A. Signal propagation of NNGP and NTK

In this section, we assume that the activation function ϕ has a continuous third derivative. Recall that the recursive formulas for NNGP $\mathcal{K}^{(l)}$ and the NTK $\Theta^{(l)}$ are given by

$$\mathcal{K}^{(l+1)}(x,x') = \sigma_w^2 \mathcal{T}(\mathcal{K}^{(l)})(x,x') + \sigma_b^2 \,. \tag{20}$$

$$\Theta^{(l+1)}(x,x') = \mathcal{K}^{(l+1)}(x,x') + \sigma_w^2 \dot{\mathcal{T}}(\mathcal{K}^{(l)})(x,x')\Theta^{(l)}(x,x') \tag{21}$$

where

$$\begin{cases} \mathcal{T}(\mathcal{K})(x,x') = \mathbb{E}\phi(u)\phi(v), \\ \dot{\mathcal{T}}(\mathcal{K})(x,x') = \mathbb{E}\dot{\phi}(u)\dot{\phi}(v), \end{cases} \qquad (u,v)^T \sim \mathcal{N}\left(0, \begin{bmatrix} q^* & \mathcal{K}(x,x') \\ \mathcal{K}(x,x') & q^* \end{bmatrix}\right) \tag{22}$$

Note that we have normalized each input to have variance q^* and the diagonals of $\mathcal{K}^{(l)}$ are equal to q^* for all l. The off-diagonal terms of $\mathcal{K}^{(l)}$ and $\Theta^{(l)}$ are denoted by $q_{ab}^{(l)}$ and $p_{ab}^{(l)}$, resp. and the diagonal terms are $q^{(l)}$ and $p^{(l)}$, resp. The above equations can be simplified to

$$q_{ab}^{(l+1)} = \sigma_w^2 \mathcal{T}(q_{ab}^{(l)}) + \sigma_b^2 \qquad\qquad p_{ab}^{(l+1)} = q_{ab}^{(l+1)} + \sigma_w^2 \dot{\mathcal{T}}(q_{ab}^{(l)})\, p_{ab}^{(l)} \tag{23}$$

$$q^{(l+1)} = q^* \qquad\qquad p^{(l+1)} = q^* + \sigma_w^2 \dot{\mathcal{T}}(q^*)\, p^{(l)} \tag{24}$$

In what follows, we compute the evolution of $q_{ab}^{(l)}, p_{ab}^{(l)}, p^{(l)}$ and the spectrum and condition numbers of $\mathcal{K}^{(l)}$ and $\Theta^{(l)}$. We will use $\lambda_{\max}(\Theta^{(l)})/\lambda_{\max}(\mathcal{K}^{(l)})$, $\lambda_{\text{bulk}}(\Theta^{(l)})/\lambda_{\text{bulk}}(\mathcal{K}^{(l)})$ and $\kappa(\Theta^{(l)})/\kappa(\mathcal{K}^{(l)})$ to denote the maximum eigenvalues, the bulk eigenvalues and the condition number of $\Theta^{(l)}/\mathcal{K}^{(l)}$, resp.

A.1. Chaotic Phase

A.1.1. Correction of the off-diagonal/diagonal

The diagonal terms are relatively simple to compute. Equation 24 gives

$$p^{(l+1)} = q^* + \chi_1 p^{(l)} \tag{25}$$

i.e.

$$p^{(l)} = \frac{1-\chi_1^{(l)}}{1-\chi_1} q^* \tag{26}$$

In the chaotic phase, $\chi_1 > 1$ and $p^{(l)} \approx \chi_1^{l-1} q^*$, i.e. diverges exponentially quickly.

Now we compute the off-diagonal terms. Since $\chi_{c^*} = \sigma_\omega^2 \dot{\mathcal{T}}(q_{ab}^*) < 1$ in the chaotic, p_{ab}^* exists and is finite. Indeed, letting $l \to \infty$ in equation 23, we have

$$q_{ab}^* = \sigma_w^2 \mathcal{T}(q_{ab}^*) + \sigma_b^2 \tag{27}$$

$$p_{ab}^* = q_{ab}^* + \sigma_w^2 \dot{\mathcal{T}}(q_{ab}^*)\, p_{ab}^* \tag{28}$$

