DISENTANGLING TRAINABILITY AND GENERALIZATION IN DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A fundamental goal in deep learning is the characterization of trainability and generalization of neural networks as a function of their architecture and hyperparameters. In this paper, we discuss these challenging issues in the context of wide neural networks at large depths where we will see that the situation simplifies considerably. To do this, we leverage recent advances that have separately shown: (1) that in the wide network limit, random networks before training are Gaussian Processes governed by a kernel known as the Neural Network Gaussian Process (NNGP) kernel, (2) that at large depths the spectrum of the NNGP kernel simplifies considerably and becomes "weakly data-dependent", and (3) that gradient descent training of wide neural networks is described by a kernel called the Neural Tangent Kernel (NTK) that is related to the NNGP. Here we show that by combining the in the large depth limit the spectrum of the NTK simplifies in much the same way as that of the NNGP kernel. By analyzing this spectrum, we arrive at a precise characterization of trainability and generalization across a range of architectures including Fully Connected Networks (FCNs) and Convolutional Neural Networks (CNNs). We find that there are large regions of hyperparameter space where networks will train but will fail to generalize, in contrast with several recent results. By comparing CNNs with- and without-global average pooling, we show that CNNs without average pooling have very nearly identical learning dynamics to FCNs while CNNs with pooling contain a correction that alters its generalization performance. We perform a thorough empirical investigation of these theoretical results and finding excellent agreement on real datasets.

1 INTRODUCTION

Machine learning models based on deep neural networks have attained state-of-the-art performance across a dizzying array of tasks including vision (Cubuk et al., 2019), speech recognition (Park et al., 2019), machine translation (Bahdanau et al., 2014), chemical property prediction Gilmer et al. (2017), diagnosing medical conditions Raghu et al. (2019), and playing games Silver et al. (2018). Historically, the rampant success of deep learning models has lacked a sturdy theoretical foundation; architectures, hyperparameters, and learning algorithms are more often than not selected by brute force search Bergstra & Bengio (2012) and heuristics Glorot & Bengio (2010). Recently, significant theoretical progress has been made on several fronts that have shown promise in making neural network design more systematic. In particular, in the infinite width (or channel) limit, the distribution of functions induced by neural networks with random weights and biases has been precisely characterized before, during, and after training.

The study of infinite networks dates back to seminal work by Neal (1994) who showed that the distribution of functions given by single hidden-layer networks with random weights and biases in the infinite-width limit are Gaussian Processes (GPs). Recently, there has been renewed interest in studying random, infinite, networks starting with concurrent work on "conjugate kernels" (Daniely et al., 2016; Daniely, 2017) and "mean-field theory" (Poole et al., 2016; Schoenholz et al., 2017). The former set of papers argued that the empirical covariance matrix of pre-activations became deterministic in the infinite-width limit and called this the conjugate kernel of the network while the latter papers studied the properties of these limiting kernels along with the kernel describing distribution of gradients. In particular, it was shown that the spectrum of the conjugate kernel of wide fully-connected networks approached a well-defined, data-independent, limit when the depth

exceeds a certain scale, ξ. Networks with $tanh$-nonlinearities (among other bounded activations) exhibit a phase transition between two limiting spectral distributions of the conjugate kernel as a function of their hyperparameters with ξ diverging at the transition. It was additionally hypothesized that networks were un-trainable when the conjugate kernel was sufficiently close to its limit.

Since then this analysis has been pushed to a wide range for architectures such as convolutions (Xiao et al., 2018), recurrent networks (Chen et al., 2018; Gilboa et al., 2019), networks with residual connections (Yang & Schoenholz, 2017), networks with quantized activations (Blumenfeld et al., 2019), the spectrum of the fisher (Karakida et al., 2018), a range of activation functions Hayou et al. (2018), and batch normalization (Yang et al., 2019). In each case, it was observed that the spectra of the kernels correlated strongly with whether or not the architectures were trainable. While these papers studied the properties of the conjugate kernels, especially the spectrum in the large-depth limit, another branch of concurrent work made a stronger statement: that fully-connected networks (Lee et al., 2018; Matthews et al., 2018), convolutional networks (Novak et al., 2019), and then more generally architectures that could be mapped to "tensor programs" (Yang, 2019) converged to Gaussian Processes. In this case, the Conjugate Kernel was known as the Neural Network Gaussian Process (NNGP) kernel.

Together this work offered a significant advance to our understanding of wide neural networks; however, this theoretical progress was limited to networks at initialization or after Bayesian posterior estimation and provided no link to gradient descent. Moreover, there was some preliminary evidence that suggested the situation might be more nuanced than the qualitative link between the NNGP spectrum and trainability might suggest. For example, Philipp et al. (2017) showed that deep fully-connected $tanh$-networks could be trained after the kernel reached its large-depth, data-independent, limit but that these networks did not generalize to unseen data.

In the last year, significant theoretical clarity has been reached regarding the relationship between the GP prior and the distribution following gradient descent. In particular, Jacot et al. (2018) along with followup work (Lee et al., 2019; Chizat et al., 2019) showed that the distribution of functions induced by gradient descent for infinite-width networks is a Gaussian Process with a particular compositional kernel known as the Neural Tangent Kernel (NTK). In addition to characterizing the distribution over functions following gradient descent in the wide network limit, the learning dynamics can be solved analytically throughout optimization.

In this paper, we leverage these developments and revisit the relationship between architecture, hyperparameters, trainability, and generalization in the large-depth limit for a variety of neural networks. In particular, we make the following contributions:

1. We compute the large-depth asymptotics of several quantities key quantities related to trainability, including the largest eigenvalue of the NTK, λ_{max}, and the condition number $\kappa = \lambda_{\text{max}}/\lambda_{\text{min}}$, where λ_{min} is the smallest eigenvalue; see the table in Section C.

2. We examine the "eigenvector correlation", namely the overlap between the eigenvectors of the NTK evaluated on the training set and the right singular vectors of the test-train NTK, which is related to the model's ability to generalize.

3. We show that the *ordered* and *chaotic* phases identified in Poole et al. (2016) lead to markedly different limiting spectra of the NTK.

4. We examine the differences in the above quantities for fully-connected networks (FCNs) and convolutional networks (CNNs) with and without pooling.

5. We provide substantial experimental evidence supporting these claims, includes experiments that densely vary the hyperparameters of FCNs and CNNs with and without pooling.

Together these results provide a complete, analytically tractable, and dataset-independent theory for learning in very deep and wide networks. In addition to being interesting in its own right our theory provides a strong test of the NTK theory. Finally, our results provides clarity regarding the observation that for linear networks the learning rate must be decreased linearly in the depth of the network Saxe et al. (2013). Here, we note that this is true only for networks that are initialized *critically*, i.e. on the order-to-chaos phase boundary.

