks. *arXiv preprint arXiv:1704.03971*, 2017.

Lechao Xiao, Jeffrey Pennington, and Samuel Schoenholz. Disentangling trainability and generalization in deep neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10462–10472. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/xiao20b.html.

Greg Yang. Scaling Limits of Wide Neural Networks with Weight Sharing: Gaussian Process Behavior, Gradient Independence, and Neural Tangent Kernel Derivation. *arXiv e-prints*, art. arXiv:1902.04760, Feb 2019.

Greg Yang and Edward J. Hu. Feature learning in infinite-width neural networks, 2020.

Greg Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 7103–7114. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6879-mean-field-residual-networks-on-the-edge-of-chaos.pdf.

Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. A mean field theory of batch normalization. *CoRR*, abs/1902.08129, 2019. URL http://arxiv.org/abs/1902.08129.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

Organisation of the appendix.

- Appendix A: More explanations are provided regarding the parametrization of FC-NNs we considered, such as how it relates to other standard parametrization.
- Appendix B: The proofs of the claims of Section 3 about the order-chaos transition for FCNNs are given, namely Theorem 3 and Theorem 4.
- Appendix C: The proofs of the claims of Section 4 about the effects of layer normalization and nonlinearity normalization are given, stated in Proposition 5 and Proposition 6 in the main text and proven respectively in Appendix C.1 and C.2.
- Appendix D: The proof of Lemma 7 about batch normalization is given.
- Appendix E: The expression of the NTK in the infinite width limit is given and derived for the GB-NNs defined in Section 5.1.1.
- Appendix F: The proof of Theorem 9 establishing the order-chaos transition for borderless DCNNs is given in Section F.1. In Section F.2. we provide more details on the effect of our layer-dependent learning rates introduced in Section 5.2.3 on the NTK, by stating Proposition 21 and making its proof, which explains how it allows to avoid checkerboard patterns in the order regime.
- Appendix G: We provide more details on border effects. In Proposition 22 we exhibit border effects on the activation kernels and the limiting NTK for a special case of a DCNN with NTK-parametrization, whereas it is not the case anymore when using the graph-based parametrization, as stated in the main text Proposition 8, whose proof is provided in this Appendix.
- Appendix H: The proof of Proposition 10 about the contribution of a given layer on checkerboard patterns is given.

Appendix A. Choice of Parametrization

The NTK parametrization for FC-NNs introduced in Section 2 differs slightly from the one commonly used, yet it ensures that the training is consistent as the size of the layers grows. In the standard parametrization, for $\ell = 0..L-1$, the activations are defined by

$$\begin{aligned}\alpha^{(0)}(x) &= x\\ \tilde{\alpha}^{(\ell+1)}(x) &= W^{(\ell)}\alpha^{(\ell)}(x) + b^{(\ell)}\\ \alpha^{(\ell+1)}(x) &= \sigma\left(\tilde{\alpha}^{(\ell+1)}(x)\right).\end{aligned}$$

Let denote by g_θ the output function of the FC-NN thus parametrized, where θ is the concise notation for the vector of free parameters of the FC-NN, and f_θ that of the FC-NN with NTK parametrization. Note the absence of $\frac{1}{\sqrt{n_\ell}}$ in comparison to the NTK parametrization. With LeCun/He initialization LeCun et al. (2012), the parameters $W^{(\ell)}$ have standard deviation $\frac{1}{\sqrt{n_\ell}}$ (or $\frac{\sqrt{2}}{\sqrt{n_\ell}}$ for the ReLU but this does not change the general analysis). Using this initialization, the activations stay stochastically bounded as the widths of the FC-NN get large. In the forward pass, there is almost no difference between the two parametrizations and for each choice of parameters θ, we can scale down the connection weights by $\frac{\sqrt{1-\beta^2}}{\sqrt{n_\ell}}$ and the bias weights by β to obtain a new set of parameters $\hat{\theta}$ such that

$$f_\theta = g_{\hat{\theta}}.$$

The two parametrizations will exhibit a difference during backpropagation since:

$$\partial_{W_{ij}^{(\ell)}} g_{\hat{\theta}}(x) = \frac{\sqrt{n_\ell}}{\sqrt{1-\beta^2}} \partial_{W_{ij}^{(\ell)}} f_\theta(x), \qquad \partial_{b_j^{(\ell)}} g_{\hat{\theta}}(x) = \frac{1}{\beta} \partial_{b_j^{(\ell)}} f_\theta(x).$$

The NTK is a sum of products of these derivatives over all parameters:

$$\Theta^{(L)} = \Theta^{(L:W^{(0)})} + \Theta^{(L:b^{(0)})} + \Theta^{(L:W^{(1)})} + \Theta^{(L:b^{(1)})} + \ldots + \Theta^{(L:W^{(L-1)})} + \Theta^{(L:b^{(L-1)})}.$$

With our parametrization, all summands converge to a finite limit, while with the Le Cun or He parameterization we obtain

$$\hat{\Theta}^{(L)} = \frac{n_0}{1-\beta^2} \Theta^{(L:W^{(0)})} + \frac{1}{\beta^2} \Theta^{(L:b^{(0)})} + \ldots + \frac{n_{L-1}}{1-\beta^2} \Theta^{(L:W^{(L-1)})} + \frac{1}{\beta^2} \Theta^{(L:b^{(L-1)})},$$

where some summands, namely the $\left(\frac{n_i}{1-\beta^2}\Theta^{(L:W^{(i)})}\right)_i$, explode in the infinite width limit. One must therefore take a learning rate of order $\frac{1}{\max(n_1,\ldots n_{L-1})}$ Karakida et al. (2018); Park et al. (2018) to obtain a meaningful training dynamics, but in this case the contributions to the NTK of the first layers connections $W^{(0)}$ and the bias of all layers $b^{(\ell)}$ vanish, which implies that training these parameters has less and less effect on the function as the width of the network grows. As a result, the dynamics of the output function during training can still be described by a modified kernel gradient descent: the modified learning rate compensates for the absence of normalization in the usual parametrization.

The NTK parametrization is hence more natural for large networks, as it solves both the problem of having meaningful forward and backward passes, and to avoid tuning the learning rate, which is the problem that sparked multiple alternative initialization strategies in deep learning Glorot and Bengio (2010). Note that in the standard parametrization, the importance of the bias parameters shrinks as the width gets large; this can be implemented in the NTK parametrization by taking a small value for the parameter β.

Appendix B. FC-NN Order and Chaos

In this section, we prove the existence of two regimes,'order' and 'chaos', in FC-NNs. First, we improve some results of Daniely et al. (2016), and study the rate of convergence of the activation kernels as the depth grows to infinity. In a second step, this allows us to characterise the behavior of the NTK for large depth.

Let us consider a standardized differentiable nonlinearity σ, i.e. satisfying $\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma^2(x)\right] = 1$. Recall that the the activation kernels are defined recursively by $\Sigma^{(1)}(x,y) = \frac{1-\beta^2}{n_0}x^T y + \beta^2$ and $\Sigma^{(\ell+1)}(x,y) = (1-\beta^2)\mathbb{L}^\sigma_{\Sigma^{(\ell)}}(x,y) + \beta^2$, where $\mathbb{L}^\sigma_{\Sigma^{(L)}}$ was introduced in Section 2.2. By induction,

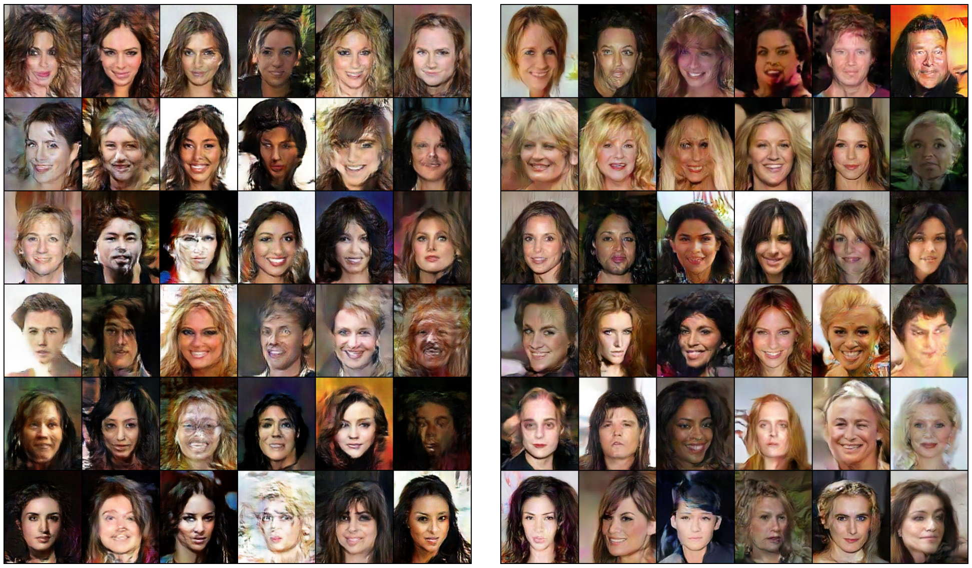

Figure 3: Result of two GANs on CelebA. (Left) with Nonlinearity Normalization and (Right) with Batch Normalization. In both cases the discriminator uses a Normalized ReLU.

for any $x, y \in \mathbb{S}_{n_0}$, $\Sigma^{(\ell+1)}(x,y)$ is uniquely determined by $\rho_{x,y} = \frac{1}{n_0}x^T y$. Defining the two functions $R_\sigma, B_\beta : [-1,1] \to [-1,1]$ by:

$$\begin{aligned} R_\sigma(\rho) &= \mathbb{E}_{v\sim\mathcal{N}\left(0,\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)}\left[\sigma(v_0)\sigma(v_1)\right], \\ B_\beta(\rho) &= \beta^2 + (1-\beta^2)\rho, \end{aligned}$$

one can formulate the activation kernels as an alternate composition of B_β and R_σ:

$$\Sigma^{(\ell)}(x,y) = \left(B_\beta \circ R_\sigma\right)^{\circ \ell - 1} \circ B_\beta\left(\rho_{x,y}\right).$$

In particular, this shows that for any $x, y \in \mathbb{S}_{n_0}$, $\Sigma^{(\ell)}(x,y) \leq 1$. Since the activation kernels are obtained by iterating the same function, we first study the fixed points of the composition $B_\beta \circ R_\sigma : [-1,1] \to [-1,1]$. When σ is a standardized nonlinearity, the function R_σ, named the dual of σ, satisfies the following key properties proven in Daniely et al. (2016):

1. $R_\sigma(1) = 1$,

2. For any $\rho \in (-1,0)$, $R_\sigma(\rho) > \rho$,

3. R_σ is convex in $[0,1)$,

4. $R'_\sigma(1) = \mathbb{E}\left[\dot{\sigma}(x)^2\right]$, where R'_σ denotes the derivative of R_σ,

5. $R'_\sigma = R_{\dot{\sigma}}$.

