解开|抽出|松手...

Disentangling Trainability and Generalization in Deep Neural Networks

Lechao Xiao [1] **Jeffrey Pennington** [1] **Samuel S. Schoenholz** [1]

Abstract

A longstanding goal in the theory of deep learning is to characterize the conditions under which a given neural network architecture will be trainable, and if so, how well it might generalize to unseen data. In this work, we provide such a characterization in the limit of very wide and very deep networks, for which the analysis simplifies considerably. For wide networks, the trajectory under gradient descent is governed by the Neural Tangent Kernel (NTK), and for deep networks the NTK itself maintains only weak data dependence. By analyzing the spectrum of the NTK, we formulate necessary conditions for trainability and generalization across a range of architectures, including Fully Connected Networks (FCNs) and Convolutional Neural Networks (CNNs). We identify large regions of hyperparameter space for which networks can memorize the training set but completely fail to generalize. We find that CNNs without global average pooling behave almost identically to FCNs, but that CNNs with pooling have markedly different and often better generalization performance. These theoretical results are corroborated experimentally on CIFAR10 for a variety of network architectures and we include a colab[1] notebook that reproduces the essential results of the paper.

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 often 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). Among numerous contributions, the pair of papers by Daniely *et al.* argued that the empirical covariance matrix of pre-activations becomes deterministic in the infinite-width limit and called this the conjugate kernel of the network. Meanwhile, from a mean-field perspective, the latter two papers studied the properties of these limiting kernels. In particular, the spectrum of the conjugate kernel of wide, fully-connected, networks approaches a well-defined and 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 extended to include 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,

[1]Google Research, Brain Team. Correspondence to: Lechao Xiao <xlc@google.com>, Samuel S. Schoenholz <schsam@google.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]Available at: https://tinyurl.com/ybsfxk5y.

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, a branch of concurrent work took a Bayesian perspective: that many networks converge to Gaussian Processes as their width becomes large (Lee et al., 2018; Matthews et al., 2018; Novak et al., 2019b; Garriga-Alonso et al., 2018; Yang, 2019). In this case, the Conjugate Kernel was referred to as the Neural Network Gaussian Process (NNGP) kernel, which is used to train neural networks in a fully Bayesian fashion. As such, the NNGP kernel characterizes performance of the corresponding NNGP.

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 $\tanh$ FCNs could be trained after the kernel reached its large-depth, data-independent, limit but that these networks did not generalize to unseen data.

Recently, 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:

- **Trainability.** We compute the large-depth asymptotics of several quantities related to trainability, including the largest/smallest eigenvalue of the NTK, $\lambda_{\text{max/min}}$, and the condition number $\kappa = \lambda_{\text{max}}/\lambda_{\text{min}}$; see Table 1.

- **Generalization.** We characterize the *mean predictor* $P(\Theta)$, which is intimately related to the prediction of wide neural networks on the test set following gradient descent training. As such, the mean predictor is intimately related to the model's ability to generalize. In particular, we argue that networks fail to generalize if the mean predictor becomes data-independent.

	NTK $\Theta^{(l)}$ of FC/CNN-F, CNN-P		
	Ordered $\chi_1 < 1$	Critical $\chi_1 = 1$	Chaotic $\chi_1 > 1$
$\lambda^{(l)}_{\text{max}}$	$mp^* + m\mathcal{O}(l\chi_1^l)$	$\frac{md+2}{3d} lq^* + m\mathcal{O}(1)$	$\Theta(\chi_1^l)/d$
$\lambda^{(l)}_{\text{bulk}}$	$\mathcal{O}(l\chi_1^l)/d$	$\frac{2}{3d} lq^* + \frac{1}{d}\mathcal{O}(1)$	$\Theta(\chi_1^l)/d$
$\kappa^{(l)}$	$dmp^*\Omega(\chi_1^{-l}/l)$	$\frac{md+2}{2} + dm\mathcal{O}(l^{-1})$	$1 + \mathcal{O}(d\chi_1^{-l})$
$P(\Theta^{(l)})Y_{\text{train}}$	$\mathcal{O}(1)$	$d\mathcal{O}(l^{-1})$	$d\mathcal{O}(l(\chi_{c^*}/\chi_1)^l)$

Table 1. **Evolution of the NTK spectra and $P(\Theta^{(l)})$ as a function of depth l.** The NTKs of FCN and CNN without pooling (CNN-F) are essentially the same and the scaling of $\lambda^{(l)}_{\text{max}}$, $\lambda^{(l)}_{\text{bulk}}$, $\kappa^{(l)}$, and $\Delta^{(l)}$ for these networks is written in black. Corrections to these quantities due to the addition of an average pooling layer (CNN-P) with window size d is written in blue.

- We show that the *ordered* and *chaotic* phases identified in Poole et al. (2016) lead to markedly different limiting spectra of the NTK. In the ordered phase the trainability of neural networks degrades at large depths, but their ability to generalize persists. By contrast, in the chaotic phase we show that trainability improves with depth, but generalization degrades and neural networks behave like hash functions.

  A corollary of these differences in the spectra is that, as a function of depth, the optimal learning rates ought to decay exponentially in the chaotic phase, linearly on the order-to-chase trainsition line, and remain roughly a constant in the ordered phase.

- We examine the differences in the above quantities for fully-connected networks (FCNs) and convolutional networks (CNNs) with and without pooling and precisely characterize the effect of pooling on the interplay between trainability, generalization, and depth.

In each case, we provide empirical evidence to support our theoretical conclusions. Together these results provide a complete, analytically tractable, and dataset-independent theory for learning in very deep and wide networks. Philosophically, we find that trainability and generalization are distinct notions that are, at least in this case, at odds with one another. Indeed, good conditioning of the NTK (which is a necessary condition for training) seems necessarily to lead to poor generalization performance. It will be interesting to see whether these results carry over in shallower and narrower networks. The tractable nature of the wide and deep regime leads us to conclude that these models will be an interesting testbed to investigate various theories of generalization in deep learning.

2. Related Work

Recent work Jacot et al. (2018); Du et al. (2018b); Allen-Zhu et al. (2018); Du et al. (2018a); Zou et al. (2018) and many others proved global convergence of over-parameterized deep networks by showing that the NTK essentailly remains a constant over the course of training. However, in a different scaling limit the NTK changes over the course of training and global convergence is much more difficult to obtain and is known for neural networks with one hidden layer Mei et al. (2018); Chizat & Bach (2018); Sirignano & Spiliopoulos (2018); Rotskoff & Vanden-Eijnden (2018). Therefore, understanding the training and generalization properties in this scaling limit remains a very challenging open question.