2 BACKGROUND

We summarize recent developments in the study of wide random networks. We will keep our discussion relatively informal; see (Lee et al., 2018; Matthews et al., 2018; Novak et al., 2019) for a more rigorous version of these arguments. To simplify this discussion and as a warmup for the main text, we will consider the case of FCNs. Consider a fully-connected network of depth L where each layer has a width N_l and an activation function $\phi : \mathbb{R} \to \mathbb{R}$. In this work we will take $\phi = \text{erf}$ however, most of the results will hold for a wide range of non-linearities though specifics - such as the phase diagram - can vary substantially. For simplicity we will take $N_l = N$ to be a layer-independent constant that is much larger than the depth and the dataset size. The network is parameterized by weights and biases that we take to be randomly initialized with $W^l_{ij}, b^l_i \sim \mathcal{N}(0, 1)$ along with hyperparameters, σ_w and σ_b that set the scale of the weights and biases. Letting the pre-activations in layer l due to an input x be given by $z^l_i(x)$, the network is then described by the recursion,

$$z^{l+1}_i(x) = \frac{\sigma_w}{\sqrt{N}} \sum_j W^l_{ij} \phi(z^l_j(x)) + \sigma_b b^l_i \qquad 1 \le l \le L. \tag{1}$$

Notice that as the width goes to infinity, the sum ends up being over a large number of random variables and we can invoke the central limit theorem to conclude that the z^l_i are Gaussian distributed with zero mean and zero covariance between different neurons. Given a dataset of M points, the distribution over pre-activations can therefore be described completely by the covariance matrix between neurons in different inputs $\mathcal{K}^l(x, x') = \mathbb{E}[z^l_{ia}(x) z^l_{ib}(x')]$. Inspecting Equation 1, we see that $\mathcal{K}^l(x, x')$ can be computed in terms of $\mathcal{K}^{l-1}(x, x')$ as

$$\mathcal{K}^l(x, x') = \sigma^2_w \mathbb{E}_{(z,z') \sim \mathcal{N}(0, \mathcal{K}(x,x'))}[\phi(z)\phi(z')] + \sigma^2_b = \mathcal{T}(\mathcal{K}(x, x')) \tag{2}$$

for $\mathcal{T}$, an appropriately defined point-wise nonlinear function on $\mathcal{K}(x, x')$.

Equation 2 describes a dynamical system on $\mathcal{K}(x, x')$. It was shown in Poole et al. (2016) that fixed points, $\mathcal{K}^*(x, x')$, of these dynamics exist such that $\lim_{l\to\infty} \mathcal{K}^l(x, x') \to \mathcal{K}^*(x, x')$ with $\mathcal{K}^*(x, x') = q^*[\delta_{x,x'} + c^*(1 - \delta_{x,x'})]$ independent of the inputs x and x'. The values of q^* and c^* are determined by the hyperparameters, σ_w and σ_b. However Equation 2 admits multiple fixed points (e.g. $c^* = 0, 1$) and the stability of these fixed points plays a significant role in determining the properties of the network. Generically, there are large regions of the (σ_w, σ_b) plane in which the fixed-point structure is constant punctuated by curves, called phase transitions, where the structure changes. The rate at which $\mathcal{K}(x, x')$ approaches $\mathcal{K}^*(x, x')$ can be determined by expanding Equation 2 about its stable fixed point, $\delta\mathcal{K}(x, x') = \mathcal{K}(x, x') - \mathcal{K}^*(x, x')$ to find

$$\delta\mathcal{K}^{l+1}(x, x') = \sigma^2_w \dot{\mathcal{T}}(\mathcal{K}^*(x, x'))\delta\mathcal{K}^l(x, x') \tag{3}$$

which exhibits exponential convergence as $\delta\mathcal{K}(x, x') \sim e^{-l/\xi(x,x')}$ over a depth $\xi(x, x') = -1/\log(\sigma^2_w \dot{\mathcal{T}}(\mathcal{K}^*(x, x')))$. Since $\mathcal{K}^*(x, x')$ does not depend on x or x' it follows that $\xi(x, x')$ can take on only two values, ξ_q when $x = x'$ and ξ_{c^*} otherwise.

This provides a precise characterization of the spectrum of the NNGP kernel at large depths. As discussed above, recent work (Jacot et al., 2018; Lee et al., 2019; Chizat et al., 2019) has connected the prior described by the NNGP with the result of gradient descent training using a quantity called the NTK. To construct the NTK, suppose we enumerate all the parameters in the fully-connected network described above by θ_α. The NTK is defined by $\Theta(x, x') = J(x)J(x')^T$ where $J_{i\alpha}(x) = \partial_{\theta_\alpha} z^L_i(x)$ is the Jacobian evaluated at a point x. The main result in Jacot et al. (2018) was to show that in the infinite-width limit, the NTK becomes deterministic and remains constant over the course of training. This implies that at a time t during gradient descent training with an MSE loss, the expected outputs of the network, $\mu_t(x) = \mathbb{E}[z^L_i(x)]$, evolve as

$$\mu(X_{\text{train}}) = (I - e^{-\eta\Theta(X_{\text{train}}, X_{\text{train}})t})Y_{\text{train}} \tag{4}$$

$$\mu(X_{\text{test}}) = \Theta(X_{\text{test}}, X_{\text{train}})\Theta(X_{\text{train}}, X_{\text{train}})^{-1}(I - e^{-\eta\Theta(X_{\text{train}}, X_{\text{train}})t})Y_{\text{train}} \tag{5}$$

for train and test points respectively. As the training time, t tends to infinity we note that these equations reduce to $\mu(X_{\text{train}}) = Y_{\text{train}}$ and $\mu(X_{\text{test}}) = \Theta(X_{\text{test}}, X_{\text{train}})\Theta(X_{\text{train}}, X_{\text{train}})^{-1}Y_{\text{train}}$. Consequently we call $\Theta(X_{\text{test}}, X_{\text{train}})\Theta(X_{\text{train}}, X_{\text{train}})^{-1}$ the "mean predictor". In addition to showing

that the NTK describes networks during gradient descent, Jacot et al. (2018) showed that the NTK could be computed in closed form in terms of $\mathcal{T}$, $\dot{\mathcal{T}}$, and the NNGP as,

$$\Theta^{(l+1)}(x, x') = \sigma_w^2 \mathcal{T}(\mathcal{K}^{(l)}(x, x')) + \sigma_b^2 + \sigma_w^2 \dot{\mathcal{T}}(\mathcal{K}^{(l)}(x, x'))\Theta^{(l)}(x, x'). \tag{6}$$

where $\Theta^{(l)}$ is the NTK for the pre-activations at layer-l.