By definition $B_\beta(1) = 1$, thus 1 is a trivial fixed point: $B_\beta \circ R_\sigma(1) = 1$. This shows that for any $x \in \mathbb{S}_{n_0}$ and any $\ell \geq 1$:

$$\Sigma^{(\ell)}(x, x) = 1.$$

It appears that -1 is also a fixed point of $B_\beta \circ R_\sigma$ if and only if the nonlinearity σ is antisymmetric and $\beta = 0$. From now on, we will focus on the region $(-1, 1)$. From the property 2. of R_σ and since B_β is non decreasing, any non trivial fixed point must lie in $[0, 1)$. Since $B_\beta \circ R_\sigma(0) > 0$, $B_\beta \circ R_\sigma(1) = 1$ and R_σ is convex in $[0, 1)$, there exists a non trivial fixed point of $B_\beta \circ R_\sigma$ if $(B_\beta \circ R_\sigma)'(1) > 1$ whereas if $(B_\beta \circ R_\sigma)'(1) < 1$ there is no fixed point in $(-1, 1)$. This leads to two regimes shown in Daniely et al. (2016), depending on the value of $r_{\sigma,\beta} = \left(1 - \beta^2\right) \mathbb{E}_{x\sim\mathcal{N}(0,1)} \left[\dot{\sigma}^2(x)\right]$:

1. "Order" when $r_{\sigma,\beta} < 1$: $B_\beta \circ R_\sigma$ has a unique fixed point equal to 1 and the activation kernels become constant at an exponential rate,

2. "Chaos" when $r_{\sigma,\beta} > 1$: $B_\beta \circ R_\sigma$ has another fixed point $0 \leq a < 1$ and the activation kernels converge to a kernel equal to 1 if $x = y$ and to a if $x \neq y$ and, if the nonlinearity is antisymmetric and $\beta = 0$, it converges to -1 if and only if $x = -y$.

To establish the existence of the two regimes for the NTK, we need the following bounds on the rate of convergence of $\Sigma^{(\ell)}(x, y)$ in the "order" region and on its values in the "chaos" region:

Lemma 11 *Suppose that σ is a standardized differentiable nonlinearity.*

If $r_{\sigma,\beta} < 1$, then for any $x, y \in \mathbb{S}_{n_0}$,

$$1 \geq \Sigma^{(\ell)}(x, y) \geq 1 - 2r_{\sigma,\beta}^{\ell-1}(1 - \beta^2).$$

If $r_{\sigma,\beta} > 1$, then there exists a fixed point $a \in [0, 1)$ of $B_\beta \circ R_\sigma$ such that for any $x, y \in \mathbb{S}_{n_0}$,

$$\left|\Sigma^{(\ell)}(x, y)\right| \leq \max\left\{\left|\beta^2 + \frac{1 - \beta^2}{n_0} x^T y\right|, a\right\}.$$

Proof Let us denote $r = r_{\sigma,\beta}$ and suppose first that $r < 1$. By Daniely et al. (2016), we know that $R'_\sigma = R_{\dot{\sigma}}$ and $R_{\dot{\sigma}}(\rho) \in \left[-\mathbb{E}\left[\dot{\sigma}(z)^2\right], \mathbb{E}\left[\dot{\sigma}(z)^2\right]\right]$ where $z \sim \mathcal{N}(0, 1)$. From now on, we will omit to specify the distribution asumption on z. The previous equalities and inequalities imply that $R_\sigma(\rho) \geq 1 - \mathbb{E}\left[\dot{\sigma}(v)^2\right](1 - \rho)$, thus we obtain:

$$B_\beta \circ R_\sigma(\rho) \geq \beta^2 + (1 - \beta^2)(1 - \mathbb{E}\left[\dot{\sigma}(z)^2\right](1 - \rho)) = 1 - r(1 - \rho).$$

By definition, we then have $\Sigma^{(\ell)}(x, y) = (B_\beta \circ R_\sigma)^{\circ\ell-1} \circ B_\beta\left(\frac{1}{n_0} x^T y\right) \geq 1 - 2(1 - \beta^2)r^{\ell-1}$. Using the bound $\Sigma^{(\ell)}(x, y) \leq 1$, this proves the first assertion.

When $r > 1$, there exists a fixed point a of $B_\beta \circ R_\sigma$ in $[0, 1)$. By a convexity argument, for any ρ in $[a, 1)$, $a \leq B_\beta \circ R_\sigma(\rho) \leq \rho$ and because $R_\sigma(\rho)$ is increasing in $[0, 1)$, for all $\rho \in [0, a]$, $0 \leq B_\beta \circ R_\sigma(\rho) \leq a$.

For negative ρ, we claim that $|B_\beta \circ R_\sigma(\rho)| \leq B_\beta \circ R_\sigma(|\rho|)$,which entails the second assertion. Since $R_\sigma(\rho) = \sum_{i=0}^\infty b_i \rho^i$ for positive b_is Daniely et al. (2016), and the composition $B_\beta \circ R_\sigma(\rho) = \sum_{i=0}^\infty c_i \rho^i$ for $c_0 = b_0(1 - \beta^2) + \beta^2 \geq 0$ and $c_i = b_i(1 - \beta^2) \geq 0$ when $i > 0$, we have

$$|B_\beta \circ R_\sigma(\rho)| = \left|\sum_{i=0}^\infty c_i \rho^i\right| \leq \sum_{i=0}^\infty c_i |\rho|^i = B_\beta \circ R_\sigma(|\rho|).$$

This leads to the inequality in the chaos regime. ■

Before studying the normalized NTK, let us remark that the NTK on the diagonal (with $x = y$ in $\mathbb{S}_{n_0}$) is equal to:

$$\begin{aligned}\Theta_{\infty}^{(L)}(x,x) &= \sum_{\ell=1}^{L} \Sigma^{(\ell)}(x,x) \prod_{k=\ell+1}^{L} \dot{\Sigma}^{(k)}(x,x) = \sum_{\ell=1}^{L} \left((1-\beta^2)\mathbb{E}\left[\dot{\sigma}(x)^2\right]\right)^{L-\ell} \\ &= \frac{1-r^L}{1-r}.\end{aligned}$$

This shows that in the ordered regime, $\Theta_{\infty}^{(L)}(x,x) \underset{L\to\infty}{\longrightarrow} \frac{1}{1-r}$ and in the chaotic regime $\Theta_{\infty}^{(L)}(x,x)$ grows exponentially. At the transition, $r = 1$ and thus $\Theta_{\infty}^{(L)}(x,x) = L$. Besides, if $x, y \in \mathbb{S}_{n_0}$, using the Cauchy-Schwarz inequality, for any ℓ , $\left|\Sigma^{(\ell)}(x,y)\right| \leq \left|\Sigma^{(\ell)}(x,x)\right|$ and $\left|\dot{\Sigma}^{(\ell+1)}(x,y)\right| \leq \left|\dot{\Sigma}^{(\ell+1)}(x,x)\right|$. This implies the following inequality: $\Theta_{\infty}^{(L)}(x,y) \leq \Theta_{\infty}^{(L)}(x,x)$.

We now study the normalized NTK $\vartheta_L\left(x,y\right) = \frac{\Theta_{\infty}^{(L)}(x,y)}{\Theta_{\infty}^{(L)}(x,x)} \leq 1$.

Theorem 12 (Theorem 3 in the main) *Suppose that σ is twice differentiable and standardized.*

If $r < 1$, we are in the ordered regime: there exists C_1 such that for $x, y \in \mathbb{S}_{n_0}$,

$$1 - C_1 L r^L \leq \vartheta^{(L)}\left(x,y\right) \leq 1.$$

If $r > 1$, we are in the chaotic regime: for $x \neq y$ in $\mathbb{S}_{n_0}$, there exist $s < 1$ and C_2, such that

$$\left|\vartheta^{(L)}\left(x,y\right)\right| \leq C_2 s^L.$$

Proof First, let us suppose that $r < 1$. Recall that the NTK is defined as

$$\Theta_{\infty}^{(L)}(x,y) = \sum_{\ell=1}^{L} \Sigma^{(\ell)}(x,y)\dot{\Sigma}^{(\ell+1)}(x,y)\ldots\dot{\Sigma}^{(L)}(x,y).$$

Several times in the appendix, we will use the following fact: for any $a_1, \cdots, a_k \in (0,1)$, we have

$$\prod_{i=1}^{k}(1-a_i) \geq 1 - \sum_{i=1}^{k} a_i. \tag{3}$$

For all $\ell = 1..L$, $\Sigma^{(\ell)}(x,y) \leq \Sigma^{(\ell)}(x,x) = 1$ and $\dot{\Sigma}^{(\ell)}(x,y) \leq \dot{\Sigma}^{(\ell)}(x,x) = r$. Writing $\Sigma^{(\ell)}(x,y) = 1 - \epsilon^{(\ell)}$ and $\dot{\Sigma}^{(\ell)}(x,y) = r - \dot{\epsilon}^{(\ell)}$ for $\epsilon^{(\ell)}, \dot{\epsilon}^{(\ell)} \geq 0$, we have that

$$\begin{aligned}\Theta_{\infty}^{(L)}(x,y) &= \sum_{\ell=1}^{L}\left(1-\epsilon^{(\ell)}\right)\prod_{k=\ell+1}^{L}\left(r-\dot{\epsilon}^{(\ell)}\right) \\ &\geq \sum_{\ell=1}^{L} r^{L-\ell} - r^{L-\ell}\epsilon^{(\ell)} - \sum_{k=\ell+1}^{L} r^{L-\ell-1}\dot{\epsilon}^{(\ell)},\end{aligned}$$

by (3). Using the bound of Lemma 11 and the fact that for any $x, y \in \mathbb{S}_{n_0}$, $\dot{\Sigma}^{(\ell)}(x,y) = (1-\beta^2)R_{\dot{\sigma}}(\Sigma^{(\ell-1)}(x,y)) \geq r - \psi\epsilon^{(\ell-1)}$ for $\psi = (1-\beta^2)\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\ddot{\sigma}(z)\right]$, we obtain $\epsilon^{(\ell)} < 2(1-\beta^2)r^{\ell-1}$ and $\dot{\epsilon}^{(\ell)} \leq 2(1-\beta^2)\psi r^{\ell-2}$. As a result:

$$\begin{aligned}
\Theta_\infty^{(L)}(x,y) &\geq \sum_{\ell=1}^{L} r^{L-\ell} - 2(1-\beta^2)r^{L-\ell}r^{\ell-1} - \sum_{k=\ell+1}^{L} 2(1-\beta^2)\psi r^{L-\ell-1}r^{k-2} \\
&= \Theta_\infty^{(L)}(x,x) - 2(1-\beta^2)\sum_{\ell=1}^{L} r^{L-1} + \psi \sum_{k=\ell+1}^{L} r^{L-\ell+k-3} \\
&= \Theta_\infty^{(L)}(x,x) - 2(1-\beta^2)\left[Lr^{L-1} + \psi r^{L-2}\sum_{\ell=1}^{L}\frac{1-r^{L-\ell}}{1-r}\right] \\
&\geq \Theta_\infty^{(L)}(x,x) - 2(1-\beta^2)\left[r + \psi\frac{1}{1-r}\right]Lr^{L-2} \\
&\geq \Theta_\infty^{(L)}(x,x) - CLr^{L}.
\end{aligned}$$