Another two excellent recent works (Hayou et al., 2019; Jacot et al., 2019) also study the dynamics of $\Theta^{(l)}(x, x')$ for FCNs (and deconvolutions in (Jacot et al., 2019)) as a function of depth and variances of the weights and biases. (Hayou et al., 2019) investigates role of activation functions (smooth v.s. non-smooth) and skip-connection. (Jacot et al., 2019) demonstrate that batch normalization helps remove the "ordered phase" (as in (Yang et al., 2019)) and a layer-dependent learning rate allows every layer in a network to contribute to learning.

3. Background

We summarize recent developments in the study of wide random networks. We will keep our discussion relatively informal; see e.g. (Novak et al., 2019b) for a more rigorous version of these arguments. To simplify this discussion and as a warm-up 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 the main text we will restrict our discussion to $\phi = \text{erf}$ or $\tanh$ for clarity, however we include results for a range of architectures including $\phi = \text{ReLU}$ with and without skip connections and layer normalization in the supplementary material (see Sec. B). We find that the high level picture described here applies to a wide range of architectural components, though important specifics - such as the phase diagram - can vary substantially. For simplicity, we will take the width of the hidden layers to infinity sequentially: $N^{(1)} \to \infty, \ldots, N^{(L-1)} \to \infty$. The network is parameterized by weights and biases that we take to be randomly initialized with $W_{ij}^{(l)}, b_i^{(l)} \sim \mathcal{N}(0, 1)$ along with hyperparameters, σ_w and σ_b that set the scale of the weights and biases respectively. Letting the i^{th} pre-activation in the l^{th} layer due to an input x be given by $z_i^{(l)}(x)$, the network is then described by the recursion, for $0 \leq l \leq L-1$,

$$z_i^{(l+1)}(x) = \frac{\sigma_w}{\sqrt{N^{(l)}}} \sum_{j=1}^{N^{(l)}} W_{ij}^{(l+1)} \phi(z_j^{(l)}(x)) + \sigma_b b_i^{(l+1)} \quad (1)$$

Notice that as $N^{(l)} \to \infty$, the sum ends up being over a large number of random variables and we can invoke the central limit theorem to conclude that the $\{z_i^{(l+1)}\}_{i \in [N^{(l+1)}]}$ are i.i.d. Gaussian with zero mean. Given a dataset of m points, the distribution over pre-activations can therefore be described completely by the covariance matrix, i.e. the NNGP kernel, between neurons in different inputs $\mathcal{K}^{(l)}(x, x') = \mathbb{E}[z_i^{(l)}(x) z_i^{(l)}(x')]$. Inspecting Equation 1, we see that $\mathcal{K}^{(l+1)}$ can be computed in terms of $\mathcal{K}^{(l)}$ as

$$\mathcal{K}^{(l+1)}(x, x') \equiv \sigma_w^2 \mathcal{T}(\mathcal{K}^{(l)})(x, x') + \sigma_b^2 \quad (2)$$

$$\mathcal{T}(\mathcal{K}) \equiv \mathbb{E}_{z \sim \mathcal{N}(0, \mathcal{K})}[\phi(z)\phi(z)^T] \quad (3)$$

Equation 2 describes a dynamical system on positive semi-definite matrices $\mathcal{K}$. It was shown in Poole et al. (2016) that fixed points, $\mathcal{K}^*(x, x')$, of these dynamics exist such that $\lim_{l \to \infty} \mathcal{K}^{(l)}(x, x') = \mathcal{K}^*(x, x')$ 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; see Fig 5 for $\tanh$-networks.

The rate at which $\mathcal{K}(x, x')$ approaches or departs $\mathcal{K}^*(x, x')$ can be determined by expanding Equation 2 about its fixed point, $\delta\mathcal{K}(x, x') = \mathcal{K}(x, x') - \mathcal{K}^*(x, x')$ to find

$$\delta\mathcal{K}^{(l+1)}(x, x') \approx \sigma_w^2 \dot{\mathcal{T}}(\mathcal{K}^*(x, x')) \delta\mathcal{K}^{(l)}(x, x') \quad (4)$$

with $\dot{\mathcal{T}}(\mathcal{K}) = \mathbb{E}_{(z_1, z_2) \sim \mathcal{N}(0, \mathcal{K})}[\dot{\phi}(z_1)\dot{\phi}(z_2)]$ and $\dot{\phi}$ is the derivative of ϕ. This expansion naturally exhibits exponential convergence to - or divergence from - the fixed-point as $\delta\mathcal{K}^{(l)}(x, x') \sim \chi(x, x')^l$ where $\chi(x, x') = \sigma_w^2 \dot{\mathcal{T}}(\mathcal{K}^*(x, x'))$. Since $\mathcal{K}^*(x, x')$ does not depend on x or x' it follows that $\chi(x, x')$ will take on a single value, χ_{c^*}, whenever $x \neq x'$. If $\chi_{c^*} < 1$ then this $\mathcal{K}^*$ fixed point is stable, but if $\chi_{c^*} > 1$ then the fixed point is unstable and, as discussed above, the system will converge to a different fixed point. If $\chi_{c^*} = 1$ then the hyperparameters lie at a phase transition and convergence is non-exponential. As was shown in Poole et al. (2016), there is always a fixed-point at $c^* = 1$ whose stability is determined by χ_1. This is the so-called ordered phase since any pair of inputs will converge to identical outputs. The line defined by $\chi_1 = 1$ defines the order-to-chaos transition separating the ordered

phase from the "chaotic" phase (where $c^* > 1$). Note, that χ_{c^*} can be used to define a depth-scale, $\xi_{c^*} = -1/\log(\chi_{c^*})$ that describes the number of layers over which $\mathcal{K}^{(l)}$ approaches $\mathcal{K}^*$.