3 Large-depth asymptotics of the NNGP and NTK

In the large-width limit, $\Theta^{(l)}$ converges to Θ^* independent of the inputs to the network. As such, the mean prediction defined by Equation 5 completely fails to generalize. The fundamental quantity that captures the generalization is therefore the finite depth correction to the infinite depth predictor

$$\Delta^{(l)}\mathcal{Y} \equiv \left(\Theta^{(l)}_{\text{test,train}} \left(\Theta^{(l)}_{\text{train,train}}\right)^{-1} - \Theta^{*}_{\text{test,train}} \left(\Theta^{*}_{\text{train,train}}\right)^{-1}\right) \mathcal{Y} \tag{7}$$

We follow the methodology outlined in Schoenholz et al. (2017); Xiao et al. (2018) to explore the spectrum of the NTK as a function of depth. We will use this to make precise predictions relating trainability and generalization to the hyperparameters σ_w and σ_b. In order to simplifty the notation, we mainly focus on the fully-connected network setting and then extend the results to CNN with pooling (CNN-P) and without pooling (CNN-F). Details can be found in the appendix.

To gain insight into Equation 7, we now analyze the large depth behavior of the NNGP and the NTK. We will focus on the NTK here since Xiao et al. (2018) contains a detailed description of the NNGP in this case. As in sec. 2, we will be concerned with the fixed points of Θ as well as the linearization of Equation 6 about its fixed point. Recall that the fixed point structure is fixed within a phase so it suffices to consider the ordered phase, the chaotic phase, and the critical line separately. In cases where a stable fixed point exists, we will describe how Θ converges to the fixed point. We will see that in the chaotic phase and on the critical line, Θ has no stable fixed point and in that case we will describe its divergence. As above, in each case the fixed points of Θ have a simple structure with $\Theta^* = q^*((1 - c^*)I + c^*\mathbf{1}\mathbf{1}^T)$. To simplify the forthcoming analysis, without a loss of generality, we assume the inputs are normalized to have variance q^*. As such, we can treat $\mathcal{T}$ and $\dot{\mathcal{T}}$, restricted on $\{\mathcal{K}^{(l)}\}_l$, as a point-wise functions, since

$$\mathcal{T}(\mathcal{K})(x, x') = \mathbb{E}\phi(u)\phi(v), \quad (u, v)^T \sim \mathcal{N}\left(0, \begin{bmatrix} q^* & \mathcal{K}(x, x') \\ \mathcal{K}(x, x') & q^* \end{bmatrix}\right). \tag{8}$$

Since the off-diagonal elements approach the same fixed point at the same rate, we use $q^{(l)}_{ab}$ and $p^{(l)}_{ab}$ to denote any off diagonal entry of $\mathcal{K}^{(l)}$ and $\Theta^{(l)}$ respectively. We will similarly use q^*_{ab} and p^*_{ab} to denote the limits, $\lim_{l\to\infty} q^{(l)}_{ab} = q^*_{ab}$ and $\lim_{l\to\infty} p^{(l)}_{ab} = p^*_{ab}$. In what follows, we split the discussion into three parts according to the values of $\chi_1 \equiv \sigma_\omega^2 \mathcal{T}(q^*) + \sigma_b^2$ recalling that in Poole et al. (2016); Schoenholz et al. (2017) it was shown that χ_1 controls the fixed point structure.

3.1 The Chaotic Phase $\chi_1(\sigma_\omega, \sigma_b) > 1$:

The chaotic phase is so-named because $q^*_{ab}/q^* < 1$ so that similar inputs become more uncorrelated as they pass through the network. In this phase, the diagonal entries of Θ grow exponentially and the off-diagonal entries converge to a fixed value. Indeed, Equation 6 implies that the diagonal entries have the large-depth limit,

$$p^{(l+1)} = q^* + \chi_1 q^* p^{(l)} \qquad \Longrightarrow \qquad p^{(l)} = q^* \frac{\chi_1^{l+1} - 1}{\chi_1 - 1}, \tag{9}$$

which diverges exponentially. To find the limit of the off-diagonal terms, let $l \to \infty$ in Equation 6, we find that $p^*_{ab} = \frac{q^*_{ab}}{1-\chi_c} < \infty$. Here we have defined $\chi_c = \sigma_\omega^2 \dot{\mathcal{T}}(q^*_{ab})$ which was shown to control convergence of the NNGP kernel in (Schoenholz et al., 2017; Xiao et al., 2018). The rate of convergence of p^*_{ab} is $\mathcal{O}(l\chi_c^l)$ (see Equation 20 in the appendix). Since the diagonal terms diverge and the off-diagonal terms are finite it follows that in very deep networks in the chaotic phase, $(p^{(l)})^{-1}\Theta^{(l)} \to Id$. Thus, the spectrum of the NTK for very deep networks in the chaotic

phase approaches the diverging constant multiplying the identity. From Equation 4 this implies that optimization in the chaotic phase should be easy since the NTK has perfect conditioning (provided numerical precision issues from the prefactor do not become problematic). However, computing the mean prediction on test points we find,

$$\Delta^{(l)}\mathcal{Y} \approx (p^{(l)})^{-1}\left(\Theta^*_{\text{test,train}} + \mathcal{O}(l\chi_c^l) + \mathcal{O}(\chi_1^{-l})\right)\mathcal{Y} = O(\chi_1^{-l})\mathcal{Y} \to \mathbf{0}. \tag{10}$$

It follows that in the chaotic phase the networks predictions on unseen data to converge to 0 exponentially quickly in the depth. In summary, for wide networks, in the chaotic phase as the depth increases optimization becomes increasingly easy (in the sense the condition number of the NTK approaches 1). However, the generalization performance degrades and eventually the network fails completely away from the training set. Since the data dependent term in 10 decays like $(p^{(l)})^{-1}\left(\mathcal{O}(l\chi_c^l) + \mathcal{O}(\chi_1^{-l})\right)$, we expect the network incapable to generalize after $\mathcal{O}(\xi_*)$ layers [1], where $\xi_* = -1/(\log\chi_c - \log\chi_1)$. We will confirm this prediction in the experimental results to follow.

3.2 The Ordered Phase $\chi_1 = \sigma_\omega^2\dot{\mathcal{T}}(q^*) < 1$:

The ordered phase is defined by the stable fixed point with $q^*_{ab}/q^* = 1$; in this case, disparate inputs will end up converging to the same output at the end of the network. In the ordered phase, Equation 9 implies that all the diagonal entries of Θ converge to the same value,

$$p^{(l)} = q^*\frac{\chi_1^{l+1}-1}{\chi_1 - 1} \xrightarrow{l\to\infty} q^*\frac{1}{1-\chi_1} < \infty \tag{11}$$

However, as with the NNGP kernel, the off-diagonal terms of the NTK, $p^{(l)}_{ab}$, will also converge to the value on the diagonal, p^*_{ab}. It follows that the limiting kernels have the form $\Theta^* = p^*\mathbf{1}\mathbf{1}^\mathbf{T}$ and $\mathcal{K}^* = q^*\mathbf{1}\mathbf{1}^\mathbf{T}$. Thus, the limiting kernels are highly singular and feature only one non-zero eigenvalue. Since the limit is singular, we must linearize the dynamics about the fixed point to gain insight into the limiting behavior of the network. To compute the corrections, let

$$\epsilon^{(l)}_{ab} = q^{(l)}_{ab} - q^*_{ab} \qquad \delta^{(l)}_{ab} = p^{(l)}_{ab} - p^*_{ab} \tag{12}$$

$$\epsilon^{(l)} = q^{(l)} - q^* \qquad \delta^{(l)} = p^{(l)} - p^* \tag{13}$$

The diagonal correction can be obtained directly from Equation 11 and we find that $\epsilon^{(l)} = 0$ and $\delta^{(l)} = \frac{\chi_1^{l+1}}{1-\chi_1}$. To compute correction of the off-diagonals, we linearize the equation around the fixed point to find that asymptotically,