which gives

$$p_{ab}^* = \frac{q_{ab}^*}{1-\chi_{c^*}} \tag{29}$$

To compute the finite depth correction, let

$$\epsilon_{ab}^{(l)} = q_{ab}^{(l)} - q_{ab}^* \tag{30}$$

$$\delta_{ab}^{(l)} = p_{ab}^{(l)} - p_{ab}^* \tag{31}$$

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

	NNGP $\mathcal{K}^{(l)}$ of FC/CNN-F, CNN-P		
	Ordered $\chi_1 < 1$	Critical $\chi_1 = 1$	Chaotic $\chi_1 > 1$
$\lambda_{\mathrm{max}}^{(l)}$	$mq^* + m\mathcal{O}(\chi_1^l)$	$mq^* + \mathcal{O}(l^{-1})$	$((1-c^*)/d + mc^*)q^* + \mathcal{O}(\chi_{c^*}^l)$
$\lambda_{\mathrm{bulk}}^{(l)}$	$\mathcal{O}(\chi_1^l)/d$	$\mathcal{O}(l^{-1})/d$	$(1-c^*)q^*/d + \mathcal{O}(\chi_{c^*}^l)$
$\kappa^{(l)}$	$dmq^*\Omega(\chi_1^{-l})$	$dm\Omega(l)$	$1 + dm\frac{c^*}{1-c^*} + d\mathcal{O}(\chi_{c^*}^l)$
$P(\mathcal{K}^{(l)})Y_{\text{Train}}$	$\mathcal{O}(1)$	$\mathcal{O}(l^{-1})$	$d\mathcal{O}(\chi_{c^*}^l)$

Table 2. **Evolution of the NTK/NNGP spectrum and $P(\Theta^{(l)})Y_{\text{train}}/P(\mathcal{K}^{(l)})Y_{\text{train}}$ as a function of depth l.** The NTKs of FCN and CNN without pooling (CNN-F) are essentially the same and the scaling of $\lambda_{\mathrm{max}}^{(l)}$, $\lambda_{\mathrm{bulk}}^{(l)}$, $\kappa^{(l)}$, 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.

Applying Taylor's expansion to the first equation of 23 gives

$$\begin{aligned} q_{ab}^* + \epsilon_{ab}^{(l+1)} &= \sigma_\omega^2 \mathcal{T}(q_{ab}^* + \epsilon_{ab}^{(l)}) + \sigma_b^2 && (32)\\ &= \sigma_\omega^2 \mathcal{T}(q_{ab}^*) + \sigma_b^2 + \sigma_\omega^2 \dot{\mathcal{T}}(q_{ab}^*)\epsilon_{ab}^{(l)} + \mathcal{O}((\epsilon_{ab}^{(l)})^2) && (33)\\ &= q_{ab}^* + \sigma_\omega^2 \dot{\mathcal{T}}(q_{ab}^*)\epsilon_{ab}^{(l)} + \mathcal{O}((\epsilon_{ab}^{(l)})^2) && (34)\end{aligned}$$

That is

$$\epsilon_{ab}^{(l+1)} = \chi_{c^*}\epsilon_{ab}^{(l)} + \mathcal{O}((\epsilon_{ab}^{(l)})^2) \tag{35}$$

Thus $q_{ab}^{(l)}$ converges to q_{ab}^* exponentially quickly with

$$\epsilon_{ab}^{(l+1)} \approx \chi_{c^*}\epsilon_{ab}^{(l)} \approx \chi_{c^*}^{l+1}\epsilon_{ab}^{(0)} \tag{36}$$

Similarly, applying Taylor's expansion to the second equation of 23 gives

$$\delta_{ab}^{(l+1)} = (1 + \frac{\chi_{c^*,2}}{\chi_{c^*}}p_{ab}^*)\epsilon_{ab}^{(l+1)} + \chi_{c^*}\delta_{ab}^{(l)} + \mathcal{O}((\epsilon_{ab}^{(l)})^2) \tag{37}$$

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

$$\begin{aligned} \epsilon_{ab}^{(l)} &\approx \chi_{c^*}^l\,\epsilon_{ab}^{(0)} && (38)\\ \delta_{ab}^{(l)} &\approx \chi_{c^*}^l\left[\delta_{ab}^{(0)} + l\left(1 + \frac{\chi_{c^*,2}}{\chi_{c^*}}p_{ab}^*\right)\epsilon_{ab}^{(0)}\right]. && (39)\end{aligned}$$

Note that $\delta_{ab}^{(l)}$ contains a polynomial correction term and decays like $l\chi_{c^*}^l$.