Now, let us suppose that $r > 1$. Recall that $B_\beta \circ R_\sigma$ has a unique fixed point a on $[0,1)$. For any x and y in $\mathbb{S}_{n_0}$, the kernels $\Sigma^{(\ell)}(x,y)$ are bounded in norm by $v = max\left\{\left|\beta^2 + \frac{1-\beta^2}{n_0}x^T y\right|, a\right\}$ from Lemma 11. For the kernels $\dot{\Sigma}^{(\ell)}$ we have $\left|\dot{\Sigma}^{(\ell)}(x,y)\right| = (1-\beta^2)\left|R_{\dot{\sigma}}(\Sigma^{(\ell-1)}(x,y))\right| \leq (1-\beta^2)R_{\dot{\sigma}}(\left|\Sigma^{(\ell-1)}(x,y)\right|) \leq (1-\beta^2)R_{\dot{\sigma}}(v) =: w$ where the first inequality follows from the fact that $R_{\dot{\sigma}}(\rho) = \sum_i b_i\rho^i$ for $b_i \geq 0$ and the second follows from the monotonicity of $R_{\dot{\sigma}}$ in $[0,1]$. Applying these two bounds, we obtain:

$$\left|\Theta_\infty^{(L)}(x,y)\right| \leq \sum_{\ell=1}^{L} v \prod_{k=\ell+1}^{L} w = v\frac{1-w^L}{1-w}.$$

Since $\Theta_\infty^{(L)}(x,y) = \frac{1-r^L}{1-r}$, we have that $\left|\vartheta_L\left(x,y\right)\right| \leq v\frac{1-w^L}{1-r^L}$. If $x \neq y$ then $v < 1$ and since σ is nonlinear, $w = (1-\beta^2)R_{\dot{\sigma}}(v) < (1-\beta^2)R_{\dot{\sigma}}(1) = r$. This implies that $\left|\vartheta_L\left(x,y\right)\right|$ converges to zero at an exponential rate, as $L \to \infty$. ■

B.1. ReLU FC-NN

For the standardized ReLU nonlinearity, $\sigma\left(x\right) = \sqrt{2}\max\left(x,0\right)$, the dual activation is computed in Daniely et al. (2016):

$$R_\sigma(\rho) = \frac{\sqrt{1-\rho^2} + \left(\pi - \cos^{-1}(\rho)\right)\rho}{\pi},$$

and the dual activation of its derivative is given by:

$$R_{\dot{\sigma}}(\rho) = \frac{\pi - \cos^{-1}(\rho)}{\pi}.$$

The characteristic value $r = r_{\sigma,\beta}$ of the standardized ReLU is equal to $1-\beta^2$: the ReLU nonlinearity therefore lies in the “order” regime as soon as $\beta > 0$. More explicitly, Lemma 11 still holds of the standardized ReLU and the following inequalities hold for any $x, y \in \mathbb{S}_{n_0}$:

$$1 \geq \Sigma^{(\ell)}(x,y) \geq 1 - 2r^{\ell}.$$

Using these bounds, we can now prove Theorem 4.

Theorem 13 (Theorem 4 in the main) *With the same notation as in Theorem 12, taking σ to be the standardized ReLU and $\beta > 0$, we are in the weakly ordered regime: there exists a constant C such that $1 - CLr^{L/2} \leq \vartheta^{(L)}\left(x, y\right) \leq 1$.*

Proof The first inequality $\vartheta_L\left(x, y\right) \leq 1$ follows the same proof as in the differentiable case.

For the lower bound, using the fact that $\left(1-\beta\right)r = 1$, we have $\epsilon^{(\ell)} = 1 - \Sigma^{(\ell)}(x, y) \leq 2r^{\ell}$ and using the explicit value of $R_{\dot\sigma}(\rho)$, we get that $R_{\dot\sigma}(\rho) \geq 1 - \sqrt{1-\rho}$ which implies that $\dot\epsilon^{(\ell)} = r - \dot\Sigma^{(\ell)}(x, y) \leq r\sqrt{2}r^{\frac{\ell-1}{2}}$: using (3), we write

$$\Theta_{\infty}^{(L)}(x, y) = \sum_{\ell=1}^{L}\left(1 - \epsilon^{(\ell)}\right)\prod_{k=\ell+1}^{L}\left(r - \dot\epsilon^{(k)}\right) \geq \sum_{\ell=1}^{L} r^{L-\ell} - 2r^{L-\ell}r^{\ell} - \sqrt{2}\sum_{k=\ell+1}^{L} r^{L-\ell-1+\frac{k-1}{2}}$$
$$\geq \Theta_{\infty}^{(L)}(x, x) - 2Lr^{L} - \sqrt{2}\sum_{\ell=1}^{L} r^{L-\frac{\ell}{2}-1}\sum_{k=0}^{L-\ell-1} r^{\frac{k}{2}}.$$

Focusing on bounding the double sum from above, we have

$$\sqrt{2}\sum_{\ell=1}^{L} r^{L-\frac{\ell}{2}-1}\sum_{k=0}^{L-\ell-1} r^{\frac{k}{2}} \leq \frac{\sqrt{2}}{1-\sqrt{r}} r^{\frac{L}{2}-1}\sum_{\ell=0}^{L-1} r^{\frac{\ell}{2}}\frac{\sqrt{2}}{1-\sqrt{r}} r^{\frac{L}{2}-1}\frac{1}{1-\sqrt{r}}$$
$$\leq \frac{\sqrt{2}}{r\left(1-\sqrt{r}\right)^{2}} r^{\frac{L}{2}}$$

Hence, we see that

$$\Theta_{\infty}^{(L)}(x, y) \geq \Theta_{\infty}^{(L)}(x, x) - \left[2Lr^{\frac{L}{2}} - \frac{\sqrt{2}}{r\left(1-\sqrt{r}\right)^{2}}\right] r^{\frac{L}{2}}.$$

Recall that for any $x \in \mathbb{S}_{n_0}$, $\Theta_{\infty}^{(L)}(x, x) = \frac{1-r^L}{1-r}$ is bounded in L. Dividing the previous inequality by $\Theta_{\infty}^{(L)}(x, x)$ we get: $1 - Cr^{L/2} \leq \vartheta_L\left(x, y\right) \leq 1$, as claimed, where the constant C is explicit. ■

Appendix C. Layer Normalization and Nonlinearity Normalization

This section of the Appendix is devoted to the proof of Proposition 6.

C.1. Layer normalization is asymptotically equivalent to nonlinearity normalization.

With Layer Normalization (LN), the coordinates of the normalized vectors of activations are $\breve\alpha_j^{(\ell)}(x) = \sqrt{n_\ell}\frac{\alpha_j^{(\ell)}(x) - \mu^{(\ell)}(x)}{||\alpha^{(\ell)}(x) - \underline{\mu}^{(\ell)}(x)||}$, where $\mu^{(\ell)} := \frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\alpha_i^{(\ell)}(x)$ and $\underline{\mu}^{(\ell)} := \begin{pmatrix}\mu^{(\ell)} \\ \vdots \\ \mu^{(\ell)}\end{pmatrix}$. We simplify the notation by making the dependence on x implicite and denote the standardized nonlinearity $\underline{\sigma}(\cdot) := \frac{\sigma(\cdot) - \mathbb{E}(\sigma(Z))}{\sqrt{\mathrm{Var}(\sigma(Z))}}$, where $Z \overset{d}{\sim} \mathcal{N}(0, 1)$.

Suppose that $L = 2$, that is we have a single hidden layer after which the LN is applied. More precisely, the output of the network function with LN is $\widetilde{\alpha}^{(2)}(\check{\alpha}^{(1)}(x))$. We rewrite

$$\check{\alpha}^{(1)} = \sqrt{n_1}\frac{\sigma(\widetilde{\alpha}^{(1)}) - \underline{\mu}^{(1)}}{||\sigma(\widetilde{\alpha}^{(1)}) - \underline{\mu}^{(1)}||} = \underline{\sigma}(\widetilde{\alpha}^{(1)})C_1 + C_2,$$

$$\text{where} \qquad C_1 = \sqrt{n_1}\frac{\sqrt{\mathrm{Var}(\sigma(\mathrm{Z}))}}{||\sigma(\widetilde{\alpha}^{(1)}) - \underline{\mu}^{(1)}||}, \quad \text{and} \quad C_2 = \sqrt{n_1}\frac{\mathbb{E}(\sigma(Z)) - \mu^{(1)}}{||\sigma(\widetilde{\alpha}^{(1)}) - \underline{\mu}^{(1)}||}.$$

Note that $C_1 \to 1$ and $C_2 \to 0$ almost surely, as $n_1 \to \infty$. Indeed, since the $\widetilde{\alpha}_i^{(1)}$'s are independent standard Gaussian variables at initialization (recall that we assume that the inputs belong to $\mathbb{S}_{n_0}$), the law of large numbers entails that $\mu^{(1)} \to \mathbb{E}(\sigma(Z))$ almost surely, as $n_1 \to \infty$, and similarly for $\frac{||\sigma(\widetilde{\alpha}^{(1)}) - \underline{\mu}^{(1)}||^2}{n_1} \to \mathrm{Var}(\sigma(Z))$.

To show that LN is asymptotically equivalent to centering and standardizing the nonlinearity, we now establish that C_1 and C_2 are constant during training. We have

$$\frac{\partial}{\partial \widetilde{\alpha}_j^{(1)}}||\sigma(\widetilde{\alpha}^{(1)}) - \underline{\mu}^{(1)}|| = \frac{\dot{\sigma}(\widetilde{\alpha}_j^{(1)})\sum_{i=1}^{n_1}(\delta_{ij} - 1/n_1)(\sigma(\widetilde{\alpha}_i^{(1)}) - \mu^{(1)})}{||\sigma(\widetilde{\alpha}^{(1)}) - \underline{\mu}^{(1)}||} = \frac{\dot{\sigma}(\widetilde{\alpha}_j^{(1)})(\sigma(\widetilde{\alpha}_j^{(1)}) - \mu^{(1)})}{||\sigma(\widetilde{\alpha}^{(1)}) - \underline{\mu}^{(1)}||}. \tag{4}$$