This provides a precise characterization 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 finite width NTK is defined by $\hat{\Theta}(x, x') = J(x)J(x')^T$ where $J_{i\alpha}(x) = \partial_{\theta_\alpha} z_i^L(x)$ 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 converges to a deterministic kernel Θ and remains constant over the course of training. As such, at a time t during gradient descent training with an MSE loss, the expected outputs of an infinitely wide network, $\mu_t(x) = \mathbb{E}[z_i^L(x)]$, evolve as

$$\mu_t(X_{\text{train}}) = (\mathbf{Id} - e^{-\eta\Theta_{\text{train, train}}t})Y_{\text{train}} \tag{5}$$

$$\mu_t(X_{\text{test}}) = \Theta_{\text{test, train}}\Theta^{-1}_{\text{train, train}}(\mathbf{Id} - e^{-\eta\Theta_{\text{train, train}}t})Y_{\text{train}} \tag{6}$$

for train and test points respectively; see Section 2 in Lee et al. (2019). Here $\Theta_{\text{test, train}}$ denotes the NTK between the test inputs X_{test} and training inputs X_{train} and $\Theta_{\text{train, train}}$ is defined similarly. Since $\hat{\Theta}$ converges to Θ as the network's width approaches infinity, the gradient flow dynamics of real network also converge to the dynamics described by Equation 5 and Equation 6 (Jacot et al., 2018; Lee et al., 2019; Chizat et al., 2019; Yang, 2019; Arora et al., 2019; Huang & Yau, 2019). As the training time, t, tends to infinity we note that these equations reduce to $\mu(X_{\text{train}}) = Y_{\text{train}}$ and $\mu(X_{\text{test}}) = \Theta_{\text{test, train}}\Theta^{-1}_{\text{train, train}}Y_{\text{train}}$. Consequently we call

$$P(\Theta) \equiv \Theta_{\text{test, train}}\Theta^{-1}_{\text{train, train}} \tag{7}$$

the "mean predictor". We can also compute the mean predictor of the NNGP kernel, $P(\mathcal{K})$, which analogously can be used to find the mean of the posterior after Bayesian inference. We will discuss the connection between the mean predictor and generalization in the next section.

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,

$$\Theta^{(l+1)}(x, x') = \mathcal{K}^{(l+1)}(x, x') + \sigma_w^2 \dot{\mathcal{T}}(\mathcal{K}^{(l)})(x, x')\Theta^{(l)}(x, x')\,. \tag{8}$$

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

4. Metrics for Trainability and Generalization at Large Depth

We begin by discussing the interplay between the conditioning of $\Theta_{\text{train, train}}$ and the trainability of wide networks. We can write Equation 5 in terms of the spectrum of $\Theta_{\text{train, train}}$. To do this we write the eigendecomposition of $\Theta_{\text{train, train}}$ as $\Theta_{\text{train, train}} = U^T DU$ with D a diagonal matrix of eigenvalues and U a unitary matrix. In this case Equation 5 can be written as,

$$\tilde{\mu}_t(X_{\text{train}})_i = (\text{Id} - e^{-\eta\lambda_i t})\tilde{Y}_{\text{train},i} \tag{9}$$

where λ_i are the eigenvalues of $\Theta_{\text{train, train}}$ and $\tilde{\mu}_t(X_{\text{train}}) = U\mu_t(X_{\text{train}})$, $\tilde{Y}_{\text{train}} = UY_{\text{train}}$ are the mean prediction and the labels respectively written in the eigenbasis of $\Theta_{\text{train,train}}$. If we order the eigenvalues such that $\lambda_0 \geq \cdots \geq \lambda_m$ then it has been hypothesized[2] in e.g. Lee et al. (2019) that the maximum feasible learning rate scales as $\eta \sim 2/\lambda_0$ as we verify empirically in section 4. Plugging this scaling for η into Equation 9 we see that the smallest eigenvalue will converge exponentially at a rate given by $1/\kappa$, where $\kappa = \lambda_0/\lambda_m$ is the condition number. It follows that if the condition number of the NTK associated with a neural network diverges then it will become untrainable and so we use κ as a metric for trainability.

We will see that at large depths, the spectrum of $\Theta_{\text{train, train}}$ typically features a single large eigenvalue, λ_{max}, and then a gap that is large compared with the rest of the spectrum. We therefore will often refer to a typical eigenvalue in the bulk as λ_{bulk} and approximate the condition number as $\kappa = \lambda_{\text{max}}/\lambda_{\text{bulk}}$.

We now turn our attention to generalization. At large depths, we will see that $\Theta^{(l)}_{\text{test, train}}$ and $\Theta^{(l)}_{\text{train, train}}$ converge their fixed points independent of the data distribution. Consequently it is often the case that $P(\Theta^*)$ will be data-independent and the network will fail to generalize. In this case, by symmetry, it is necessarily true that $P(\Theta^*)$ will be a constant matrix. Contracting this matrix with a vector of labels Y_{train} that have been standardized to have zero mean it will follow that $P(\Theta^*)Y_{\text{train}} = 0$ and the network will output zero in expectation on all test points. Clearly, in this setting the network will not be able to generalize. At large, but finite, depths the generalization performance of the network can be quantified by considering the rate at which $P(\Theta^{(l)})Y_{\text{train}}$ decays to zero. There are cases, however, where despite the data-independence of Θ^*, $\lim_{l\to\infty} P(\Theta^{(l)})Y_{\text{train}}$ remains nonzero and the network can continue to generalize even in the asymptotic limit. In either case, we will show that precisely characterizing $P(\Theta^{(l)})Y_{\text{train}}$ allows us to understand exactly where networks can, and cannot, generalize.

[2]For finite width, the optimization problem is non-convex and there are not rigorous bounds on the maximum learning rate.

Our goal is therefore to characterize the evolution of the two metrics $\kappa^{(l)}$ and $P(\Theta^{(l)})$ in l. 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, σ_b, l). Our main results are summarized in Table 1 which describes the evolution of $\lambda^{(l)}_{\max}$ (the largest eigenvalue of $\Theta^{(l)}$), $\lambda^{(l)}_{\text{bulk}}$ (the remaining eigenvalues), $\kappa^{(l)}$, and $P(\Theta^{(l)})$ as a function of depth for three different network configurations (the ordered phase, the chaotic phase, and the phase transition). We study the dependence on: the size of the training set, m; the choices of architecture including fully-connected networks (FCN), convolutional networks with flattening (CNN-F), and convolutions with pooling (CNN-P); and the size, d, of the window in the pooling layer (which we always take to be the penultimate layer).