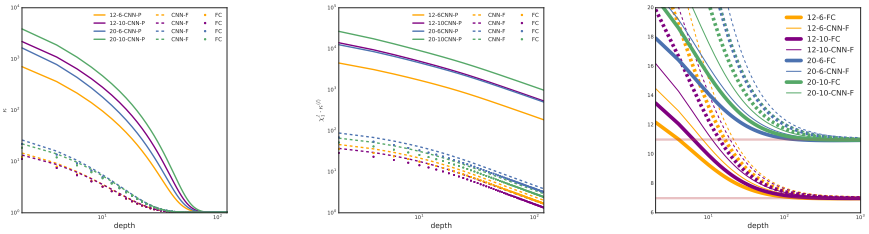
$$\epsilon^l_{ab} \approx \chi_1^l\epsilon^0_{ab} \qquad \delta^l_{ab} \approx \chi_1^l\left[\delta^0_{ab} + l\left(1 + \frac{\chi_2}{\chi_1}p^*_{ab}\right)\epsilon^0_{ab}\right] \tag{14}$$

where $\chi_2 = \sigma_\omega^2\ddot{\mathcal{T}}(p^*_{ab})$. While the NNGP and NTK feature the same exponential rate of convergence set by χ_1, we see that terms in the off-diagonal terms of the NTK feature polynomial corrections. Θ has (approximately) two eigenspaces. The first eigenspace comes from the single non-zero eigenvalue at the fixed point and it is very close to the DC mode with eigenvalue

$$\lambda^{(l)}_{\max} \approx (m-1)(p^*_{ab} - \delta^{(l)}_{ab}) + (p^*_{ab} - \delta^{(l)}) \to mp^*_{ab} = \frac{mq^*}{1-\chi_1} \tag{15}$$

i.e. is the sum of one row, where m is the batch size. The second eigenspace comes from lifting the degenerate zero-modes when $l < \infty$ and it has dimension $(m-1)$ with eigenvalue $\lambda^{(l)}_{\text{rest}} \approx -\delta^{(l)} + \delta^{(l)}_{ab} = \mathcal{O}(l\chi_1^l) \to 0$, which goes to zero exponentially over a depth χ_1. The eigenvalues of $\mathcal{K}^{(l)}$ have a similar distribution with $\lambda^{(l)}_{\max} \approx mq^* - (m-1)\epsilon^{(l)}_{ab}$ and $\lambda^{(l)}_{\text{rest}} = O(\chi_1^l)$. Thus the conditioning number, κ, of both $\Theta^{(l)}$ and $\mathcal{K}^{(l)}$ diverges exponentially as $\mathcal{O}(\chi_1^{-l}l^{-1})$ and $\mathcal{O}(\chi_1^{-l})$ respectively. As discussed above, there is a polynomial correction in the conditioning number of the NTK that slightly improves its conditioning. Since Θ^* is singular, we insert a diagonal regularization σId into both predictors in Equation 7, where σ is a positive constant that independent of l and χ_1. We find $\Delta^{(l)} = \mathcal{O}_\sigma(l\chi_1^l)$. Therefore, in the ordered phase, $\xi_1 = -\log\chi_1$ (for simplicity, we ignore the polynomial correction) governs both trainability and generalizability of the predictor .

[1] Preliminary numerical experiments indicate $\chi_c\chi_1 \geq 1$



(a) Chaotic (b) Ordered (c) Critical: FC and CNN-F

Figure 1: Condition numbers of NTKs in different phases. We use different colors represent different image size. For example, in the yellow "12-6" , "12" represent the batch size and "6" represent the dimension number ($6 * 6 * 3$ for FC and $(6, 6, 3)$ for CNN) (a) In the chaotic phase, κ converges to 1 for all architectures. (b) We plot $\chi_1^l \kappa^{(l)}$ confirming κ explodes with rate χ_1^l / l. In (c), the dashed lines representing the condition number and solid lines the ratio between first and second eigenvalues. We see that these two numbers converge to $\frac{m+2}{2}$ for FC and CNN-F; see Figure 4 for CNN-P.

3.3 The Critical Line $\chi_1 = \sigma_\omega^2 \dot{\mathcal{T}}(q^*) = 1$

On the critical line both the diagonal and the off-diagonal terms of $\Theta^{(l)}$ diverge linearly in the depth while $\mathcal{K}$ converges to $q^* \mathbf{1}\mathbf{1}^T$. From Equation 6 we see immediately that the diagonal terms are given by $p^{(l)} = q^*$ and $p^{(l)} = lq^*$. To compute the correction of the off-diagonals, we define $\delta_{ab}^{(l)}$ and $\epsilon_{ab}^{(l)}$ slightly differently to the above as $\delta_{ab}^{(l)} = p_{ab}^{(l)} - lq^*$ and $\epsilon_{ab}^{(l)} = q_{ab}^{(l)} - q^*$ to take into account the linear divergence at large depths. Taylor expanding to second order we find,

$$\epsilon_{ab}^{(l)} = -\frac{2}{\chi_2} \frac{1}{l} + o(\frac{1}{l}), \quad \delta_{ab}^{(l)} = -\frac{2}{3} lq^* + \mathcal{O}(1) \tag{16}$$

Thus for large l, $\Theta^{(l)}$ has the following form $p^{(l)} = lq^*$ and $p_{ab}^{(l)} = \frac{1}{3} lq^* + \mathcal{O}(1)$. As in the ordered phase, for large l it follows that Θ essentially has two eigenspaces: one has dimension one and the other has dimension $(m-1)$ with

$$\lambda_{\text{max}}^{(l)} = \frac{(m+2)q^*}{3} l + m\mathcal{O}(1), \quad \lambda_{\text{rest}}^{(l)} = \frac{2}{3} q^* l + \mathcal{O}(1) \tag{17}$$

and the condition number $\kappa^{(l)} = \frac{m+2}{2} + m\mathcal{O}(l^{-1}) \to \frac{m+2}{2}$ as $l \to \infty$. Unlike the chaotic and ordered phases, $\kappa^{(l)}$ converges with rate $\mathcal{O}(l^{-1})$. The $\mathcal{K}^{(l)}$ has $\lambda_{\text{max}}^{(l)} = mq^* + m\mathcal{O}(l^{-1})$ and $\lambda_{\text{rest}}^{(l)} \approx \frac{2}{\chi_2} l^{-1}$ and the condition number $\kappa^{(l)}$ diverges linearly with slope $m\chi_2/2$. A similar calculation to the chaotic phase gives $\Delta^{(l)} = \mathcal{O}(l^{-1})$.