Note that the absolute value of the latter is bounded by $2||\dot{\sigma}||_\infty$. We write $g(t)$ for any function g that depends on the parameters $\theta(t)$ at time $t \geq 0$. Using twice the triangle inequality yields that

$$\begin{aligned}\Big| ||\sigma(\widetilde{\alpha}^{(1)}(t)) - \underline{\mu}^{(1)}(t)|| - ||\sigma(\widetilde{\alpha}^{(1)}(0)) - \underline{\mu}^{(1)}(0)|| \Big| &\leq ||\sigma(\widetilde{\alpha}^{(1)}(t)) - \sigma(\widetilde{\alpha}^{(1)}(0))|| + ||\underline{\mu}^{(1)}(t) - \underline{\mu}^{(1)}(0)|| \\ &\leq ||\dot{\sigma}||_\infty \left(\left(\sum_{i=1}^{n_1} (\widetilde{\alpha}_i^{(1)}(t) - \widetilde{\alpha}_i^{(1)}(0))^2 \right)^{1/2} + \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \left| \widetilde{\alpha}_i^{(1)}(t) - \widetilde{\alpha}_i^{(1)}(0) \right| \right) \leq ct, \end{aligned} \tag{5}$$

for some constant $c > 0$, where we used that $|\widetilde{\alpha}_i^{(1)}(t) - \widetilde{\alpha}_i^{(1)}(0)| = \mathcal{O}(t/\sqrt{n_1})$, see Appendix A.2 of Jacot et al. (2018). Since $||\sigma(\widetilde{\alpha}^{(1)}(0)) - \underline{\mu}^{(1)}(0)|| \sim \sqrt{n_1}$ by the law of large numbers, we can always write $||\sigma(\widetilde{\alpha}^{(1)}(t)) - \underline{\mu}^{(1)}(t)|| > ||\sigma(\widetilde{\alpha}^{(1)}(0)) - \underline{\mu}^{(1)}(0)|| - ct > 0$. Hence, using (4) then (5), we get

$$\begin{aligned}\left| \frac{\partial C_1(t)}{\partial \widetilde{\alpha}_j^{(1)}(t)} \right| &= \frac{\sqrt{n_1}\mathrm{Var}(\sigma(Z))}{||\sigma(\widetilde{\alpha}^{(1)}(t)) - \underline{\mu}^{(1)}(t)||^2} \cdot \left| \frac{\dot{\sigma}(\widetilde{\alpha}_j^{(1)}(t))(\sigma(\widetilde{\alpha}_j^{(1)}(t)) - \mu^{(1)}(t))}{||\sigma(\widetilde{\alpha}^{(1)}(t)) - \underline{\mu}^{(1)}(t)||} \right| \\ &\leq \frac{\sqrt{n_1}\mathrm{Var}(\sigma(Z))}{(||\sigma(\widetilde{\alpha}^{(1)}(0) - \underline{\mu}^{(1)}(0))|| - ct)^2}||\dot{\sigma}||_\infty = \mathcal{O}(1/\sqrt{n_1}), \end{aligned} \tag{6}$$

by the law of large numbers. The case of C_2 is similar:

$$\begin{aligned}\frac{\partial C_2(t)}{\partial \widetilde{\alpha}_j^{(1)}(t)} &= \frac{-\dot{\sigma}(\widetilde{\alpha}_j^{(1)}(t))}{\sqrt{n_1}||\sigma(\widetilde{\alpha}^{(1)}(t)) - \underline{\mu}^{(1)}(t)||} - \sqrt{n_1}\frac{(\mathbb{E}(\sigma(Z)) - \mu^{(1)}(t))\dot{\sigma}(\widetilde{\alpha}_j^{(1)}(t))(\sigma(\widetilde{\alpha}_j^{(1)}(t)) - \mu^{(1)}(t))}{||\sigma(\widetilde{\alpha}^{(1)}(t)) - \underline{\mu}^{(1)}(t)||^3} \\ &\leq ||\dot{\sigma}||_\infty \left(\frac{1}{n_1} \frac{\sqrt{n_1}}{||\sigma(\widetilde{\alpha}^{(1)}(0)) - \underline{\mu}^{(1)}(0)|| - ct} - \frac{1}{\sqrt{n_1}} \frac{n_1(\mathbb{E}(\sigma(Z)) - \mu^{(1)}(0) + ct)}{(||\sigma(\widetilde{\alpha}^{(1)}(0)) - \underline{\mu}^{(1)}(0)|| - ct)^2} \right) = \mathcal{O}(1/\sqrt{n_1}), \end{aligned} \tag{7}$$

again by the law of large numbers. For $i = 1, 2$, we now write $\frac{\partial C_i(t)}{\partial t} = \frac{\partial \widetilde{\alpha}_j^{(1)}(t)}{\partial t} \frac{\partial C_i(t)}{\partial \widetilde{\alpha}_j^{(1)}(t)}$ and recall that the first term is changing at rate $\mathcal{O}(1/\sqrt{n_1})$. Therefore, $|C_i(t) - C_i(0)| \leq \mathcal{O}(t/n_1)$. The claim for $L \geq 3$ follows by induction.

C.2. Pre-layer normalization has asymptotically no effect.

Normalizing the preactivations has asymptotically no effect on the network at initialization as well as during training. The output of the ℓ-th layer becomes $\check{\alpha}_j^{(\ell)} = \sigma\big(\sqrt{n_\ell}\frac{\widetilde{\alpha}_j^{(\ell)} - \mu^{(\ell)}}{||\widetilde{\alpha}^{(\ell)} - \underline{\mu}^{(\ell)}||}\big)$ where $\mu^{(\ell)}$ and $\underline{\mu}^{(\ell)}$ are computed similarly as before with $\widetilde{\alpha}^{(\ell)}$ in place of $\alpha^{(\ell)}$. As before, we assume $L = 2$ and deduce the general case by induction. We write $\check{\alpha}_j^{(1)} = \sigma(\widetilde{\alpha}_j^{(1)} C_1 + C_2)$, with $C_1 = \sqrt{n_1}/||\widetilde{\alpha}^{(\ell)} - \underline{\mu}^{(\ell)}||$ and $C_2 = -\sqrt{n_1}\mu^{(1)}/||\widetilde{\alpha}^{(\ell)} - \underline{\mu}^{(\ell)}||$. Again, the law of large numbers show that $C_1 \to 1$ and $C_2 \to 0$ almost surely, as $n_1 \to \infty$. Moreover, similarily as (4) and (5), we have that

$$\frac{\partial}{\partial \widetilde{\alpha}_j^{(1)}} ||\widetilde{\alpha}^{(1)} - \underline{\mu}^{(1)}|| = \frac{\widetilde{\alpha}_j^{(1)} - \mu^{(1)}}{||\widetilde{\alpha}^{(1)} - \underline{\mu}^{(1)}||},$$

$$\Big| ||\widetilde{\alpha}^{(1)}(t) - \underline{\mu}^{(1)}(t)|| - ||\widetilde{\alpha}^{(1)}(0) - \underline{\mu}^{(1)}(0)|| \Big| \leq ct,$$

for some constant $c > 0$. Using the same argument as in (6) and (7), one can thus show for $i = 1, 2$ that

$$\left| \frac{\partial C_i(t)}{\partial \widetilde{\alpha}_j^{(1)}} \right| = \mathcal{O}(1/\sqrt{n_1}).$$

We conclude as previously, noting that

$$\frac{\partial \check{\alpha}_j^{(1)}(t)}{\partial t} = \dot{\sigma}\left(\widetilde{\alpha}_j^{(1)}(t) C_1(t) + C_2(t)\right) \left(\frac{\partial \widetilde{\alpha}_j^{(1)}(t)}{\partial t} C_1(t) + \widetilde{\alpha}_j^{(1)}(t) \frac{\partial C_1(t)}{\partial t} + \frac{\partial C_2(t)}{\partial t}\right).$$

Appendix D. Batch Normalization

If one adds a BatchNorm layer after the nonlinearity of the last hidden layer, we have:

Lemma 14 (Lemma 7 in the main) *Consider a FC-NN with L layers, with a PN-BN after the last nonlinearity. For any $k, k' \in \{1, \ldots, n_L\}$ and any parameter θ_p, we have $\sum_{i=1}^N \Theta_{\theta_p}^{(L)}(\cdot, x_i) = \beta^2 \mathrm{Id}_{n_L}$.*

Proof This is an direct consequence of the definition of the NTK and of the following claim:

Claim. *For a fully-connected DNN with a BatchNorm layer after the nonlinearity of the last hidden layer then $\frac{1}{N}\sum_{i=1}^N \partial_{\theta_p} f_{\theta,k}(x_i)$ is equal to β if θ_p is $b_k^{(L-1)}$, the bias parameter of the last layer, and equal to 0 otherwise.*

The average of $f_{\theta,k}$ on the training set, $\frac{1}{N}\sum_{i=1}^N \partial_{\theta_p} f_{\theta,k}(x_i)$, only depends on the bias of the last layer:

$$\frac{1}{N}\sum_{i=1}^N f_{\theta,k}(x_i) = \frac{\sqrt{1-\beta^2}}{\sqrt{n_{L-1}}} W^{(L-1)} \frac{1}{N}\sum_{i=1}^N \hat{\alpha}^{(L-1)}(x_i) + \beta b_k^{(L-1)} = \beta b_k^{(L-1)}.$$

Thus for any parameter θ_p, $\frac{1}{N}\sum_{i=1}^{N}\partial_{\theta_p} f_{\theta,k}(x_i) = \partial_{\theta_p}\left(\beta b_k^{(L-1)}\right)$ is equal to β if the parameter is the bias $b_k^{(L-1)}$ and zero otherwise. ■

Appendix E. Graph-based Neural Networks

Recall the definition of GB-NNs from Section 5.1.1) and notation therein.

In this section, we prove the convergence of the NTK at initialization for a general family of DNNs which contain in particular CNNs and DC-NNs. We will consider the graph-based parametrization introduced in the main.

For any layer $\ell+1$ and neurons $p, p' \in I_{\ell+1}$ and $q \in P(p)$, $q' \in P(p')$, we denote by $\chi(q \to p, q' \to p')$ the map that is equal to 1 if and only if $W^{(\ell,q\to p)}$ and $W^{(\ell,q'\to p')}$ are shared (in the sense that the two matrices are forced to be equal at initialization and during training) and 0 otherwise. It satisfies $\chi(q \to p, q \to p) = 1$ for any neuron p and any $q \in P(p)$ and it is transitive. The assumption that no pair of connections leading to the same neuron are shared, stated in Section 5.1.1 reads as follows: $\forall \ell = 1..L$, $\forall p \in I_{\ell+1}$, $\forall q, q' \in P(p)$, it holds that $\chi(q \to p, q' \to p) = \delta_{qq'}$.

Remark 15 *Note that the parametrization (2) is different from the traditional one: we divide by $\sqrt{|P(p)|\, n_\ell}$ instead of dividing by $\sqrt{n_\ell \frac{|\omega|}{s_1 \dots s_d}}$. This does not lead to any difference when one consider infinite-sized images as in Section F since in this case the number of parents is constant, equal to $\frac{|\omega|}{s_1 \dots s_d}$. The key difference between the two parametrizations will be investigated in Appendix G.*

Before we can derive an explicit formula for the NTK, we compute the infinite width limit of the feature kernels. Recall, that for a kernel $K : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}$, and for any $z_0, z_1 \in \mathbb{R}^{n_0}$, we defined:

$$\mathbb{L}_K^g(z_0, z_1) = \mathbb{E}_{(y_0,y_1)\sim\mathcal{N}\left(0,(K(z_i,z_j))_{i,j=0,1}\right)}\left[g(y_0)\, g(y_1)\right].$$

Proposition 16 *In the above setting, as $n_1 \to \infty$, ...,$n_{\ell-1} \to \infty$ sequentially, the preactivations $\left(\tilde{\alpha}_i^{(\ell,p)}(x)\right)_{i=1,\dots,n_\ell, p\in I_\ell}$ of the ℓ^{th} layer converge to a centered Gaussian process with covariance $\Sigma^{(\ell,pp')}(x,y)\delta_{ii'}$, where $\Sigma^{(\ell,pp')}(x,y)$ is defined recursively as*

$$\Sigma^{(1,pp')}(x,y) = \beta^2 + \frac{1-\beta^2}{\sqrt{|P(p)|\,|P(p')|} n_0}\sum_{q\in P(p)}\sum_{q'\in P(p')}\chi(q \to p, q' \to p')\,(x_q)^T\, y_{q'},$$

$$\Sigma^{(\ell+1,pp')}(x,y) = \beta^2 + \frac{1-\beta^2}{\sqrt{|P(p)|\,|P(p')|}}\sum_{q\in P(p)}\sum_{q'\in P(p')}\chi(q \to p, q' \to p')\mathbb{L}^\sigma_{\Sigma^{(\ell,qq')}}(x,y)\,.$$

Proof The proof is done by induction on ℓ. For $\ell = 1$ and any $i \in \{1, \dots, n_1\}$, the preactivation

$$\tilde{\alpha}_i^{(1,p)}(x) = \beta b_i^{(0)} + \frac{\sqrt{1-\beta^2}}{\sqrt{|P(p)|\, n_0}}\sum_{q\in P(p)}\left(W_p^{(0,q\to p)} x_q\right)_i$$

is a random affine function of x and its coefficients are centered Gaussian: it is hence a centered Gaussian process whose covariance is easily shown to be equal to $\mathbb{E}\left[\tilde{\alpha}_i^{(1,p)}(x)\tilde{\alpha}_{i'}^{(1,p')}(y)\right] = \Sigma^{(1,pp')}(x,y)\delta_{ii'}$.