Before discussing the methodology it is useful to first give a qualitative overview of the phenomenology. We find identical phenomenology between FCNs and CNN-F architectures. In the ordered phase, $\Theta^{(l)} \to p^* \mathbf{1}\mathbf{1}^T$, $\lambda^{(l)}_{\max} \to mp^*$ and $\lambda^{(l)}_{\text{bulk}} = \mathcal{O}(l\chi_1^l)$. At large depths since $\chi_1 < 1$ it follows that $\kappa^{(l)} \gtrsim mp^*/(l\chi_1^l)$ and so the condition number diverges exponentially quickly. Thus, in the ordered phase we expect networks not to be trainable (or, specifically, the time they take to learn will grow exponentially in their depth). Here $P(\Theta^{(l)})$ converges to a data dependent constant independent of depth; thus, in the ordered phase networks fail to train but can generalize indefinitely.

By contrast, in the chaotic phase we see that there is no gap between $\lambda^{(l)}_{\max}$ and $\lambda^{(l)}_{\text{bulk}}$ and networks become perfectly conditioned and are trainable everywhere. However, in this regime we see that the mean predictor scales as $l(\chi_{c^*}/\chi_1)^l$. Since in the chaotic phase $\chi_{c^*} < 1$ and $\chi_1 > 1$ it follows that $P(\Theta^{(l)}) \to 0$ over a depth $\xi_* = -1/\log(\chi_{c^*}/\chi_1)$. Thus, in the chaotic phase, networks fail to generalize at a finite depth but remain trainable indefinitely. Finally, introducing pooling modestly augments the depth over which networks can generalize in the chaotic phase but reduces the depth in the ordered phase. We will explore all of these predictions in detail in section 7.

5. A Toy Example: RBF Kernel

To provide more intuition about our analysis, we present a toy example using RBF kernels which already shares some core observations for deep neural networks. Consider a Gaussian process along with the RBF kernel given by,

$$K_h(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{h}\right) \tag{10}$$

where $x, x' \in X_{\text{train}}$ along with a bandwidth $h > 0$. Note that $K_h(x, x) = 1$ for all h and x. Considering the following two cases.

If the bandwidth is given by $h = 2^l$ and $l \to \infty$, then $K_h(x, x') \approx 1 - 2^{-l}\|x - x'\|_2^2$ which converges to 1 exponentially fast. Thus, the largest eigenvalue of K_h is $\lambda_{\text{max}} \approx |X_{\text{train}}|$ and the bulk is of order $\lambda_{\text{bulk}} \approx 2^{-l}$. Thus the condition number $\kappa \gtrsim 2^l$ which diverges with l. We will see in the **Ordered Phase** $\Theta^{(l)}$ behaves qualitatively similar to this setting.

On the other hand, if the bandwidth is given by $h = 1/l$ and $l \to \infty$ then the off-diagonals $K_h(x, x') = \exp(-l\|x - x'\|_2^2) \to 0$. For large l, K_h is very close to the identity matrix and the condition number of it is almost 1. In the **Chaotic Phase**, $\Theta^{(l)}$ is qualitatively similar to K_h.

6. Large-Depth Asymptotics of the NNGP and NTK

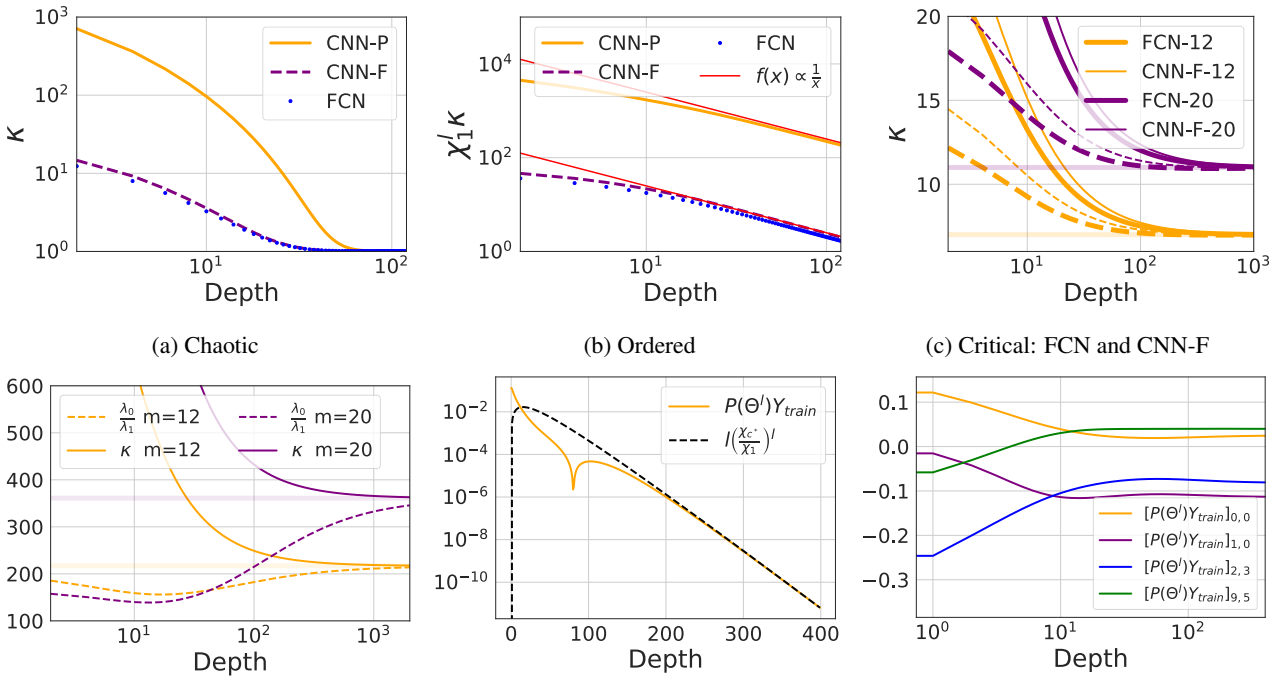
We now give a brief derivation of the results in Table 1. Details can be found in Sec.A, C in the appendix. To simplify notation we will discuss fully-connected networks and then extend the results to CNNs with pooling (CNN-P) and without pooling (CNN-F).

As in Sec. 3, we will be concerned with the fixed points of Θ as well as the linearization of Equation 8 about its fixed point. Recall that the fixed point structure is invariant 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 $\Theta^* = p^*((1 - \hat{c}^*)\mathbf{Id} + \hat{c}^* \mathbf{1}\mathbf{1}^T)$.