Lemma 1. *There exist a finite number ζ_{ab} such that*

$$|\chi_{c^*}^{-l}\epsilon_{ab}^{(l)} - \zeta_{ab}| \lesssim \chi_{c^*}^l \quad \textit{and} \quad |\chi_{c^*}^{-l}\delta_{ab}^{(l+1)} - l(1 + \frac{\chi_{c^*,2}}{\chi_{c^*}}p_{ab}^*)\zeta_{ab}| \lesssim \chi_{c^*}^l. \tag{40}$$

We want to emphasize that the limits are data-dependent, which was verified in Fig. 1e and 1f empirically.

Proof. Let $\zeta_{ab}^{(l)} = \chi_{c^*}^{-l}\epsilon_{ab}^{(l)}$. We will show $\zeta_{ab}^{(l)}$ is a Cauchy sequence. For any $k > l$

$$|\chi_{c^*}^{-l}\epsilon_{ab}^{(l)} - \chi_{c^*}^{-k}\epsilon_{ab}^{(k)}| \leq \sum_{j=l}^{\infty}|\chi_{c^*}^{-j}\epsilon^{(j)} - \chi_{c^*}^{-j-1}\epsilon_{ab}^{(j+1)}| = \mathcal{O}(\sum_{j=l}^{\infty}\chi_{c^*}^{-(j+1)}(\epsilon_{ab}^{(j)})^2) \lesssim \epsilon_{ab}^{(0)}\sum_{j=l}^{\infty}\chi_{c^*}^{j-1} \lesssim \chi_{c^*}^l \tag{41}$$

Thus $\zeta_{ab} \equiv \lim_{l\to\infty}\chi_{c^*}^{-l}\epsilon_{ab}^{(l)}$ exists and

$$|\zeta_{ab}^{(l)} - \zeta_{ab}| \lesssim \chi_{c^*}^l. \tag{42}$$

Equation 37 gives

$$\chi_{c^*}^{-(l+1)}\delta_{ab}^{(l+1)} = (1 + \frac{\chi_{c^*,2}}{\chi_{c^*}}p_{ab}^*)(\chi_{c^*}^{-(l+1)})\epsilon_{ab}^{(l+1)} + \chi_{c^*}^{-l}\delta_{ab}^{(l)} + \chi_{c^*}^{-(l+1)}\mathcal{O}((\epsilon_{ab}^{(l)})^2) \tag{43}$$

Let $\eta_{ab}^{(l)} = \chi_{c^*}^{-l}\delta_{ab}^{(l)} - l(1 + \frac{\chi_{c^*,2}}{\chi_{c^*}}p_{ab}^*)\zeta_{ab}$. Coupled the above equation with Equation 41, we have

$$|\eta_{ab}^{(l+1)} - \eta_{ab}^{(l)}| \lesssim \chi_{c^*}^l \tag{44}$$

Summing over all l implies

$$|\chi_{c^*}^{-l}\delta_{ab}^{(l)} - l(1 + \frac{\chi_{c^*,2}}{\chi_{c^*}}p_{ab}^*)\zeta_{ab}| \lesssim \chi_{c^*}^l. \tag{45}$$

□

A.1.2. The Spectrum of the NNGP and NTK

We consider the spectrum of $\mathcal{K}$ and Θ in this phase. For $\mathcal{K}^{(l)}$, we have $q_{ab}^* = c^*q^*$ (with $c^* < 1$), $q^{(l)} = q^*$ and $q_{ab}^{(l)} = q_{ab}^* + \mathcal{O}(\chi_{c^*}^l)$. Thus

$$\mathcal{K}^{(l)} = \mathcal{K}^* + \mathcal{E}^{(l)} \tag{46}$$

where

$$\mathcal{K}^* = q^*(c^*\mathbf{1}\mathbf{1}^T + (1 - c^*\mathbf{Id})) \tag{47}$$

$$\mathcal{E}_{ij}^{(l)} = \mathcal{O}(\chi_{c^*}^l) \tag{48}$$

The NNGP $\mathcal{K}^*$ has two different eigenvalues: $q^*(1 + (m-1)c^*)$ of order 1 and $q^*(1 - c^*)$ of order $(m-1)$, where m is the size of the dataset. For large l, since the spectral norm of $\mathcal{E}^l$ is $\mathcal{O}(\chi_{c^*}^l)$, the spectrum and condition number of $\mathcal{K}^{(l)}$ are