For the induction step, we assume that the result holds for the pre-activations of the layer ℓ. The pre-activations of the next layer are of the form

$$\tilde{\alpha}_i^{(\ell+1,p)}(x) = \beta b_i^{(0)} + \frac{\sqrt{1-\beta^2}}{\sqrt{|P(p)|\, n_\ell}} \sum_{q \in P(p)} \left(W^{(\ell,q\to p)} \alpha^{(\ell,q)}(x) \right)_i .$$

Conditioned on the activations $\alpha^{(\ell,q)}$ of the last layer, $\tilde{\alpha}^{(\ell+1,p)}$ is a centered Gaussian process: in other terms, it is a mixture of centered Gaussians with a random covariance determined by the activations of the last layer. The random covariance between $\tilde{\alpha}_{i_0}^{(\ell+1,p_0)}(x)$ and $\tilde{\alpha}_{i_1}^{(\ell+1,p_1)}(y)$ is equal to

$$\beta^2 \delta_{i_0 i_1} + \frac{1-\beta^2}{\sqrt{|P(p)|\,|P(p')| n_\ell}} \sum_{\substack{q_0 \in P(p_0) \\ q_1 \in P(p_1)}} \sum_{j_0,j_1=1}^{n_\ell} \mathbb{E}\left[W_{i_0 j_0}^{(\ell,q_0\to p_0)} W_{i_1 j_1}^{(\ell,q_1\to p_1)} \right] \alpha_{j_0}^{(\ell,q_0)}(x) \alpha_{j_1}^{(\ell,q_1)}(y)$$

$$= \delta_{i_0 i_1} \left[\beta^2 + \frac{1-\beta^2}{\sqrt{|P(p)|\,|P(p')|}} \sum_{\substack{q_0 \in P(p_0) \\ q_1 \in P(p_1)}} \chi(q_0 \to p_0, q_1 \to p_1) \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \sigma\left(\tilde{\alpha}_j^{(\ell,q_0)}(x) \right) \sigma\left(\tilde{\alpha}_j^{(\ell,q_1)}(y) \right) \right],$$

where we used the fact that $\mathbb{E}\left[W_{i_0 j_0}^{(\ell,q_0\to p_0)} W_{i_1 j_1}^{(\ell,q_1\to p_1)} \right] = \chi(q_0 \to p_0, q_1 \to p_1)\delta_{i_0 i_1}\delta_{j_0 j_1}$. Using the induction hypothesis, as $n_1 \to \infty, \ldots, n_{\ell-1} \to \infty$ sequentially, the preactivations $\left(\tilde{\alpha}_j^{(\ell,q_0)}(x), \tilde{\alpha}_j^{(\ell,q_1)}(y) \right)_j$ converge to independant centered Gaussian pairs. As $n_\ell \to \infty$, by the law of large numbers, the sum over j along with the $\frac{1}{n_\ell}$ converges to $\mathbb{L}_\sigma^{\Sigma^{(\ell,qq')}}(x,y)$. In this limit, the random covariance of the Gaussian mixture becomes deterministic and as a consequence, the mixture of Gaussian processes tends to a centered Gaussian process with the right covariance. ■

Similarly to the activation kernels, one can prove that the NTK converges at initialization.

Proposition 17 *As $n_1 \to \infty, \ldots, n_{L-1} \to \infty$ sequentially, the NTK $\Theta^{(L,p_0p_1)}$ of a general convolutional network converges to $\Theta_{\infty,p_0p_1}^{(L)} \otimes \mathrm{Id}_{n_L}$ where $\Theta_\infty^{(L,p_0p_1)}(x,y)$ is defined recursively by:*

$$\begin{aligned}
\Theta_\infty^{(1,p_0p_1)}(x,y) &= \Sigma^{(1,p_0p_1)}(x,y), \\
\Theta_\infty^{(L,p_0p_1)}(x,y) &= \frac{1-\beta^2}{\sqrt{|P(p_0)|\,|P(p_1)|}} \sum_{\substack{q_0 \in P(p_0) \\ q_1 \in P(p_1)}} \chi(q_0 \to p_0, q_1 \to p_1) \Theta_\infty^{(L-1,q_0q_1)}(x,y) \mathbb{L}_{\Sigma^{(L-1,q_0q_1)}}^{\dot{\sigma}}(x,y) \\
&\quad + \Sigma^{(L,p_0p_1)}(x,y).
\end{aligned}$$

Proof The proof by induction on L follows the one of Jacot et al. (2018) for fully-connected DNNs. We present the induction step and assume that the result holds for a general convolutional network with $L-1$ hidden layers. Following the same computations as in Jacot et al. (2018), the NTK $\Theta_{p_0p_1,jj'}^{(L+1)}(x,y)$ is equal to

$$\frac{1-\beta^2}{\sqrt{|P(p_0)|\,|P(p_1)|}n_L}\sum_{q_0\in P(p_0)}\sum_{q_1\in P(p_1)}\sum_{ii'}\Theta_{ii'}^{(L,q_0q_1)}(x,y)\dot{\sigma}\left(\tilde{\alpha}_i^{(L,q_0)}(x)\right)\dot{\sigma}\left(\tilde{\alpha}_{i'}^{(L,q_1)}(y)\right)$$
$$W_{ij}^{(L,q_0\to p_0)}W_{i'j'}^{(L,q_1\to p_1)}$$
$$+\delta_{jj'}\beta^2+\delta_{jj'}\frac{1-\beta^2}{\sqrt{|P(p_0)|\,|P(p_1)|}n_L}\sum_{q_0\in P(p_0)}\sum_{q_1\in P(p_1)}\chi(q_0\to p_0,q_1\to p_1)\sum_i\alpha_i^{(L,q_0)}(x)\alpha_i^{(L,q_1)}(y)$$

which, by assumption, converges as $n_1\to\infty,\ldots,n_{L-1}\to\infty$ to

$$\frac{1-\beta^2}{\sqrt{|P(p_0)|\,|P(p_1)|}n_L}\sum_{q_0\in P(p_0)}\sum_{q_1\in P(p_1)}\sum_{i}\Theta_{\infty}^{(L,q_0q_1)}(x,y)\dot{\sigma}\left(\tilde{\alpha}_i^{(L,q_0)}(x)\right)\dot{\sigma}\left(\tilde{\alpha}_{i}^{(L,q_1)}(y)\right)$$
$$W_{ij}^{(L,q_0\to p_0)}W_{ij'}^{(L,q_1\to p_1)}$$
$$+\delta_{jj'}\beta^2+\delta_{jj'}\frac{1-\beta^2}{\sqrt{|P(p_0)|\,|P(p_1)|}n_L}\sum_{q_0\in P(p_0)}\sum_{q_1\in P(p_1)}\chi(q_0\to p_0,q_1\to p_1)\sum_i\alpha_i^{(L,q_0)}(x)\alpha_i^{(L,q_1)}(y).$$

As $n_L\to\infty$, using the previous results on the preactivations and the law of large number, the NTK converges to

$$\frac{1-\beta^2}{\sqrt{|P(p_0)|\,|P(p_1)|}}\sum_{q_0\in P(p_0)}\sum_{q_1\in P(p_1)}\Theta_{\infty}^{(L,q_0q_1)}(x,y)\mathbb{L}^{\dot{\sigma}}_{\Sigma^{(L,q_0q_1)}}(x,y)\,\mathbb{E}\left[W_{ij}^{(L,q_0\to p_0)}W_{ij'}^{(L,q_1\to p_1)}\right]$$

$$+\delta_{jj'}\beta^2+\delta_{jj'}\frac{1-\beta^2}{\sqrt{|P(p_0)|\,|P(p_1)|}}\sum_{q_0\in P(p_0)}\sum_{q_1\in P(p_1)}\chi(q_0\to p_0,q_1\to p_1)\mathbb{L}^{\sigma}_{\Sigma^{(L,q_0q_1)}}(x,y)\,,$$

which can be simplified–using the fact that $\mathbb{E}\left[W_{ij}^{(L,q_0\to p_0)}W_{ij'}^{(L,q_1\to p_1)}\right]=\chi(q_0\to p_0,q_1\to p_1)\delta_{jj'}$–into:

$$\delta_{jj'}\frac{1-\beta^2}{\sqrt{|P(p_0)|\,|P(p_1)|}}\sum_{q_0\in P(p_0)}\sum_{q_1\in P(p_1)}\chi(q_0\to p_0,q_1\to p_1)\Theta_{\infty}^{(L,q_0q_1)}(x,y)\mathbb{L}^{\dot{\sigma}}_{\Sigma^{(L,q_0q_1)}}(x,y)$$
$$+\delta_{jj'}\Sigma^{(L+1,p_0p_1)}(x,y),$$

which proves the assertions. ■

Appendix F. Deconvolutional Neural Networks

In this section, in order to study the behaviour of DC-NNs in the bulk and to avoid dealing with border effects, studied in Section G, we assume that for all layers ℓ there is no border, i.e. the positions p are in $\mathbb{Z}^d$. Let us consider a DC-NN with up-sampling $s\in\{2,3,...\}^d$ where the window sizes for all layers are all set equal to $\pi = \omega = \{0, \cdots, w_1 s_1 - 1\} \times \cdots \times \{0, \cdots, w_d s_d - 1\}$. A position p has therefore $w_1 \cdots w_d$ parents which are given by

$$P(p) = \left\{ \left\lfloor \frac{p_0}{s_0} \right\rfloor, \left\lfloor \frac{p_0}{s_0} \right\rfloor + 1, \cdots, \left\lfloor \frac{p_0}{s_0} \right\rfloor + w_1 \right\} \times \cdots \times \left\{ \left\lfloor \frac{p_d}{s_d} \right\rfloor, \left\lfloor \frac{p_d}{s_d} \right\rfloor + 1, \cdots, \left\lfloor \frac{p_d}{s_d} \right\rfloor + w_d \right\}.$$

Two connections $q \to p$ and $q' \to p'$ are shared if and only if $s \mid p - p'$ (i.e. for any $i = 1, ..., d$, $s_i \mid p_i - p'_i$) and $q_i - q'_i = \frac{p_i - p'_i}{s_i}$ for any $i = 1, ..., d$.