To simplify the forthcoming analysis, without a loss of generality, we assume the inputs are normalized to have variance q^* [3]. As such, we can treat $\mathcal{T}$ and $\dot{\mathcal{T}}$, restricted on $\{\mathcal{K}^{(l)}\}_l$, as a point-wise functions. To see this note that with this normalization $\mathcal{K}^{(l)}(x, x) = q^*$ for all l and x. It follows that both $\mathcal{T}(\mathcal{K}^{(l+1)})(x, x')$ and $\dot{\mathcal{T}}(\mathcal{K}^{(l+1)})(x, x')$ depend only on $\mathcal{K}^{(l)}(x, x')$.

Since all of the off-diagonal elements approach the same fixed point at the same rate, we use $q^{(l)}_{ab} \equiv \mathcal{K}^{(l)}(x, x')$ and $p^{(l)}_{ab} \equiv \Theta^{(l)}(x, x')$ 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, $\lim_{l\to\infty} q^{(l)}_{ab} = q^*_{ab} = c^* q^*$ and $\lim_{l\to\infty} p^{(l)}_{ab} = p^*_{ab} = \hat{c}^* p^*$. Finally, although the diagonal entries of $\mathcal{K}^{(l)}$ are all q^*, the diagonal entries of $\Theta^{(l)}$ can

[3] It has been observed in previous works (Poole et al., 2016; Schoenholz et al., 2017) that the diagonals converge much faster than the off-diagonals for tanh- or erf- networks.



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

(d) Critical: CNN-P (e) Chaotic: $P(\Theta^{(l)})Y_{\text{train}}$ (f) Ordered: $P(\Theta^{(l)})Y_{\text{train}}$

Figure 1. **Condition number and mean predictor of NTKs and their rate of convergence for FCN, CNN-F and CNN-P.** (a) In the chaotic phase, $\kappa^{(l)}$ converges to 1 for all architectures. (b) We plot $\chi_1^l \kappa^{(l)}$, confirming that κ explodes with rate $1/l\chi_1^l$ in the ordered phase. In (c) and (d), the solid lines are $\kappa^{(l)}$ and dashed lines are the ratio between first and second eigenvalues. We see that, on the order-to-chaos transition, these two numbers converge to $\frac{m+2}{2}$ and $\frac{dm+2}{2}$ (horizontal lines) for FC/CNN-F and CNN-P respectively, where $m = 12$ or 20 is the batch size and $d = 36$ is the spatial dimension. (e) In the chaotic phase, the mean predictor decays to zero exponentially fast. (f) In the ordered phase the mean predictor converges to a data dependent value.

vary and we denote them $p^{(l)}$.

In what follows, we split the discussion into three sections according to the values of $\chi_1 \equiv \sigma_\omega^2 \dot{\mathcal{T}}(q^*)$ recalling that in Poole et al. (2016); Schoenholz et al. (2017) it was shown that χ_1 controls the fixed point structure. In each section, we analyze the evolution of (1) the entries of $\Theta^{(l)}$, i.e., $p^{(l)}$, $p_{ab}^{(l)}$, (2) the spectrum $\lambda_{\max}^{(l)}$ and $\lambda_{\text{bulk}}^{(l)}$, (3) the trainability and generalization metrics $\kappa^{(l)}$ and $P(\Theta^{(l)})$, and finally (4) discuss the impact on finite width networks.

6.1. The Chaotic Phase $\chi_1 > 1$:

The chaotic phase is so-named because it has a stable fixed-point $c^* < 1$; as such similar inputs become increasingly uncorrelated as they pass through the network. Our first result is to show that (see Sec. A.1),

$$\begin{cases} q_{ab}^{(l)} = q_{ab}^* + \mathcal{O}(\chi_{c^*}^l) \\ q^{(l)} = q^* \end{cases} \quad \begin{cases} p_{ab}^{(l)} = p_{ab}^* + \mathcal{O}(l\chi_{c^*}^l) \\ p^{(l)} = q^* \frac{\chi_1^l - 1}{\chi_1 - 1} \end{cases} \tag{11}$$

where

$$p_{ab}^* = q_{ab}^*/(1 - \chi_{c^*}) \text{ and } \chi_{c^*} = \sigma_\omega^2 \dot{\mathcal{T}}(q_{ab}^*) \tag{12}$$

Note that χ_{c^*} controls the convergence of the $q_{ab}^{(l)}$ and is always less than 1 in the chaotic phase (Poole et al., 2016; Schoenholz et al., 2017; Xiao et al., 2018). Since $\chi_1 > 1$, $p^{(l)}$ diverges with rate χ_1^l while $p_{ab}^{(l)}$ remains finite. It follows that $(p^{(l)})^{-1}\Theta^{(l)} \to \mathbf{Id}$ as $l \to \infty$. Thus, in the chaotic phase, the spectrum of the NTK for very deep networks approaches the diverging constant multiplying the identity. This implies

$$\lambda_{\max}^{(l)}, \lambda_{\text{bulk}}^{(l)} = p^{(l)} + O(1) \text{ and } \kappa^{(l)} = 1 + \mathcal{O}\left(\frac{1}{p^{(l)}}\right)$$

Figure 1a plots the evolution of $\kappa^{(l)}$ in this phase, confirming $\kappa^{(l)} \to 1$ for all three different architectures (FCN, CNN-F and CNN-P).

We now describe the asymptotic behavior of the mean predictor. Since $\Theta^l_{\text{test, train}}$ has no diagonal elements, it follows that it remains finite at large depths and so $P(\Theta^*)Y_{\text{train}} = 0$. It follows that in the chaotic phase, the predictions of asymptotically deep neural networks on unseen test points will converge to zero exponentially quickly (see Sec. C.1),

$$P(\Theta^{(l)})Y_{\text{train}} \approx \mathcal{O}(l(\chi_{c^*}/\chi_1)^l) \to \mathbf{0}. \tag{13}$$

Neglecting the relatively slowly varying polynomial term, this implies that we expect chaotic networks to fail to generalize when their depth is much larger than a scale set by $\xi_* = -1/\log(\chi_{c^*}/\chi_1)$. We confirm this scaling in Fig 1e.

We confirm these predictions for finite-width neural network training using SGD as well as gradient-flow on infinite networks in the experimental results; see Fig 2.