3.4 The effect of pooling and flattening of CNNs

With the bulk of the theory in hand, we now turn our attention to CNNs. We show in the appendix that the dominant mode in CNNs behaves exactly like the fully-connected case, however we will see that the readout can significantly affect the spectrum. The NNGP and NTK of the l-th hidden layer CNN are 4D tensors $\mathcal{K}_{\alpha,\alpha'}^{(l)}(x, x')$ and $\Theta_{\alpha,\alpha'}^{(l)}(x, x')$, where $\alpha, \alpha' \in [d] \equiv [0, 1, \ldots, d-1]$ denote the pixel locations. To perform tasks like image classification or regression, "flattening" and "pooling" (more precisely, global average pooling) are two popular readout strategies that transform the last convolution layer into the logit layer. The former strategy "flattens" an image of size (d, N) into a vector and stacks a fully-connected layer on top. The latter projects the (d, N) image into a vector of dimension N via averaging out the spatial dimension and then stacks a fully-connected layer on top. The actions of "flattening" and "pooling" on the image correspond to computing the mean of

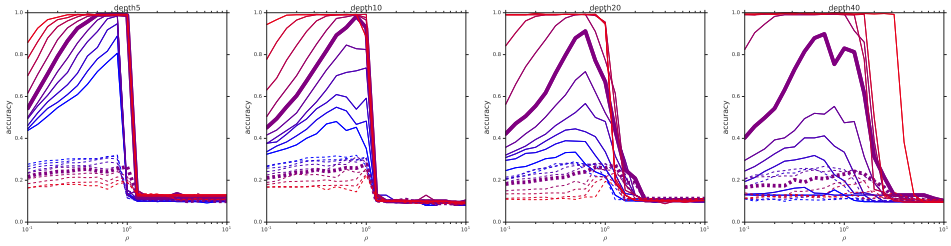


Figure 2: Maximal feasible learning rate can be calculated via the $\lambda_{\max}$ of NTK. The y-axis accuracy and x-axis multiples of η_{theory}. Each point on the solid (dashed) lines represent the best training (test) accuracy throughout training of one configure ration. From blue to purple to red, $(\sigma_\omega, \sigma_b)$ is moving from the order phase to the chaotic phase. $\rho = 1$ is the theoretical prediction.

the trace and the mean of the pixel-to-pixel covariance on the NNGP/NTK, respectively, i.e.,

$$\Theta^{(l)}_{\text{flatten}}(x, x') = \frac{1}{d}\sum_{\alpha\in[d]} \Theta^{(l)}_{\alpha,\alpha}(x, x'), \quad \Theta^{(l)}_{\text{pool}}(x, x') = \frac{1}{d^2}\sum_{\alpha,\alpha'\in[d]} \Theta^{(l)}_{\alpha,\alpha'}(x, x') \tag{18}$$

where $\Theta^{(l)}_{\text{flatten}}$ ($\Theta^{(l)}_{\text{pool}}$) denotes the NTK right after flattening (pooling) the last convolution. We will also use $\Theta^{(l)}_{\text{fc}}$ to denote the NTK of FC. $\mathcal{K}^{(l)}_{\text{flatten}}$ and $\mathcal{K}^{(l)}_{\text{pool}}$ are defined similarly.

As discussed above, in the large depth setting, all the diagonals $\Theta^{(l)}_{\alpha,\alpha}(x, x) = p^{(l)}$ (since the inputs are normalized to have variance q^* for each pixel) and similar to $\Theta^{(l)}_{\text{fc}}$, all the off-diagonals $\Theta^{(l)}_{\alpha',\alpha}(x, x')$ are almost equal (in the sense they have the same order of correction to p^*_{ab} if exists.) Without loss of generality, we assume all off-diagonals are the same and equal to $p^{(l)}_{ab}$ (the leading correction of $q^{(l)}_{ab}$ for CNN and FC are of the same order.) Applying flattening and pooling, the kernels become $\Theta^{(l)}_{\text{flatten}}(x, x') = \mathbf{1}_{x=x'}p^{(l)} + \mathbf{1}_{x\neq x'}p^{(l)}_{ab}$ and $\Theta^{(l)}_{\text{pool}}(x, x') = \frac{1}{d}\mathbf{1}_{x=x'}(p^{(l)} - p^{(l)}_{ab}) + p^{(l)}_{ab}$ respectively. As we can see, $\Theta^{(l)}_{\text{ft}}$ is essentially the same as its FC counterpart $\Theta^{(l)}_{\text{fc}}$ up to sub-dominant Fourier modes which decay exponentially faster than the dominant Fourier modes. Therefore the spectrum properties of $\Theta^{(l)}_{\text{ft}}$ and $\Theta^{(l)}_{\text{fc}}$ are essentially the same. However, pooling alters the NTK/NNGP spectrum in an interesting way. On the critical line, asymptotically, $\lambda^{(l)}_{\max} \approx (md + 2)q^*l/(3d)$ and $\lambda^{(l)}_{\text{rest}} \approx 2q^*l/(3d)$, and $\kappa^{(l)} = \frac{md+2}{2} + md\mathcal{O}(l^{-1})$. Here we use blue color to indicate the changes of such quantities against their $\Theta^{(l)}_{\text{flatten}}$ counterpart. Thus pooling decreases $\lambda^{(l)}_{\text{rest}}$ roughly by a factor of d and increases the condition number by a factor of d comparing to flattening. In the chaotic phase, pooling does not change the off-diagonals $q^{(l)}_{ab} = \mathcal{O}(1)$ but does slow down the growth of the diagonals by a factor of d, $p^{(l)} = \mathcal{O}(\chi_1^l\, d)$. This suggests, in the chaotic phase, there exists a transient regime of depths, where CNN-F hardly perform while CNN-P performs well. In the ordered phase, the pooling does not affect $\lambda^{(l)}_{\max}$ much but does decrease λ_{rest} by a factor of d and the condition number grows approximately like $d\chi_1^{-l}$, d times bigger than its flattening and fully-connected network counterpart. This suggests the existence of a transient regime of depths, in which CNN-F outperforms CNN-P. This might be surprising because it is commonly believed CNN-P usually outperforms CNN-F.

4 Experiments

In this section, we provide empirical results to support the theoretical results in Section 3. All experiments in this section are conducted using CIFAR-10 with MSE as the loss function.