$$\lambda_{\max}(\mathcal{K}^{(l)}) = q^*(1 + (m-1)c^*) + \mathcal{O}(\chi_{c^*}^l) \tag{49}$$

$$\lambda_{\mathrm{bulk}}(\mathcal{K}^{(l)}) = q^*(1 - c^*) + \mathcal{O}(\chi_{c^*}^l) \tag{50}$$

$$\kappa(\mathcal{K}^{(l)}) = \frac{(1 + (m-1)c^*)}{1 - c^*} + \mathcal{O}(\chi_{c^*}^l). \tag{51}$$

For $\Theta^{(l)}$, we have $p_{ab}^{(l)} = p_{ab}^* + \mathcal{O}(l\chi_{c^*}^l) \to p_{ab}^* < \infty$ and $p^{(l)} = \frac{1-\chi_1^l}{1-\chi_1}q^* \to \infty$, i.e.

$$(p^{(l)})^{-1}\Theta^{(l)} = \mathbf{Id} + \mathcal{O}((p^{(l)})^{-1}) \tag{52}$$

Thus $\Theta^{(l)}$ is essentially a diverging constant multiplying the identity and

$$\lambda_{\max}(\Theta^{(l)}) = p^{(l)} + \mathcal{O}(1) \tag{53}$$

$$\lambda_{\text{bulk}}(\Theta^{(l)}) = p^{(l)} + \mathcal{O}(1) \tag{54}$$

$$\kappa(\Theta^{(l)}) = 1 + \mathcal{O}((p^{(l)})^{-1}) \tag{55}$$

A.2. Ordered Phase

A.2.1. The correction of the diagonal/off-diagonal

In the ordered phase, $q_{ab}^{(l)} \to q^*$, $q^{(l)} = q^*$, $p^{(l)} \to p^*$ and $p_{ab}^{(l)} \to p^*$. Indeed, letting $l \to \infty$ in the equations 24 and 23,

$$q^* = \sigma_\omega^2 \mathcal{T}(q^*) + \sigma_b^2 \tag{56}$$

$$p^* = \frac{1}{1-\chi_1} q^* \tag{57}$$

The correction of the diagonal terms are $p^{(l)} = p^* - \frac{\chi_1^l - \chi_1}{1-\chi_1} q^*$. Same calculation as in the chaotic phase implies

$$\epsilon_{ab}^{(l)} \approx \chi_1^l \, \epsilon_{ab}^{(0)} \tag{58}$$

$$\delta_{ab}^{(l)} \approx \chi_1^l \left[\delta_{ab}^{(0)} + l\left(1 + \frac{\chi_{1,2}}{\chi_1} p_{ab}^*\right) \epsilon_{ab}^{(0)}\right]. \tag{59}$$

where $\chi_{1,2} = \sigma_\omega^2 \ddot{\mathcal{T}}(q^*)$. Note that $\delta_{ab}^{(l)}$ contains also a polynomial correction term and decays like $l\chi_1^l$.

Similar, in the ordered phase we have the following.

Lemma 2. *There exists ζ_{ab} such that*

$$|\chi_1^{-l}\epsilon_{ab}^{(l)} - \zeta_{ab}| \lesssim \chi_1^l \quad \textit{and} \quad |\chi_1^{-l} l^{-1} \delta_{ab}^{(l)} - (1 + \frac{\chi_{1,2}}{\chi_1} p_{ab}^*)\zeta_{ab}| \lesssim \chi_1^l \tag{60}$$

Therefore the following limits exist

$$\lim_{l\to\infty} \chi_1^{-l}(\mathcal{K}^{(l)} - \mathcal{K}^*) \quad \textit{and} \quad \lim_{l\to\infty} \chi_1^{-l} l^{-1}(\Theta^{(l)} - \Theta^*) \tag{61}$$

Since the proof is almost identical to Lemma 1, we omit the details.