F.1. Order and Chaos Regimes

Propositions 16 and 17 hold true in this setting. By Proposition 23, if the nonlinearity σ is standardized, $\Sigma^{(\ell,pp)}(x,x) = 1$ for any $x \in \mathbb{S}_{n_0}^{I_0}$ and any $p \in I_\ell$. The activation kernels $\Sigma^{(\ell,pp')}(x,y)$ for any two inputs $x, y \in \mathbb{S}_{n_0}^{I_0}$ and two output positions $p, p' \in \mathbb{Z}^d$ are therefore defined recursively by:

$$\Sigma^{(1,pp')}(x,y) = \beta^2 + \delta_{s|p-p'} \frac{1-\beta^2}{|P(p)|\, n_0} \sum_{q \in P(p)} (x_q)^T y_{q + \frac{p'-p}{s}},$$

$$\Sigma^{(\ell+1,pp')}(x,y) = \beta^2 + \delta_{s|p-p'} \frac{1-\beta^2}{|P(p)|} \sum_{q \in P(p)} R_\sigma \left(\Sigma^{(\ell,q,q+\frac{p'-p}{s})}(x,y) \right),$$

where $\frac{p'-p}{s} = \left(\frac{p'_i - p_i}{s_i} \right)_i$ is a valid position since $s|p - p'$. Similarly, the NTK at initialization satisfies the following recursion:

$$\Theta_\infty^{(L+1,pp')}(x,y) = \Sigma^{(L+1,pp')}(x,y) + \delta_{s|p-p'} \frac{1-\beta^2}{|P(p)|} \sum_{q \in P(p)} \Theta_\infty^{(L,q,q+\frac{p'-p}{s})}(x,y) R_{\dot\sigma} \left(\Sigma^{(L,q,q+\frac{p'-p}{s})}(x,y) \right).$$

Remark 18 *Recall that the s-valuation $v_s(n)$ of a number $n \in \mathbb{Z}^d$ is the largest $k \in \{0, 1, 2, \dots\}$ such that $s_i^k \mid n_i$ for all dimensions $i = 1, ..., d$. For two pixels $p, p' \in \mathbb{Z}^d$ and any input vectors $x, y \in \mathbb{S}_{n_0}^{I_0}$, if $v_s(p' - p) < \ell$ the activation kernel $\Sigma^{(\ell,pp')}(x,y)$ does not depend neither on x nor on y. More precisely, if $v = v_s(p' - p) = 0$, we have*

$$\Sigma^{(\ell,pp')}(x,y) = \beta^2,$$

and for a general $v < \ell$:

$$c_v := \Sigma^{(\ell,pp')}(x,y) = (B_\beta \circ R_\sigma)^{\circ v} (\beta^2).$$

In particular, if $v < L$, the NTK is therefore also equal to a constant:

$$\Theta_\infty^{(L,pp')}(x,y) = \sum_{k=0}^{v} c_k (1-\beta^2)^k \prod_{m=0}^{k-1} R_{\dot\sigma}(c_m).$$

We establish the bounds on the rate of convergence in the “order” region and on the values of the activations kernel in the chaos region for DC-NNs.

Proposition 19 *In the setting introduced above, for a standardized twice differentiable σ, for $x, y \in \mathbb{S}_{n_0}^{I_0}$, and any positions $p, p' \in I_\ell$, taking $k = \min\{v_s(p'-p), \ell\}$, we have:*

If $r_{\sigma,\beta} < 1$ then:

$$1 \geq \Sigma^{(\ell,pp')}(x,y) \geq 1 - 2(1-\beta^2) r_{\sigma,\beta}^k.$$

If $r_{\sigma,\beta} > 1$ then there exists a fixed point $a \in [0,1)$ of $B_\beta \circ R_\sigma$ such that:

- *If $k < \ell$:*
$$\left|\Sigma^{(\ell,pp')}(x,y)\right| \leq \max\left\{\beta^2, a\right\},$$

- *If $p' - p = ms^\ell$ and there is a $c \leq 1$ such that for all input positions $q \in P^{\circ\ell}(p)$, $\left|\frac{1}{n_0} x_q^T y_{q+m}\right| \leq c$, then*
$$\left|\Sigma^{(\ell,pp')}(x,y)\right| \leq \max\left\{\beta^2 + (1-\beta^2)c, a\right\}.$$

Proof Let us denote $r = r_{\sigma,\beta}$. Let us suppose that $r < 1$ and let us prove the first assertion by induction on ℓ. If $\ell = 1$, then

$$\begin{aligned}\Sigma^{(1,pp')}(x,y) &= \beta^2 + \delta_{s|p-p'} \frac{1-\beta^2}{|P(p)|\, n_0} \sum_{q \in P(p)} (x_q)^T y_{q+\frac{p'-p}{s}} \geq \beta^2 - \delta_{s|p-p'}(1-\beta^2) \\ &\geq 1 - 2(1-\beta^2)\end{aligned}$$

For the induction step, suppose that the inequality holds true for some $\ell \geq 1$, then

$$\begin{aligned}\Sigma^{(\ell+1,pp')}(x,y) &\geq \beta^2 + \delta_{s|p-p'} \frac{1-\beta^2}{|P(p)|} \sum_{q=0}^{\frac{w}{s}} R_\sigma\left(1 - 2(1-\beta^2) r^{k-1}\right) \\ &\geq \beta^2 + \delta_{s|p-p'} \frac{1-\beta^2}{|P(p)|} \sum_{q=0}^{\frac{w}{s}} 1 - 2(1-\beta^2) R_{\dot\sigma}(1) r^{k-1} \\ &\geq \beta^2 + \delta_{s|p-p'}\left(1 - \beta^2 - 2(1-\beta^2) r^k\right) \\ &= \begin{cases} 1 - (1-\beta^2) & \text{if } k = 0 \\ 1 - 2(1-\beta^2) r^k & \text{if } k > 0 \end{cases} \\ &\geq 1 - 2(1-\beta^2) r^k\end{aligned}$$

Now let us suppose that $r > 1$. If $k < \ell$, then $\left|\Sigma^{(\ell,pp')}(x,y)\right| = \left|(B_\beta \circ R_\sigma)^{\circ k}\left(\beta^2\right)\right| < \max\left\{\beta^2, a\right\}$. Let us suppose at last that $k = \ell$ and let us prove the last assertion by induction on ℓ. If $\ell = 1$, then

$$\begin{aligned}\left|\Sigma^{(1,pp')}(x,y)\right| &\leq \beta^2 + \frac{1-\beta^2}{|P(p)|\, n_0} \sum_{q \in P(p)} \left| x_q^T y_{q+\frac{p'-p}{s}}^T \right| \leq \beta^2 + \frac{1-\beta^2}{|P(p)|} \sum_{q \in P(p)} c \\ &= \beta^2 + (1-\beta^2)c.\end{aligned}$$

For the induction step, if we suppose that the inequality holds true for ℓ, then

$$\begin{aligned}\left|\Sigma^{(\ell+1,pp')}(x,y)\right| &\leq \beta^2 + \frac{(1-\beta^2)}{|P(p)|}\sum_{q\in P(p)}\left|R_\sigma\left(\Sigma^{(\ell,q,q+\frac{p'-p}{s})}(x,y)\right)\right| \\ &\leq \beta^2 + \frac{(1-\beta^2)}{|P(p)|}\sum_{q\in P(p)} R_\sigma\left(\max\{\beta^2+(1-\beta^2)c,a\}\right) \\ &= B_\beta \circ R_\sigma\left(\max\{\beta^2+(1-\beta^2)c,a\}\right) \\ &\leq \max\{\beta^2+(1-\beta^2)c,a\},\end{aligned}$$

which allows us to conclude. ■

The NTK features the same two regimes:

Theorem 20 (Theorem 9 in the main) *Take $I_0 = \mathbb{Z}^d$, and consider a DC-NN with upsampling stride $s \in \{2,3,\ldots\}^d$, windows $\pi = \omega = \{0,\ldots,w_1s_1-1\}\times\ldots\times\{0,\ldots,w_ds_d-1\}$ for $w \in \{1,2,3,\ldots\}^d$. For a standardized twice differentiable σ, there exist constants $C_1, C_2 > 0$, such that the following holds: for $x, y \in \mathbb{S}_{n_0}^{I_0}$, and any positions $p, p' \in I_L$, we have:*

Order: *When $r_{\sigma,\beta} < 1$, taking $v = \min\left(v_s\left(p-p'\right), L-1\right)$, taking $v = L-1$ if $p = p'$ and $r = r_{\sigma,\beta}$, we have*

$$\frac{1-r^{v+1}}{1-r^L} - C_1(v+1)r^v \leq \vartheta_\infty^{(L,p,p')}(x,y) \leq \frac{1-r^{v+1}}{1-r^L}.$$

Chaos: *When $r_{\sigma,\beta} > 1$, if either $v_s\left(p-p'\right) < L$ or if there exists a $c < 1$ such that for all positions $q \in I_0$ which are ancestor of p, $\left|x_q^T y_{q+\frac{p'-p}{s^L}}\right| < c$, then there exists $h < 1$ such that*

$$\left|\vartheta_\infty^{(L,p,p')}(x,y)\right| \leq C_2 h^L.$$

Proof Let us denote $r = r_{\sigma,\beta}$ and let us suppose that $r < 1$. The NTK can be bounded recursively

$$\begin{aligned}\Theta_\infty^{(L,pp')}(x,y) &= \Sigma^{(L,pp')}(x,y) + \delta_{s|p-p'}\frac{1-\beta^2}{|P(p)|}\sum_{q\in P(p)}\Theta_\infty^{(L-1;q,q+\frac{p'-p}{s})}(x,y)R_{\dot\sigma}\left(\Sigma^{(L-1;q,q+\frac{p'-p}{s})}(x,y)\right) \\ &\geq 1 - 2(1-\beta^2)r^v + \delta_{s|p-p'}\frac{1}{|P(p)|}\sum_{q\in P(p)}\Theta_\infty^{(L-1;q,q+\frac{p'-p}{s})}(x,y)\left(r - \psi 2(1-\beta^2)^2 r^{v-1}\right).\end{aligned}$$

Unrolling this inequality then using (3), we get

$$\begin{aligned}
\Theta_\infty^{(L,pp')}(x,y) &= \sum_{k=0}^{v} \left(1-2(1-\beta^2)r^k\right) \prod_{m=k+1}^{v} \left(r - \psi 2(1-\beta^2)^2 r^{m-1}\right) \\
&\geq \sum_{k=0}^{v} r^{v-k} - 2(1-\beta^2) r^{v-k} r^k - \psi 2(1-\beta^2)^2 \sum_{m=k+1}^{v} r^{v-k-1} r^{m-1} \\
&= \frac{1-r^{v+1}}{1-r} - 2(1-\beta^2)(v+1)r^v - \psi 2(1-\beta^2)^2 \sum_{k=0}^{v} r^{v-1} \sum_{m=0}^{v-k-1} r^m \\
&\geq \frac{1-r^{v+1}}{1-r} - 2(1-\beta^2)\left[r + \frac{\psi(1-\beta^2)}{1-r}\right](v+1)r^{v-1} \\
&\geq \frac{1-r^{v+1}}{1-r} - C(v+1)r^v,
\end{aligned}$$

where the constant C is allowed to depend on σ and β. For the upper bound, we have: $\Theta_\infty^{(L,pp')}(x,y) \leq \sum_{\ell=L-k}^{L} 1 \prod_{m=\ell+1}^{L} r = \frac{1-r^{v+1}}{1-r}$. Thus, we get the same bounds as in the FC-NNs case, but with respect to v, which is the maximal integer strictly smaller than L such that $s^v | p - p'$:

$$\frac{1-r^{v+1}}{1-r} \geq \Theta_\infty^{(L,pp')}(x,y) \geq \frac{1-r^{v+1}}{1-r} - C(v+1)r^v.$$

Dividing by $\Theta_\infty^{(L,pp)}(x,x)$ which is bounded in the ordered regime (see proof of Proposition 23) as $L \to \infty$, one gets the desired result.