Maximal Feasible Learning Rates (Figure 2 top). To confirm that the maximal feasible learning rates are approximately $\eta_{\text{theory}} = \frac{2}{\lambda_{\max}}$, where $\lambda_{\max}$ is the maximal eigenvalue of the theoretical NTK, we train a Fully-connected network with width 2048 using $1k$ training samples and using SGD, with (1) the std of the bias fixed $\sigma_b = 0.43$, (2) depths: $l = 5, 10, 20, 40$, (3) 10 different values

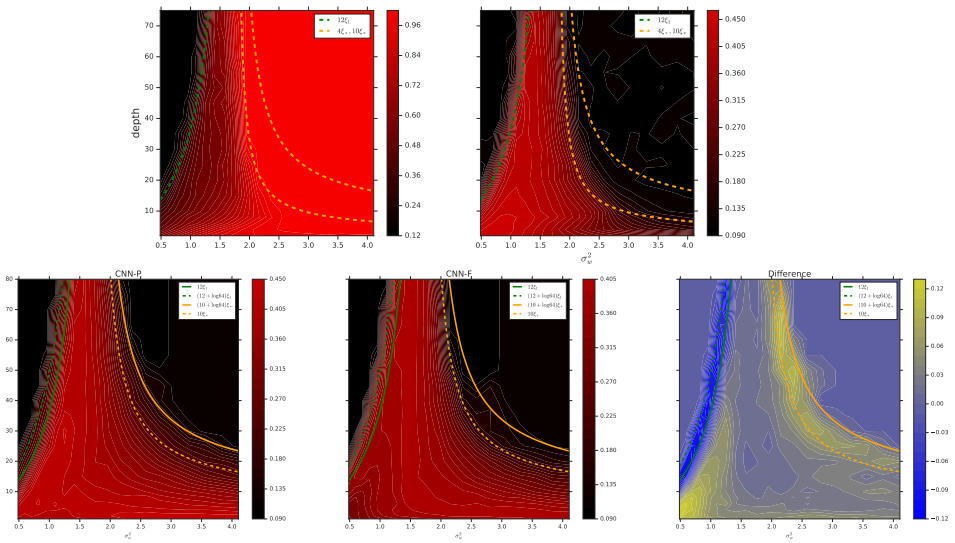


Figure 3: Top: training (left) and test accuracy of FC using SGD. Bottom: test accuracy of CNN-P, CNN-F and the difference. In the blue strip, CNN-F significantly outperforms CNN-P, due to the fact that pooling increases the spectra gap by a factor of d.

of σ_ω moving from the ordered phase (blue) to the chaotic phase (red) (4) 10 different learning rates $\eta = \rho\eta_{\text{theory}}$, with $\rho \in [10^{-1}, 10^{1}]$. Overall, we see excellent agreement for depths less or equal to 20 and reasonable good agreement for depth 40.

Trainability vs Generalization (Figure 3 bottom). Our theoretical result suggests that in the deep chaotic regime (χ_1^l is large) training becomes easier but the network can not generalize. On the other hand, the network can generalize but training becomes much more difficult as one moves towards the deep ordered region because κ blows up exponentially. To confirm this claim, we conduct an experiment using 16k training samples from CIFAR-10 with 20×20 different σ_ω-depth configurations. We train each network using SGD with batch size $b = 1024$ and learning rate $\eta = 0.3\eta_{\text{theory}}$. In the deep chaotic regime, all configurations reach perfect training accuracy but the network become un-generalizable in the sense test accuracy approaches 10%. On the other hand, the network perform much better on the ordered regime although training accuracy is much lower. The two depth scales ξ_1 and ξ_* also perfectly capture the generalizable regime.

CNN-P v.s. CNN-F: spatial correction (Figure 3). We compute the test accuracy using the analytic NTK predictor Equation 5, which corresponds to the test accuracy of ensemble of gradient descent trained neural networks taking the width to infinity. We choose $1k$ train set and choose 20×20 different configurations of σ_ω-depths with σ_b^2 fixed. We layout the test performance of CNN-P and CNN-F and performance difference in Fig 3. Remarkably, the performance of both CNN-P and CNN-F are captured by $\xi_1 = -1/\log(\chi_1)$ in the ordered phase and by $\xi_* = -1/(\log \xi_c - \log \xi_1)$ in the chaotic phase. We also layout the performance difference between CNN-P and CNN-F and identify a blue strip in the ordered phase, as it was predicted exactly by our calculation, where CNN-F outperforms CNN-P by a large margin.

5 Conclusion and Future Work

In this work, we identify several quantities ($\lambda_{\max}$, λ_{rest}, κ, etc) related to the spectrum of the NTK that control trainability and generalization of deep networks. We also provide substantial experimental evidence supporting our claims. In future work, we would like to extend our framework to other architectures, e.g., ResNet, attention model. Understanding the implication of the sub-Fourier modes in the NTK to the test performance of CNN is also an important research direction.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

Yaniv Blumenfeld, Dar Gilboa, and Daniel Soudry. A mean field theory of quantized deep networks: The quantization-depth trade-off. *arXiv preprint arXiv:1906.00771*, 2019.

Minmin Chen, Jeffrey Pennington, and Samuel Schoenholz. Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks. In *International Conference on Machine Learning*, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. 2019.

Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Amit Daniely. SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems 30*. 2017.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, 2016.

Dar Gilboa, Bo Chang, Minmin Chen, Greg Yang, Samuel S. Schoenholz, Ed H. Chi, and Jeffrey Pennington. Dynamical isometry and a mean field theory of lstms and grus. *CoRR*, abs/1901.08987, 2019. URL `http://arxiv.org/abs/1901.08987`.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1263–1272. JMLR.org, 2017. URL `http://dl.acm.org/citation.cfm?id=3305381.3305512`.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the selection of initialization and activation function for deep neural networks. *arXiv preprint arXiv:1805.08266*, 2018.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*. 2018.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: mean field approach. *arXiv preprint arXiv:1806.01316*, 2018.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Sam Schoenholz, Jeffrey Pennington, and Jascha Sohldickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 4 2018. URL `https://openreview.net/forum?id=H1-nGgWC-`.

Radford M. Neal. Priors for infinite networks (tech. rep. no. crg-tr-94-1). *University of Toronto*, 1994.

Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations*, 2019.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

George Philipp, Dawn Song, and Jaime G Carbonell. The exploding gradient problem demystified-definition, prevalence, impact, origin, tradeoffs, and solutions. *arXiv preprint arXiv:1712.05577*, 2017.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances In Neural Information Processing Systems*, pp. 3360–3368, 2016.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning with applications to medical imaging. *arXiv preprint arXiv:1902.07208*, 2019.

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

Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *International Conference on Learning Representations*, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. ISSN 0036-8075. doi: 10.1126/science.aar6404. URL `https://science.sciencemag.org/content/362/6419/1140`.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, 2018.

Ge Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. In *Advances in Neural Information Processing Systems*. 2017.

Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, and Samuel S Schoenholz. A mean field theory of batch normalization. *arXiv preprint arXiv:1902.08129*, 2019.