A.2.2. The Spectrum of the NNGP and NTK

For $\mathcal{K}^{(l)}$, we have $q_{ab}^* = q^*$, $q_{ab}^{(l)} = q^* + \mathcal{O}(\chi_1^l)$ and $q^{(l)} = q^*$. Thus

$$\mathcal{K}^{(l)} = q^* \mathbf{1}\mathbf{1}^T + \mathcal{O}(\chi_1^l) \tag{62}$$

which implies

$$\lambda_{\max}(\mathcal{K}^{(l)}) = mq^* + \mathcal{O}(\chi_1^l) \tag{63}$$

$$\lambda_{\text{bulk}}(\mathcal{K}^{(l)}) = \mathcal{O}(\chi_1^l) \tag{64}$$

$$\kappa(\mathcal{K}^{(l)}) \gtrsim \chi_1^{-l} \tag{65}$$

For $\Theta^{(l)}$, $p_{ab}^{(l)} = p^* + \mathcal{O}(l\chi_1^l)$ and $p^{(l)} = p^* - \frac{\chi_1^l - \chi_1}{1-\chi_1} q^* = p^* + \mathcal{O}(\chi_1^l)$. Thus

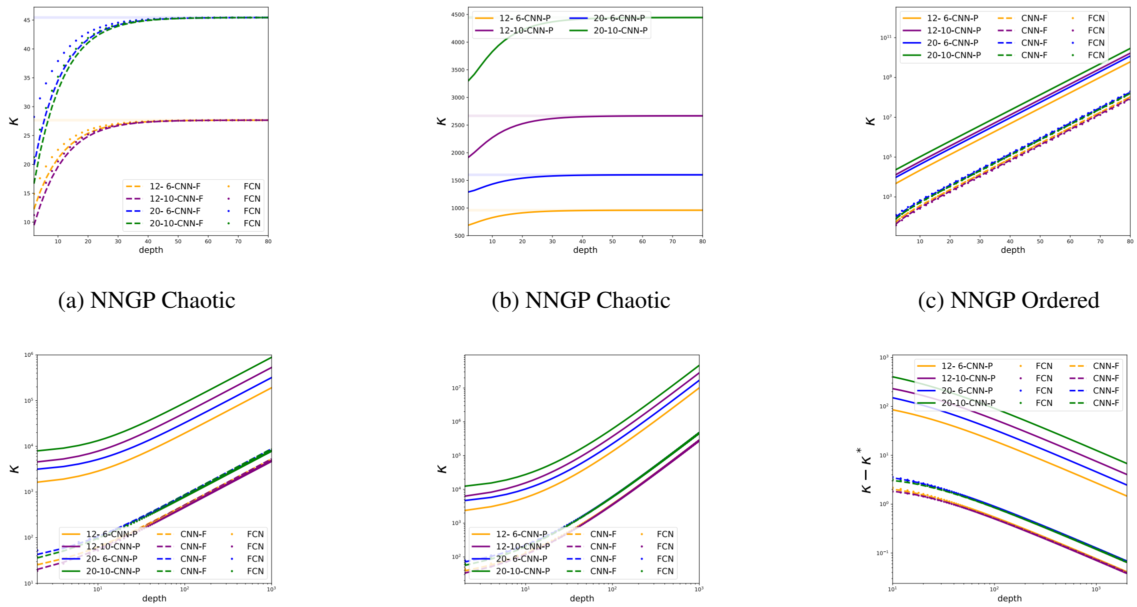
$$\Theta^{(l)} = p^* \mathbf{1}\mathbf{1}^T + \mathcal{O}(l\chi_1^l) \tag{66}$$

which implies

$$\lambda_{\max}(\Theta^{(l)}) = mp^* + \mathcal{O}(l\chi_1^l) \tag{67}$$

$$\lambda_{\text{bulk}}(\Theta^{(l)}) = \mathcal{O}(l\chi_1^l) \tag{68}$$

$$\kappa(\Theta^{(l)}) \gtrsim (l\chi_1^l)^{-1} \tag{69}$$



(a) NNGP Chaotic

(b) NNGP Chaotic

(c) NNGP Ordered

(d) NNGP Critical

(e) NNGP RELU

(f) NTK RELU

Figure 3. **Condition numbers of NNGP and their rate of convergence.** In the chaotic phase, $\kappa(\mathcal{K}^{(l)})$ converges to a constant (see Table 2) for FCN, CNN-F (a) and CNN-P (b). However, it diverges exponentially in the ordered phase (c) and linearly on the critical line (d). For critical RELU network, $\kappa(\mathcal{K}^{(l)})$ diverges quadratically (e) while $\kappa(\Theta^{(l)})$ converges to a fixed number with rate (l^{-1}) (see Equation 92) and we plot the value of $(\kappa(\Theta^{(l)}) - \kappa(\Theta^*))$ of the NTK in (f).