If $r > 1$, there are two cases. When $p' - p = ks^L$ then if there exists $c < 1$ such that $\left|x_q^T y_{q+k}\right| < cn_0$ for all ancestors q of p. Writing $z = \max\{\beta^2 + (1-\beta^2)c, a\}$ and $w = (1-\beta^2)R_{\dot\sigma}(z) < r$ such that $\left|\Sigma^{(\ell;q,q+ks^\ell)}(x,y)\right| < z$ for all position q at layer ℓ which is an ancestor of p. Then

$$\left|\Theta_\infty^{(L,pp')}(x,y)\right| \leq \sum_{\ell=1}^{L} v w^{L-\ell} = v\frac{1-w^L}{1-w}$$

such that

$$\frac{\left|\Theta_\infty^{(L,pp')}(x,y)\right|}{\left|\Theta_\infty^{(L,pp)}(x,x)\right|} \leq c\frac{1-r}{1-w}\frac{1-w^L}{1-r^L} \leq C(\sigma,\beta)\left(\frac{w}{r}\right)^L$$

which goes to zero exponentially.

If $p' - p$ is not divisible by s^L then for $z = \max\{\beta^2, a\}$ and $w = (1-\beta^2)R_{\dot\sigma}(z) < r$

$$\left|\Theta_\infty^{(L,pp')}(x,y)\right| \leq \sum_{\ell=L-v+1}^{L} z w^{L-\ell} = z\frac{1-w^v}{1-w}$$

which also converges exponentially to 0. ■

F.2. Adapting the learning rate

Let us suppose that we multiply the learning rate of the ℓ-th layer weights and bias by $S^{-\frac{\ell}{2}}$ where $S = \prod_i s_i$. This is slightly different than what we propose in the main, where the learning rate of the bias are multiplied by $S^{-\frac{\ell+1}{2}}$ instead of $S^{-\frac{\ell}{2}}$, but it greatly simplifies the formulas. Furthermore, the balance between the weights and bias can be modified with the meta-parameter β to achieve a similar result. The NTK then takes the value:

$$\Theta^{(L,pp)}(x,x) = \sum_{\ell=1}^{L} S^{-\frac{\ell}{2}} \prod_{n=\ell+1}^{L} r = \sum_{\ell=1}^{L} S^{-\frac{\ell}{2}} r^{L-\ell} = S^{-\frac{L}{2}} \frac{1 - \left(\sqrt{S}r\right)^{L}}{1 - \sqrt{S}r}$$

This leads to another transtion inside the "order" regime: if $\sqrt{S}r < 1$ the NTK $\Theta_\infty^{(L,pp)}(x,x)$ goes to zero and if $\frac{1}{\sqrt{S}} < r < 1$ it converges to a constant. If we translate the bound of Proposition 20 to the NTK with varying learning rates, the convergence to a constant is only guaranteed when $\sqrt{S}r < 1$, which suggests that adapting the learning (or changing the number of channels) does reduce the checkerboard artifacts (as confirmed by numerical experiments):

Proposition 21 *Suppose that $r < 1$. For any two inputs x, y such that for all $p \in \mathbb{Z}$, $\|x^p\| = \|y^p\| = \sqrt{n_0}$ and for any two output positions p, p' such that k is the maximal integer in $\{0, ..., L-1\}$ such that s^k divides the difference $p - p'$, it holds that*

$$\frac{1 - (\sqrt{S}r)^{k+1}}{1 - (\sqrt{S}r)^{L}} \geq \vartheta_\infty^{(L,pp')}(x,y) \geq \frac{1 - (\sqrt{S}r)^{k+1}}{1 - (\sqrt{S}r)^{L}} - \frac{C_{\sigma,\beta}(\sqrt{S}r)^{k}}{\left|1 - (\sqrt{S}r)^{L}\right|}$$

Proof The NTK can be bounded recursively

$$\begin{aligned}
\Theta_\infty^{(L,pp')}(x,y) &= S^{-\frac{L-1}{2}} \Sigma^{(L,pp')}(x,y) + \delta_{s|p-p'} \frac{1-\beta^2}{|P(p)|} \sum_{q \in P(p)} \Theta_\infty^{(L-1;q,q+\frac{p'-p}{s})}(x,y) R_{\dot{\sigma}} \left(\Sigma^{(L-1;q,q+\frac{p'-p}{s})}(x,y) \right) \\
&\geq S^{-\frac{L-1}{2}} (1 - 2(1-\beta^2) r^k) + \delta_{s|p-p'} \frac{1}{|P(p)|} \sum_{q \in P(p)} \Theta_\infty^{(L;q,q+\frac{p'-p}{s})}(x,y) \left(r - \psi 2 (1-\beta^2)^2 r^{k-1} \right)
\end{aligned}$$

unrolling then using (3), we get

$$\begin{aligned}
\Theta_\infty^{(L,pp')}(x,y) &\geq \sum_{m=0}^{k} S^{-\frac{L-k+m}{2}} \left(1 - 2(1-\beta^2) r^m\right) \prod_{n=m+1}^{k} \left(r - \psi 2 (1-\beta^2)^2 r^{n-1}\right) \\
&\geq \sum_{m=0}^{k} S^{\frac{k-m-L}{2}} r^{k-m} - S^{\frac{k-m-L}{2}} 2(1-\beta^2) r^{k-m} r^m - S^{\frac{k-m-L}{2}} \psi 2 (1-\beta^2)^2 \sum_{n=m+1}^{k} r^{k-m-1} r^{n-1} \\
&\geq S^{-\frac{L}{2}} \frac{1 - (\sqrt{S}r)^{k+1}}{1 - \sqrt{S}r} - 2 \frac{1-\beta^2}{1 - S^{-\frac{1}{2}}} S^{\frac{k-L}{2}} r^k - \psi 2 (1-\beta^2)^2 r^{k-1} \sum_{m=0}^{k} S^{\frac{k-m-L}{2}} \sum_{n=0}^{k-m-1} r^n
\end{aligned}$$

We can bound the last term:

$$\psi 2 (1-\beta^2)^2 r^{k-1} \sum_{m=0}^{k} S^{\frac{k-m-L}{2}} \sum_{n=0}^{k-m-1} r^n \leq \psi 2 (1-\beta^2)^2 r^{k-1} S^{\frac{k-L}{2}} \frac{1}{1 - S^{-\frac{1}{2}}} \frac{1}{1-r}$$

Hence, we write

$$\begin{aligned}\Theta_{\infty}^{(L,pp')}(x,y) &\geq S^{-\frac{L}{2}}\left(\frac{1-(\sqrt{S}r)^{k+1}}{1-\sqrt{S}r} - 2\frac{1-\beta^2}{1-S^{-\frac{1}{2}}}\left[1+\frac{\psi r(1-\beta^2)}{1-r}\right]\left(\sqrt{S}r\right)^{k}\right)\\ &\geq S^{-\frac{L}{2}}\left(\frac{1-(\sqrt{S}r)^{k+1}}{1-\sqrt{S}r} - C_{\sigma,\beta}\left(\sqrt{S}r\right)^{k}\right).\end{aligned}$$

For the upper bound, we have that

$$\Theta_{\infty}^{(L,pp')}(x,y) \leq \sum_{m=0}^{k} S^{-\frac{L-k+m}{2}} \prod_{n=m+1}^{k} r = S^{-\frac{L}{2}}\frac{1-(\sqrt{S}r)^{k+1}}{1-\sqrt{S}r}.$$

Dividing by $\Theta_{\infty}^{(L,pp)}(x,x)$ we obtain

$$\frac{1-(\sqrt{S}r)^{k+1}}{1-(\sqrt{S}r)^{L}} \geq \vartheta_{\infty}^{(L,pp')}(x,y) \geq \frac{1-(\sqrt{S}r)^{k+1}}{1-(\sqrt{S}r)^{L}} - \frac{C_{\sigma,\beta}(\sqrt{S}r)^{k}}{\left|1-(\sqrt{S}r)^{L}\right|},$$

as claimed. ■

Appendix G. Border Effects

With the usual scaling of $\frac{1}{\sqrt{\frac{|\omega|}{s_1\cdots s_d}}}$, in a general convolutional network, the positions on the border have less parents and hence a lower activation variance. In this section, we show, in a special example, how this parametrization leads to border effects in the limiting activation kernels and NTK. This could be generalized to a more general setting, yet, our main purpose is to show that with the graph-based parametrization–as defined in Section E–no border artifact is present in both kernels in this general setting.

The following proposition illustrates the border artifact present in the usual NTK-parametrization. Let us consider a DC-NN with a standardized ReLU nonlinearity, with $I_0 = I_1 \ldots = \mathbb{N}$, with up-sampling stride of 2, and windows $\pi_0 = \omega_0 = \pi_1 = \omega_1 = \ldots = \{-3,-2,-1,0\}$. In particular, there is only one border at position 0. Using the formalism of Section E, the set of parents of a position p is $P(p) = \{\lfloor\frac{p}{2}\rfloor - 1, \lfloor\frac{p}{2}\rfloor\} \cap \mathbb{N}$. In particular, any generic position in any hidden or last layer has 2 parents except for the border $p = 0$ for which $P(0) = \{0\}$.

Proposition 22 *In the setting introduced above, for any $x \in \mathbb{S}_{n_0}^{I_0}$, the kernels satisfy:*

$$\Sigma^{(\ell,00)}(x,x) = \frac{\beta^2 + \left(\frac{r}{2}\right)^{\ell+1}}{1-\frac{r}{2}} \text{ and } \Theta_{\infty}^{(L,00)}(x,x) = \frac{\beta^2\left(1-\left(\frac{r}{2}\right)^{L}\right)}{\left(1-\frac{r}{2}\right)^2} + L\frac{\left(\frac{r}{2}\right)^{L+1}}{1-\frac{r}{2}}.$$

In particular $\Sigma^{(\ell,00)}(x,x)$ is smaller than the "bulk-value" $\lim_{p\to\infty}\Sigma^{(\ell,pp)}(x,x) = 1$ and $\Theta_{\infty}^{(L,00)}(x,x)$ is smaller than the "bulk-value" $\lim_{p\to\infty}\Theta_{\infty}^{(L,pp)}(x,x) = \frac{1-r^L}{1-r}$.