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

The ordered phase is defined by the stability of the $c^* = 1$ fixed point. Here disparate inputs will end up converging to the same output at the end of the network. We show in Sec. A.2 that elements of the NNGP kernel and NTK have asymptotic dynamics given by,

$$\begin{cases} q_{ab}^{(l)} = q^* + \mathcal{O}(\chi_1^l) \\ q^{(l)} = q^* \end{cases} \quad \begin{cases} p_{ab}^{(l)} = p^* + \mathcal{O}(l\chi_1^l) \\ p^{(l)} = p^* + \mathcal{O}(\chi_1^l) \end{cases} \tag{14}$$

where $p^* = q^*/(1-\chi_1)$. Here all of the entries of $\Theta^{(l)}$ converge to the same value, p^*, and the limiting kernel has the form $\Theta^* = p^* \mathbf{1_n} \mathbf{1_m^T}$ where $\mathbf{1_m}$ is the all-ones vector of dimension m (typically m will correspond to the number of datapoints in the training set). The NNGP kernel has the same structure with $p^* \leftrightarrow q^*$. Consequently both the NNGP kernel and the NTK are highly singular and feature a single non-zero eigenvalue, $\lambda_{\text{max}} = mp^*$, with eigenvector $\mathbf{1_m}$.

For large-but-finite depths, $\Theta^{(l)}$ has (approximately) two eigenspaces: the first eigenspace corresponds to finite-depth corrections to λ_{max},

$$\lambda_{\text{max}}^{(l)} \approx (m-1)p_{ab}^{(l)} + p^{(l)} = mp^* + \mathcal{O}(l\chi_1^l). \tag{15}$$

The second eigenspace comes from lifting the degenerate zero-modes has dimension $(m-1)$ with eigenvalues that scale like $\lambda_{\text{bulk}}^{(l)} = \mathcal{O}(p^{(l)} - p_{ab}^{(l)}) = \mathcal{O}(l\chi_1^l)$. It follows that $\kappa^{(l)} \gtrsim (l\chi_1^l)^{-1}$ and so the conditioning number explodes exponentially quickly. We confirm the presence of the $1/l$ correction term in $\kappa^{(l)}$ by plotting $\chi_1^l \kappa^{(l)}$ against l in Figure 1b. Neglecting this correction, we expect networks in the ordered phase to become untrainable when their depth exceeds a scale given by $\xi_1 = -1/\log \chi_1$.

We now turn our discussion to the mean predictor. Equation 14 shows that we can write the finite-depth corrections to the NTK as $\Theta^{(l)} = p^* \mathbf{1}\mathbf{1}^T + \boldsymbol{A}^{(l)} l\chi_1^l$. Here $\boldsymbol{A}^{(l)}$ is the data-dependent piece that lifts the zero eigenvalues. In the appendix, $\boldsymbol{A}^{(l)}$ converges to $\boldsymbol{A}$ as $l \to \infty$; see Lemma 2. In Sec. C.3 we show that despite the singular nature of Θ^*, the mean has a well-defined limit as,

$$\lim_{l\to\infty} P(\Theta^{(l)}) Y_{\text{train}} = (\boldsymbol{A}_{\text{test, train}} \boldsymbol{A}_{\text{train, train}}^{-1} + \hat{\boldsymbol{A}}) Y_{\text{train}}, \tag{16}$$

where $\hat{\boldsymbol{A}}$ is some correction term. Thus, the mean predictor remains well-behaved and data dependent even in the infinite-depth limit. Thus, we suspect that networks in the ordered phase should be able to generalize whenever they can be trained. We confirm the asymptotic data-dependence of the mean predictor in Fig 1f.

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

On the critical line the $c^* = 1$ fixed point is marginally stable and dynamics become powerlaw. Here, both the diagonal and the off-diagonal elements of $\Theta^{(l)}$ diverge linearly in the depth with $\frac{1}{l}\Theta^{(l)} \to \frac{q^*}{3}(\mathbf{1}\mathbf{1}^T + 2\mathbf{Id})$. The condition number $\kappa^{(l)}$ converges to a finite value and the network is always trainable. However, the mean predictor decreases linearly with depth. In particular we show in Sec. A.3,

$$\begin{cases} q_{ab}^{(l)} = q^* + \mathcal{O}(\frac{1}{l}) \\ q^{(l)} = q^* \end{cases} \quad \begin{cases} p_{ab}^{(l)} = \frac{1}{3} l p^* + \mathcal{O}(1) \\ p^{(l)} = l p^* \end{cases} \tag{17}$$

For large l it follows that $\Theta^{(l)}$ essentially has two eigenspaces: one has dimension one and the other has dimension $(m-1)$ with

$$\lambda_{\text{max}}^{(l)} = \frac{(m+2)q^* l}{3} + \mathcal{O}(1), \; \lambda_{\text{bulk}}^{(l)} = \frac{2q^* l}{3} + \mathcal{O}(1). \tag{18}$$

It follows that the condition number $\kappa^{(l)} = \frac{m+2}{2} + m\mathcal{O}(l^{-1}) \to \frac{m+2}{2}$ as $l \to \infty$. Unlike in the chaotic and ordered phases, here $\kappa^{(l)}$ converges with rate $\mathcal{O}(l^{-1})$. Figure 1c confirms the $\kappa^{(l)} \to \frac{m+2}{2}$ for both FCN and CNN-F (the global average pooling in CNN introduces a correction term that we will discuss below). A similar calculation gives $P(\Theta^{(l)}) = \mathcal{O}(l^{-1})$ on the critical line.

In summary, $\kappa^{(l)}$ converges to a finite number and the network ought to be trainable for arbitrary depth but the mean predictor $P(\Theta^{(l)})$ decays as a powerlaw. Decay as l^{-1} is much slower than exponential and is slow on the scale of neural networks. This explains why critically initialized networks with thousands of layers could still generalize (Xiao et al., 2018).