A.3. The critical line.

A.3.1. Correction of the diagonals/off-diagonals.

We have $\chi_1 = 1$ on the critical line. Equation 24 implies $p^{(l)} = lq^*$, i.e. the diagonal terms diverge linearly. To capture the linear divergence of $p_{ab}^{(l)}$, define

$$\epsilon_{ab}^{(l)} = q_{ab}^{(l)} - q_{ab}^* \tag{70}$$

$$\delta_{ab}^{(l)} = p_{ab}^{(l)} - lq^* \tag{71}$$

We need to expand the first equation of 23 to the second order

$$\epsilon_{ab}^{(l+1)} = \epsilon_{ab}^{(l)} + \frac{1}{2}\chi_{1,2}\left(\epsilon_{ab}^{(l)}\right)^2 + \mathcal{O}((\epsilon_{ab}^{(l)})^3) \tag{72}$$

Here we assume $\mathcal{T}$ has a continuous third derivative (which is sufficient to assume the activation ϕ to have a continuous third derivative.) The above equation implies

$$\epsilon_{ab}^{(l)} = -\frac{2}{\chi_{1,2}}\frac{1}{l} + o(\frac{1}{l}). \tag{73}$$

Then

$$\delta_{ab}^{(l+1)} = q_{ab}^{(l+1)} - q^* + \sigma_\omega^2 \dot{\mathcal{T}}(q^* + \epsilon_{ab}^{(l)})p_{ab}^{(l)} - lq^* \tag{74}$$

$$= \epsilon_{ab}^{(l+1)} + (\chi_1 + \chi_{1,2}\epsilon_{ab}^{(l)} + \mathcal{O}(\epsilon_{ab}^{(l)})^2)(lq^* + \delta_{ab}^{(l)}) - lq* \tag{75}$$

$$= \epsilon_{ab}^{(l+1)} + (1 + \chi_{1,2}\epsilon_{ab}^{(l)})\delta_{ab}^{(l)} + lq^*\chi_{1,2}\epsilon_{ab}^{(l)} + \mathcal{O}((\epsilon_{ab}^{(l)})^2)lq^* \tag{76}$$

Plugging Equation 73 into the above equation gives

$$\delta_{ab}^{(l)} = -\frac{2}{3} l q^* + \mathcal{O}(1)\,. \tag{77}$$

A.3.2. The Spectrum of NNGP and NTK

For $\mathcal{K}^{(l)}$, $q_{ab}^{(l)} = q^* + \mathcal{O}(l^{-1})$ and $q^{(l)} = q^*$. Thus

$$\lambda_{\max}(\mathcal{K}^{(l)}) = m q^* + \mathcal{O}(1/l) \tag{78}$$

$$\lambda_{\text{bulk}}(\mathcal{K}^{(l)}) = \mathcal{O}(1/l) \tag{79}$$

$$\kappa(\mathcal{K}^{(l)}) \gtrsim l \tag{80}$$

For $\Theta^{(l)}$, $p_{ab}^{(l)} = \frac{1}{3} q^* l + \mathcal{O}(1)$ and $p^{(l)} = l q^*$. Thus

$$\lambda_{\max}(\Theta^{(l)}) = \frac{m+2}{3} l q^* + \mathcal{O}(1) \tag{81}$$

$$\lambda_{\text{bulk}}(\Theta^{(l)}) = \frac{2}{3} l q^* + \mathcal{O}(1) \tag{82}$$

$$\kappa(\Theta^{(l)}) = \frac{m+2}{2} + \mathcal{O}(1/l) \tag{83}$$

B. NNGP and NTK of Relu networks.

B.1. Critical Relu.