Proof Recall that for the standardized ReLU, $r_{\sigma,\beta} = 1 - \beta^2$. From now on, we denote $r = r_{\sigma,\beta}$ and x is an element of $\mathbb{S}^{I_0}_{n_0}$. For any $\ell = 0, 1 \ldots$, we have:

$$\Sigma^{(\ell+1,00)}(x,x) = \beta^2 + \frac{1-\beta^2}{2} \sum_{q \in P(0)} \mathbb{E}_{z \sim \mathcal{N}(0, \Sigma^{(\ell)}_{qq}(x,x))} \left[\sigma(x)^2\right] = \beta^2 + \frac{1-\beta^2}{2} \Sigma^{(\ell,00)}(x,x).$$

Since $x \in \mathbb{S}^{I_0}_{n_0}$, we get $\Sigma^{(1)}(x,x) = \beta^2 + \frac{r}{2}$: this implies the following equalities:

$$\begin{aligned}\Sigma^{(\ell,00)}(x,x) &= \left(\frac{r}{2}\right)^{\ell} + \sum_{k=0}^{\ell-1} \beta^2 \left(\frac{r}{2}\right)^{k} = \left(\frac{r}{2}\right)^{\ell} + \beta^2 \frac{1 - \left(\frac{r}{2}\right)^{\ell}}{1 - \frac{r}{2}} \\ &= \frac{\beta^2}{1 - \frac{r}{2}} + \frac{\left(\frac{r}{2}\right)^{\ell} - \left(\frac{r}{2}\right)^{\ell+1} - \beta^2 \left(\frac{r}{2}\right)^{\ell}}{1 - \frac{r}{2}} = \frac{\beta^2 + \left(\frac{r}{2}\right)^{\ell+1}}{1 - \frac{r}{2}}.\end{aligned}$$

For the limiting NTK, with the usual NTK parametrization, the following recursion holds:

$$\Theta^{(L+1,00)}_{\infty}(x,x) = \Sigma^{(L+1,00)}(x,x) + \frac{r}{2} \Theta^{(L,00)}_{\infty}(x,x) \mathbb{L}^{\dot{\sigma}}{}_{\Sigma^{(L,00)}}(x,x).$$

Note that for the standardized ReLU, $\dot{\sigma}$ is a rescaled Heaviside, thus

$$\mathbb{L}^{\dot{\sigma}}{}_{\Sigma^{(L,00)}}(x,x) = \mathbb{E}_{x \sim \mathcal{N}(0, \Sigma^{(L,00)}(x,x))} \left[\dot{\sigma}(x)^2\right] = 2\mathbb{E}_{x \sim \mathcal{N}(0,1)}[\mathbb{I}_{x \geq 0}] = 1.$$

This implies:

$$\begin{aligned}\Theta^{(L,00)}(x,x) &= \sum_{\ell=1}^{L} \Sigma^{(\ell,00)}(x,x) \left(\frac{r}{2}\right)^{L-\ell} = \sum_{\ell=1}^{L} \left(\frac{\beta^2}{1-\frac{r}{2}} + \frac{\left(\frac{r}{2}\right)^{\ell+1}}{1 - \frac{r}{2}} \right) \left(\frac{r}{2}\right)^{L-\ell} \\ &= \frac{\beta^2 (1 - \left(\frac{r}{2}\right)^{L})}{\left(1 - \frac{r}{2}\right)^2} + L \frac{\left(\frac{r}{2}\right)^{L+1}}{1 - \frac{r}{2}}.\end{aligned}$$

The "bulk-values" for the activation kernels and the limiting NTK kernel can be deduced from the proof of Proposition 23. A tedious study of variation of functions allows to prove the assertion on the boundary/bulk comparison. ■

As a consequence of the previous proposition, in the limits as ℓ and L goes to infinity, the ratio boundary/bulk value is bounded by $\max\left(1, c\beta^2\right)$: the smaller β is, the stronger the boundary effect will be.

In the graph-based parametrization, the variance of the neurons throughout the network is always equal to 1 and the NTK $\Theta^{(L)}_{\infty,pp}(x,x)$ becomes independent of the position p: the border artifacts disappear.

Proposition 23 (Proposition 8 in the main) *For the graph-based parametrization of DC-NNs, if the nonlinearity is standardized, $\left(\Sigma^{(L)}\right)_{pp}(x)$ and $\left(\Theta^{(L)}_{\infty}\right)_{pp}(x)$ do not depend neither on $p \in I_L$ nor on $x \in \mathbb{S}^{I_0}_{n_0}$.*

Proof Actually, we will prove the stronger statement: for any General Convolutional Network, as defined in Section E, for any standardized nonlinearity, for any $x \in \mathbb{S}_{n_0}^{I_0}$ and any $p \in I_L$,

$$\Sigma^{(L,pp)}(x,x) = 1, \quad \text{and } \Theta_\infty^{(L,pp)}(x,x) = \frac{1-r^L}{1-r}.$$

For the activation kernels,this is proven by induction on ℓ . For any $x \in \mathbb{S}_{n_0}^{I_0}$ and any $p \in I_1$:

$$\begin{aligned}\Sigma^{(1,pp)}(x,x) &= \beta^2 + \frac{1-\beta^2}{|P(p)|\, n_0} \sum_{q\in P(p)} \sum_{q'\in P(p)} \chi(q \to p, q' \to p) x_q^T x_{q'} \\ &= \beta^2 + \frac{1-\beta^2}{|P(p)|\, n_0} \sum_{q\in P(p)} x_q^T x_q = \beta^2 + (1-\beta^2) = 1,\end{aligned}$$

and if the assertion holds true for L, then:

$$\begin{aligned}\Sigma^{(L+1,pp)}(x,x) &= \beta^2 + \frac{1-\beta^2}{|P(p)|\, n_0} \sum_{q\in P(p)} \sum_{q'\in P(p)} \chi(q \to p, q' \to p) \Sigma^{(L,qq')}(x,x) \\ &= \beta^2 + \frac{1-\beta^2}{|P(p)|\, n_0} \sum_{q\in P(p)} \Sigma^{(L,qq)}(x,x) = 1.\end{aligned}$$

For the activation kernels, this is proven by induction on L. It is easy to see that $\Theta_\infty^{(1,pp)}(x,x) = 1$ is valid for any $x \in \mathbb{S}_{n_0}^{I_0}$ and any $p \in I_L$. Let us show the induction step:

$$\begin{aligned}\Theta_\infty^{(L+1,pp)}(x,x) &= \Sigma^{(L+1,pp)}(x,x) + \frac{1-\beta^2}{|P(p)|} \sum_{q\in P(p)} \Theta_\infty^{(L,qq)}(x,x) R_{\dot\sigma}\left(\Sigma^{(L,qq)}(x,x)\right) \\ &= 1 + r\Theta_\infty^{(L,qq)}(x,x).\end{aligned}$$

Thus, $\Theta_\infty^{(L,pp)}(x,x) = \sum_{\ell=1}^{L} r^{L-\ell} = \frac{1-r^L}{1-r}$. ■

Appendix H. Layerwise Contributions to the NTK and Checkerboard Patterns

In a DC-NN with stride $s \in \{2,3,...\}^d$, two connection weight matrices $W^{(\ell,q\to p)}$ and $W^{(\ell,q'\to p')}$ are shared if and only if $s \mid p'-p$. That is, $\chi(q \to p, q' \to p') = 0 \Leftrightarrow s \nmid p'-p$. The limiting contribution of the weights $\Theta_\infty^{(L:W^{(\ell)})}$ and bias $\Theta_\infty^{(L:b^{(\ell)})}$ to the limiting NTK can be formulated recursively. For the last layer $L-1$ we have

$$\begin{aligned}\Theta_\infty^{(L:b^{(L-1)},pp')} &= \beta^2 \\ \Theta_\infty^{(1:W^{(0)},pp')} &= \delta_{s|p-p'} \frac{1-\beta^2}{|P(p)|\, n_0} \sum_{q\in P(p)} x_q^T y_{q+\frac{p'-p}{s}} \\ \Theta_\infty^{(L:W^{(L-1)},pp')} &= \delta_{s|p-p'} \frac{1-\beta^2}{|P(p)|} \sum_{q\in P(p)} R_\sigma\left(\Sigma^{(L-1,q,q+\frac{p'-p}{s})}(x,y)\right) \text{ for } L>1\end{aligned}$$

and for the other layers, we have

$$\Theta_\infty^{(L+1:b^{(\ell)},pp')} = \delta_{s|p-p'}\frac{1-\beta^2}{|P(p)|}\sum_{q\in P(p)}\Theta_\infty^{(L;b^{(\ell)},q,q+\frac{p'-p}{s})}(x,y)R_{\dot\sigma}\left(\Sigma^{(L,q,q+\frac{p'-p}{s})}(x,y)\right)$$

$$\Theta_\infty^{(L+1:W^{(\ell)},pp')} = \delta_{s|p-p'}\frac{1-\beta^2}{|P(p)|}\sum_{q\in P(p)}\Theta_\infty^{(L;W^{(\ell)},q,q+\frac{p'-p}{s})}(x,y)R_{\dot\sigma}\left(\Sigma^{(L,q,q+\frac{p'-p}{s})}(x,y)\right).$$

Proposition 24 (Proposition 10 in the main) *In a DC-NN with stride $s \in \{2,3,...\}^d$, we have $\Theta_\infty^{(L:W^{(\ell)},pp')}(x,y) = 0$ if $s^{L-\ell} \nmid p'-p$ and $\Theta_\infty^{(L:b^{(\ell)},pp')}(x,y) = 0$ if $s^{L-\ell-1} \nmid p'-p$.*

Proof From the formulas of the limiting contributions $\Theta^{(L:W^{(\ell)})}$ and $\Theta^{(L:b^{(\ell)})}$, we see that the bias of the last layer contribute to all pairs p,p' while the bias only contribute to pairs such that $s \mid p'-p$. Now by induction on L, if $\Theta^{(L:b^{(\ell)},qq')}$ and $\Theta^{(L:W^{(\ell)},qq')}$ only contribute to pairs q,q' such that $s^{L-\ell-1} \mid q'-q$ and $s^{L-\ell} \mid q'-q$ then

$$\Theta_\infty^{(L+1:b^{(\ell)},pp')} = \delta_{s|p-p'}\frac{1-\beta^2}{|P(p)|}\sum_{q\in P(p)}\Theta_\infty^{(L;b^{(\ell)},q,q+\frac{p'-p}{s})}(x,y)R_{\dot\sigma}\left(\Sigma^{(L,q,q+\frac{p'-p}{s})}(x,y)\right)$$

$$\Theta_\infty^{(L+1:W^{(\ell)},pp')} = \delta_{s|p-p'}\frac{1-\beta^2}{|P(p)|}\sum_{q\in P(p)}\Theta_\infty^{(L;W^{(\ell)},q,q+\frac{p'-p}{s})}(x,y)R_{\dot\sigma}\left(\Sigma^{(L,q,q+\frac{p'-p}{s})}(x,y)\right)$$

only contribute to pairs p',p such that $s^{L-\ell} \mid p'-p$ and $s^{L+1-\ell} \mid p'-p$ as needed. ■