6.4. The Effect of Convolutions

The above theory can be extended to CNNs. We will provide an informal description here, with details in Sec. E. For an input-images of size $(m, k, k, 3)$ the NTK and NNGP kernels will have shape (m, k, k, m, k, k) and will contain information about the covariance between each pair of pixels in each image. For convenience we will let $d = k^2$. In the large depth setting deviations of both kernels from their fixed point decomposes via Fourier transform in the spatial dimensions as,

$$\delta\Theta_{\text{CNN}}^{(l)} \approx \sum_q \rho_q^l \delta\Theta^{(l)}(q) \tag{19}$$

where q denotes the Fourier mode with $q = 0$ being the zero-frequency (uniform) mode and ρ_q are eigenvalues of certain

convolution operator. Here $\delta\Theta^{(l)}(q)$ are deviations from the fixed-point for the q^{th} mode with $\delta\Theta^{(l)}(q) \propto \delta\Theta^{(l)}_{\text{FCN}}$ the fully-connected deviation described above. We show that $\rho_{q=0} = 1$ and $|\rho_{q\neq 0}| < 1$ which implies that asymptotically the nonuniform modes become subleading as $\rho^l_q \to 0$. Thus, at large depths different pixels evolve identically as FCNs.

In Sec. E.2 we discuss the differences that arise when one combines a CNN with a flattening layer compared with an average pooling layer at the readout. In the case of flattening, the pixel-pixel correlations are discarded and $\Theta^{(l)}_{\text{CNN-F}} \approx \Theta^{(l)}_{\text{FCN}}$. The plots in the first row of Figure 1 confirm that the $\kappa^{(l)}$ of $\Theta^{(l)}_{\text{CNN-F}}$ and of $\Theta^{(l)}_{\text{FCN}}$ evolve almost identically in all phases. Note that this clarifies an empirical observation in Xiao et al. (2018) (Figure 3 of Xiao et al. (2018)) that test performance of critically initialized CNNs degrades towards that of FCNs as depth increases. This is because (i) in the large width limit, the prediction of neural networks is characterized by the NTK and (ii) the NTKs of the two models are almost identical for large depth. However, when CNNs are combined with global average pooling a correction to the spectrum of the NTK (NNGP) emerges oweing to pixel-pixel correlations; this alters the dynamics of $\kappa^{(l)}$ and $P(\Theta^{(l)})$. In particular, we find that global average pooling increases $\kappa^{(l)}$ by a factor of d in the ordered phase and on the critical line; see Table 1 for the exact correction as well as Figures 1d for experimental evidence of this correction.

6.5. Dropout, Relu and Skip-connection

Adding a dropout to the penultimate layer has a similar effect to adding a diagonal regularization term to the NTK, which significantly improves the conditioning of the NTK in the ordered phase. In particular, adding a single dropout layer can cause $\kappa^{(l)}$ to converge to a finite κ^* rather than diverges exponentially; see Figure 4 and Sec. D.

For critically initialized Relu networks (aka, He's initialization (He et al., 2015)), the entries of the NTK also diverges linearly and $\kappa^{(l)} \to \frac{m+3}{3}$ and $P(\Theta^{(l)}) = \mathcal{O}(1/l)$; see Table 2 and Figure 3 in SM. In addition, adding skip-connections makes all entries of the NTK to diverge exponentially, resulting exploding of gradients. However, we find that skip connections do not alter the dynamics of $\kappa^{(l)}$. Finally, layer normalization could help address the issue of exploding of gradients; see Sec. B.

7. Experiments

Evolution of $\kappa^{(l)}$ (Figure 1). We randomly sample inputs with shape $(m, k, k, 3)$ where $m \in \{12, 20\}$ and $k = 6$. We compute the exact NTK with activation function *Erf* using the *Neural Tangents* library (Novak et al., 2019a). We see excellent agreement between the theoretical calculation of $\kappa^{(l)}$ in Sec. 6 (summarized in Table 1) and the experimental results Figure 1.

Maximum Learning Rates (Figure 2 (c)). In practice, given a set of hyper-parameters of a network, knowing the range of feasible learning rates is extremely valuable. As discussed above, in the infinite width setting, Equation 5 implies the maximal convergent learning rate is given by $\eta_{\text{theory}} \equiv 2/\lambda^{(l)}_{\text{max}}$. From our theoretical results above, varying the hyperparameters of our network allows us to vary $\lambda^{(l)}_{\text{max}}$ over a wide range and test this hypothesis. This is shown for depth 10 networks varying σ^2_w with $\eta = \rho\eta_{\text{theory}}$. We see that networks become untrainable when ρ exceeds 2 as predicted.

Trainability vs Generalization (Figure 2 (a,b)). We conduct an experiment training finite-width CNN-F networks with 1k training samples from CIFAR-10 with 20×20 different (σ^2_ω, l) configurations. We train each network using SGD with batch size $b = 256$ and learning rate $\eta = 0.1\eta_{\text{theory}}$. We see in Figure 2 (a) that deep in the chaotic phase we see that all configurations reach perfect training accuracy, but the network completely fails to generalize in the sense test accuracy is around 10%. As expected, in the ordered phase we see that although the training accuracy degrades generalization improves. As expected we see that the depth-scales ξ_1 and ξ_* control trainability in the ordered phase and generalization in the chaotic phase respectively. We also conduct extra experiments for FCN with more training points (16k); see Figure 6.

CNN-P v.s. CNN-F: spatial correction (Figure 2 (d-f)). We compute the test accuracy using the analytic equations for gradient flow, Equation 6, which corresponds to the test accuracy of ensemble of gradient descent trained neural networks taking the width to infinity. As above, we use $1k$ training points and consider a 20×20 grid of configurations for (σ^2_ω, l). We plot the test performance of CNN-P and CNN-F and the performance difference in Fig 2 (d-f). As expected, we see that 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 see that the test performance difference between CNN-P and CNN-F exhibits a region in the ordered phase (a blue strip) where CNN-F outperforms CNN-P by a large margin. This performance difference is due to the correction term d as predicted by the $P(\Theta^{(l)})$-row of Table 1. We also conduct extra experiments densely varying σ^2_b; see Sec. F.4. Together these results provide an extremely stringent test of our theory.