A APPENDIX

A.1 CORRECTION OF THE OFF-DIAGONALS IN THE CHAOTIC PHASE

Similar to the ordered phase, we have

$$\epsilon_{ab}^{(l+1)} \approx \chi_c \epsilon_{ab}^{(l)}, \quad \delta_{ab}^{(l+1)} \approx (1 + \chi_{c,2} p_{ab}^*)\epsilon_{ab}^{(l+1)} + \chi_c \delta_{ab}^{(l)} \tag{19}$$

where $\chi_{c,2} = \sigma_\omega^2 \ddot{\mathcal{T}}(q_{ab}^*)$. This implies

$$\epsilon_{ab}^{(l)} \approx \chi_c^l \epsilon_{ab}, \quad \delta_{ab}^l \approx \chi_c^l \left[\delta_{ab}^0 + l \left(1 + \frac{\chi_{c,2}}{\chi_c} p_{ab}^* \right) \epsilon_{ab}^0 \right] \tag{20}$$

which also contains a polynomial correction.

A.2 CORRECTION OF THE OFF-DIAGONALS ON THE CRITICAL LINE.

Recall that

$$\epsilon_{ab}^{(l)} = -\frac{2}{\chi_2} \frac{1}{l} + o(\frac{1}{l}). \tag{21}$$

Then

$$\delta_{ab}^{(l+1)} = q_{ab}^{(l+1)} - q^* + \sigma_\omega^2 \dot{\mathcal{T}}(q^* + \epsilon_{ab}^{(l)}) p_{ab}^{(l)} - l q^* \tag{22}$$

$$\approx \epsilon_{ab}^{(l+1)} + (\chi_1 + \chi_2 \epsilon_{ab}^{(l)} + \frac{1}{2}\chi_3 (\epsilon_{ab}^{(l)})^2)(l q^* + \delta_{ab}^{(l)}) - l q* \tag{23}$$

$$\approx \epsilon_{ab}^{(l+1)} + (1 + \chi_2 \epsilon_{ab}^{(l)}) \delta_{ab}^{(l)} + l q^* \chi_2 \epsilon_{ab}^{(l)} + \frac{1}{2}\chi_3 (\epsilon_{ab}^{(l)})^2 l q^* \tag{24}$$

Solving this gives $\delta_{ab}^{(l)} = -\frac{2}{3} l q^* + \mathcal{O}(1)$.

B CONVOLUTIONS

General setup. For simplicity of presentation we consider 1D convolutional networks with circular padding as in Xiao et al. (2018). We will see that this reduces to the fully-connected case introduced above if the image size is set to one and as such we will see that many of the same concepts and equations carry over schematically from the fully-connected case. The theory of two-dimensional convolutions proceeds identically but with more indices.

Random weights and biases. The parameters of the network are the convolutional filters and biases, $\omega_{ij,\beta}^l$ and μ_i^l, respectively, with outgoing (incoming) channel index i (j) and filter relative spatial location $\beta \in [\pm k] \equiv \{-k, \ldots, 0, \ldots, k\}$.[2] As above, we will assume a Gaussian prior on both the filter weights and biases,

$$W_{ij,\beta}^l = \frac{\sigma_\omega}{\sqrt{(2k+1)n^l}} \omega_{ij,\beta}^l \qquad b_i^l = \sigma_b \mu_i^l, \qquad \omega_{ij,\beta}^l, \quad \mu_i^l \sim \mathcal{N}(0,1) \tag{25}$$

As above, σ_ω^2 and σ_b^2 are hyperparameters that control the variance of the weights and biases respectively. n^l is the number of channels (filters) in layer l, $2k+1$ is the filter size.

Inputs, pre-activations, and activations. Let $\mathcal{X}$ denote a set of input images. The network has activations $y^l(x)$ and pre-activations $z^l(x)$ for each input image $x \in \mathcal{X} \subset \mathbb{R}^{n^0 d}$, with input channel count $n^0 \in \mathbb{N}$, number of pixels $d \in \mathbb{N}$, where

$$y_{i,\alpha}^l(x) \equiv \begin{cases} x_{i,\alpha} & l = 0 \\ \phi\left(z_{i,\alpha}^{l-1}(x)\right) & l > 0 \end{cases}, \qquad z_{i,\alpha}^l(x) \equiv \sum_{j=1}^{n^l} \sum_{\beta=-k}^{k} W_{ij,\beta}^l y_{j,\alpha+\beta}^l(x) + b_i^l. \tag{26}$$

[2] We will use Roman letters to index channels and Greek letters for spatial location. We use letters i, j, i', j', etc to denote channel indices, α, α', etc to denote spatial indices and β, β', etc for filter indices.

$\phi : \mathbb{R} \to \mathbb{R}$ is a point-wise activation function. Since we assume circular padding for all the convolutional layers, the spacial size d remains constant throughout the networks until the readout layer.

For each $l > 0$, as $\min\{n^1 \ldots, n^{l-1}\} \to \infty$, for each $i \in \mathbb{N}$, the pre-activation converges in distribution to d-dimensional Gaussian with mean $\mathbf{0}$ and covariance matrix $\mathcal{K}^{(l)}$, which can be computed recursively (Novak et al., 2019; Xiao et al., 2018)

$$\mathcal{K}^{(l+1)} = \mathcal{A} \circ \mathcal{T}(\mathcal{K}^{(l)}) = (\mathcal{A} \circ \mathcal{T})^{l+1}(\mathcal{K}^0) \tag{27}$$

Here $\mathcal{K}^{(l)} \equiv [\mathcal{K}^{(l)}_{\alpha,\alpha'}(x,x')]_{\alpha,\alpha' \in [d], x, x' \in \mathcal{X}}$, $\mathcal{T}$ is a non-linear transformation related to its fully-connected counterpart, and $\mathcal{A}$ a convolution coupled with a shift term acting on $\mathcal{X}d \times \mathcal{X}d$ PSD matrices

$$\left[\mathcal{T}\left(\mathcal{K}\right)\right]_{\alpha,\alpha'}(x,x') \equiv \mathbb{E}_{u \sim \mathcal{N}(0,K)}\left[\phi\left(u_\alpha(x)\right)\phi\left(u_{\alpha'}(x')\right)\right] \tag{28}$$

$$\left[\mathcal{A}\left(\mathcal{K}\right)\right]_{\alpha,\alpha'}(x,x') \equiv \sigma_b^2 + \sigma_\omega^2 \sum_\beta \frac{1}{2k+1}\left[K\right]_{\alpha+\beta,\alpha'+\beta}(x,x')\,. \tag{29}$$