8. Conclusion and Future Work

In this work, we identify several quantities (λ_{max}, λ_{bulk}, κ, and $P(\Theta^{(l)})$) related to the spectrum of the NTK that

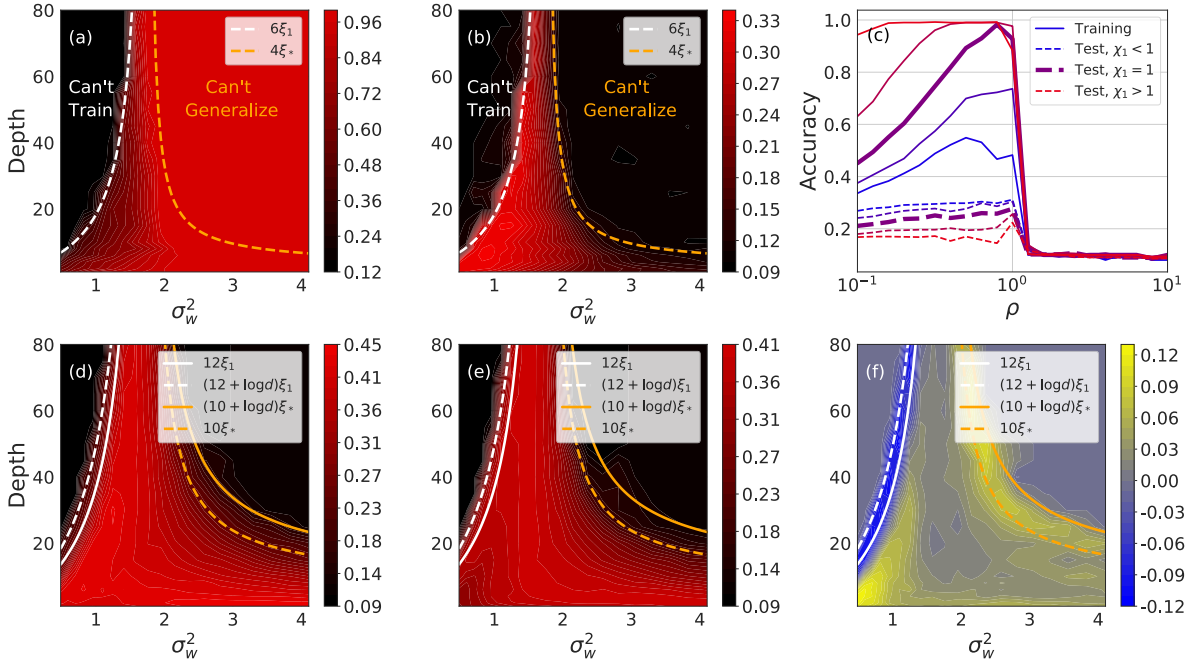


Figure 2. **Trainability and generalization are captured by $\kappa^{(l)}$ and $P(\Theta^{(l)})$.** (a,b) The training and test accuracy of CNN-F trained with SGD. The network is untrainable above the green line because $\kappa^{(l)}$ is too large and is ungeneralizable above the orange line because $P(\Theta^{(l)})$ is too small. (c) The accuracy vs learning rate for FCNs trained with SGD sweeping over the weight variance. (d,e) The test accuracy of CNN-P and CNN-F using kernel regression. (f) The difference in accuracy between CNN-P and CNN-F networks.

control trainability and generalization of deep networks. We offer a precise characterization of these quantities and provide substantial experimental evidence supporting their role in predicting the training and generalization performance of deep neural networks. Future work might extend our framework to other architectures (for example, residual networks with batch-norm or attention architectures). Understanding the role of the nonuniform Fourier modes in the NTK in determining the test performance of CNNs is also an important research direction.

In practice, the correspondence between the NTK and neural networks is often broken due to, e.g., insufficient width, using a large learning rate, or changing the parameterization. Our theory does not directly apply to this setting. As such, developing an understanding of training and generalization away from the NTK regime remains an important research direction.

Acknowledgements

We thank Jascha Sohl-dickstein, Greg Yang, Ben Adlam, Jaehoon Lee, Roman Novak and Yasaman Bahri for useful discussions and feedback. We also thank anonymous reviewers for feedback that helped improve the manuscript.

References

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

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

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

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

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal

transport. In *Advances in neural information processing systems*, pp. 3040–3050, 2018.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. 2019.

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

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

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

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018a.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks, 2018b.

Garriga-Alonso, A., Rasmussen, C. E., and Aitchison, L. Deep convolutional networks as shallow gaussian processes, 2018.

Gilboa, D., Chang, B., Chen, M., Yang, G., Schoenholz, S. S., Chi, E. H., and Pennington, J. Dynamical isometry and a mean field theory of lstms and grus. *CoRR*, abs/1901.08987, 2019. URL http://arxiv.org/abs/1901.08987.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. 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.

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

Hayou, S., Doucet, A., and Rousseau, J. On the selection of initialization and activation function for deep neural networks. *arXiv preprint arXiv:1805.08266*, 2018.

Hayou, S., Doucet, A., and Rousseau, J. Mean-field behaviour of neural tangent kernel for deep neural networks, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015. URL http://arxiv.org/abs/1502.01852.

Huang, J. and Yau, H.-T. Dynamics of deep neural networks and neural tangent hierarchy. *arXiv preprint arXiv:1909.08156*, 2019.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*. 2018.

Jacot, A., Gabriel, F., and Hongler, C. Freeze and chaos for dnns: an ntk view of batch normalization, checkerboard and boundary effects, 2019.

Karakida, R., Akaho, S., and Amari, S.-i. Universal statistics of fisher information in deep neural networks: mean field approach. *arXiv preprint arXiv:1806.01316*, 2018.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S., Pennington, J., and Sohl-dickstein, J. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

Matthews, A., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 4 2018. URL https://openreview.net/forum?id=H1-nGgWC-.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

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

Novak, R., Lee, L. X. J., Sohl-Dickstein, J., and Schoenholz, S. S. Neural tangents: Fast and easy infinite neural networks in python, 2019a. URL http://github.com/google/neural-tangents.

Novak, R., Xiao, L., Lee, J., Bahri, Y., Yang, G., Hron, J., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. Bayesian deep convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations*, 2019b.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Pennington, J., Schoenholz, S. S., and Ganguli, S. The emergence of spectral universality in deep networks. *arXiv preprint arXiv:1802.09979*, 2018.

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

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. In *Advances In Neural Information Processing Systems*, pp. 3360–3368, 2016.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. Transfusion: Understanding transfer learning with applications to medical imaging. *arXiv preprint arXiv:1902.07208*, 2019.

Rotskoff, G. M. and Vanden-Eijnden, E. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv preprint arXiv:1805.00915*, 2018.

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

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. 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.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks. *arXiv preprint arXiv:1805.01053*, 2018.

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. 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.

Yang, G. 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.

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

Yang, G., Pennington, J., Rao, V., Sohl-Dickstein, J., and Schoenholz, S. S. A mean field theory of batch normalization. *arXiv preprint arXiv:1902.08129*, 2019.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.