B.1 The Neural Tangent Kernel

To understand how the neural tangent kernel evolves with depth, we define the NTK of the l-th hidden layer to be $\hat{\Theta}^{(l)}$

$$\hat{\Theta}^{(l)}_{\alpha,\alpha'}(x,x') = \nabla_{\theta^{\leq l}} z^l_{i,\alpha}(x) \nabla_{\theta^{\leq l}} z^l_{i,\alpha'}(x') \tag{30}$$

where $\theta^{\leq l}$ denotes all of the parameters in layers at-or-above the l'th layer. It does not matter which channel index i is used because as the number of channels approach infinity, this kernel also will converge in distribution to a deterministic kernel $\Theta^{(l+1)}$ Yang (2019), which can also be computed recursively in a similar manner to the NTK for fully-connected networks as,

$$\Theta^{(l+1)} = \mathcal{K}^{(l+1)} + \mathcal{A} \circ (\dot{\mathcal{T}}(\mathcal{K}^{(l)}) \odot \Theta^{(l)}) - \sigma_b^2, \tag{31}$$

where $\dot{\mathcal{T}}$ is given by Equation 28 with replacing ϕ by its derivative ϕ' in Equation 28. We will also normalize the variance of the inputs to q^* and hence treat $\mathcal{T}$ and $\dot{\mathcal{T}}$ as pointwise functions. We will only present the treatment in the chaotic phase to showcase how to deal with the operator $\mathcal{A}$. The treatment of other phases are similar. Note that the diagonals terms of $\mathcal{K}^{(l)}$ and $\Theta^{(l)}$ are exactly the same as the fully-connected setting. We only need to consider the off-diagonal terms. Letting $l \to \infty$ in Equation 31 we see that all the off-diagonal terms also converge . Let $\bar{\mathcal{A}} = \sigma_w^{-2}(\mathcal{A} - \sigma_b^2)$, be the normalized convolution operator. Note that $\mathcal{A}$ does not mix terms from different diagonals and it suffices to each off-diagonal separately. Let $\epsilon^{(l)}_{ab}$ and $\delta^{(l)}_{ab}$ denote the correction of the j-th diagonal of $\mathcal{K}^{(l)}$ and $\Theta^{(l)}$. Linearizing Equation 27 and Equation 31 gives

$$\epsilon^{(l+1)}_{ab} \approx \chi_c \bar{\mathcal{A}}(\epsilon^{(l+1)}_{ab} + \chi_{c,2} p^*_{ab} \epsilon^{(l)}_{ab} + \delta^{(l)}_{ab})\,. \tag{32}$$

Next let $\{\rho_\alpha\}_\alpha$ be the eigenvalues of $\bar{\mathcal{A}}$ and $\epsilon^{(l)}_{ab,\alpha}$ and $\delta^{(l+1)}_{ab,\alpha}$ be the projection of $\epsilon^{(l+1)}_{ab}$ and $\delta^{(l)}_{ab}$ onto the α-th eigenvector of $\bar{\mathcal{A}}$. Then for each α,

$$\epsilon^{(l+1)}_{ab,\alpha} \approx \rho_\alpha \chi_c (\epsilon^{(l+1)}_{ab,\alpha} + \chi_{c,2} p^*_{ab} \epsilon^{(l)}_{ab,\alpha} + \delta^{(l)}_{ab,\alpha}) \tag{33}$$

which gives

$$\epsilon^{(l)}_{ab,\alpha} \approx \rho_\alpha \chi_c^l \epsilon_{ab,\alpha}, \quad \delta^{(l)}_{ab,\alpha} \approx \rho_\alpha^l \chi_c^l \left[\delta^0_{ab,\alpha} + l\left(1 + \frac{\chi_{c,2}}{\chi_c} p^*_{ab}\right)\epsilon^0_{ab,\alpha}\right] \tag{34}$$

Therefore, the correction $\Theta^{(l)} - \Theta^*$ propagates independently through different Fourier modes. In each mode, up to a scaling factor ρ_α, the correction is the same as the correction of its FC counterpart. Since the subdominant modes (with $|\rho_a lpha| < 1$) decay exponentially faster than the dominant mode (with $\rho_\alpha = 1$), for large depth, the NTK correction of CNN is essentially the same as that of its FC counterpart.

NTK	FC/CNN-F, CNN-P		
	Ordered	Critical	Chaotic
$\lambda_{\max}^{(l)}$	$mq^* - m\mathcal{O}(l\chi_1^l)$	$\frac{md+2}{3d} lq^* + m\mathcal{O}(1)$	$\mathcal{O}(\chi_1^l)/d$
$\lambda_{\text{rest}}^{(l)}$	$\mathcal{O}(l\chi_1^l)/d$	$\frac{2}{3d} q^* l + \frac{1}{d}\mathcal{O}(1)$	$\mathcal{O}(\chi_1^l)/d$
$\kappa^{(l)}$	$dmq^*\mathcal{O}(\chi_1^{-l}/l)$	$\frac{md+2}{2} + dm\mathcal{O}(l^{-1})$	$\to 1$

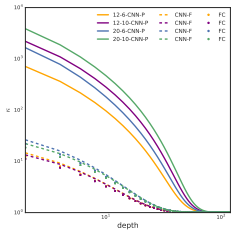
C FIGURE AND TABLE

C.1 SPECTRUM OF NTKS: FC VS CNN-F VS CNN-P

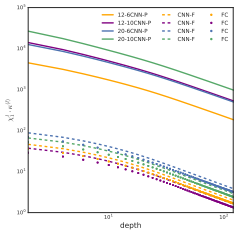
The table summarizes the grow/decay of $\lambda_{\max}^{(l)}$, $\lambda_{\text{rest}}^{(l)}$ and $\kappa^{(l)}$ in different phases.

C.2 CONVERGENCE/DIVERGENCE OF THE CONDITION NUMBERS

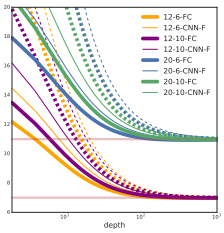
We plot the convergence/divergence of the condition numbers for fully-connected networks, convolution networks with and without pooling.



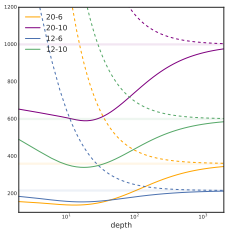
(a) Chaotic



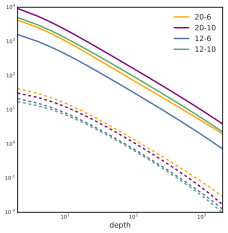
(b) Ordered



(c) Critical: FC and CNN-F



(d) Critical: CNN-P



(e) Critical: Convergent Rate

Figure 4: Condition numbers of NTKs in different phases. We use different colors represent different image size. For example, in the yellow "12-6" , "12" represent the batch size and "6" represent the dimension number ($6 * 6 * 3$ for FC and $(6, 6, 3)$ for CNN) (a) In the chaotic phase, κ converges to 1 for all architectures. (b) We plot $\chi_1^l \kappa^{(l)}$ confirming κ explodes with rate χ_1^l/l. In (c) and (d), the dashed lines representing the condition number and solid lines the ratio between first and second eigenvalues. We see that these two numbers converge to $\frac{m+2}{2}$ for FC and CNN-F and to $\frac{dm+2}{2}$ for CNN-P on the order-to-chaos transition. Finally, we plot the rates of convergence for CNN-P (solid) and CNN-F (dashed), confirming that pooling slows down the convergence by a factor of d.