We only consider the critical initialization (i.e. He's initialization (He et al., 2015)) $\sigma_\omega^2 = 2$ and $\sigma_b^2 = 0$, which preserves the norm of an input from layer to layer. We also normalize the inputs to have unit variance, i.e. $q^* = q^{(l)} = q^{(0)} = 1$. Recall that

$$\mathcal{K}^{(l+1)} = 2\mathcal{T}(\mathcal{K}^{(l)}) \tag{84}$$

$$\Theta^{(l+1)} = \mathcal{K}^{(l+1)} + 2\dot{\mathcal{T}}(\mathcal{K}^{(l)}) \odot \Theta^{(l)} \tag{85}$$

This implies

$$p^{(l+1)} = q^{(l)} + 2\dot{\mathcal{T}}(q^{(l)}) p^{(l)} = 1 + 2\dot{\mathcal{T}}(1) p^{(l)} = 1 + p^{(l)} \tag{86}$$

which gives $p^{(l)} = l$. Using the equations in Appendix C of (Lee et al., 2019) gives

$$2\mathcal{T}(1-\epsilon) = 1 - \epsilon + \frac{2\sqrt{2}}{3\pi}\epsilon^{3/2} + \mathcal{O}(\epsilon^{5/2}) \tag{87}$$

and taking the derivative w.r.t. ϵ

$$2\dot{\mathcal{T}}(1-\epsilon) = 1 - \frac{\sqrt{2}}{\pi}\epsilon^{1/2} + \mathcal{O}(\epsilon^{3/2}) \quad \text{as} \quad \epsilon \to 0^+. \tag{88}$$

Thus

$$1 - \epsilon_{ab}^{(l+1)} = 1 - \epsilon_{ab}^{(l)} + \frac{2\sqrt{2}}{3\pi}(\epsilon_{ab}^{(l)})^{3/2} + \mathcal{O}((\epsilon_{ab}^{(l)})^{5/2}) \tag{89}$$

This is enough to conclude (similar to the above calculation)

$$\epsilon_{ab}^{(l)} = (\frac{3\pi}{\sqrt{2}})^2 l^{-2} + o(l^{-2}) \tag{90}$$

and

$$p_{ab}^{(l)} - p^{(l)} = -\frac{3}{4}l + \mathcal{O}(1). \tag{91}$$

Recall that the diagonals of $\mathcal{K}^{(l)}$ and $\Theta^{(l)}$ are $q^{(l)} = 1$ and $p^{(l)} = l$, resp. Therefore the spectrum and the condition numbers of $\mathcal{K}^{(l)}$ and $\Theta^{(l)}$ for large l are

$$\begin{cases} \lambda_{\max}(\Theta^{(l)}) & = \frac{m+3}{4}l + \mathcal{O}(1) \\ \lambda_{\text{bulk}}(\Theta^{(l)}) & = \frac{3}{4}l + \mathcal{O}(1) \\ \kappa(\Theta^{(l)}) & = \frac{m+3}{3} + \mathcal{O}(1/l) \end{cases} \qquad \begin{cases} \lambda_{\max}(\mathcal{K}^{(l)}) & = m + \mathcal{O}(l^{-2}) \\ \lambda_{\text{bulk}}(\mathcal{K}^{(l)}) & = \mathcal{O}(l^{-2}) \\ \kappa(\mathcal{K}^{(l)}) & \gtrsim \mathcal{O}(l^2) \end{cases} \tag{92}$$

B.2. Residual Relu

We consider the following "continuum" residual network

$$x^{(t+dt)} = x^{(t)} + (dt)^{1/2}(W\phi(x^{(t)}) + b) \tag{93}$$

where t denotes the 'depth' and $dt > 0$ is sufficiently small and W and b are the weights and biases. We also set $\sigma_\omega^2 = 2$ (i.e. $\mathbb{E}[WW^T] = 2\textbf{Id}$) and $\sigma_b^2 = 0$ (i.e. $b = 0$). The NNGP and NTK have the following form

$$\mathcal{K}^{(t+dt)} = \mathcal{K}^{(t)} + 2dt\mathcal{T}(\mathcal{K}^{(t)}) \tag{94}$$

$$\Theta^{(t+dt)} = \Theta^{(t)} + 2dt\mathcal{T}(\mathcal{K}^{(t)}) + 2dt\dot{\mathcal{T}}(\mathcal{K}^{(t)}) \odot \Theta^{(t)} \tag{95}$$

Taking the limit $dt \to 0$ gives

$$\dot{\mathcal{K}}^{(t)} = 2\mathcal{T}(\mathcal{K}^{(t)}) \tag{96}$$

$$\dot{\Theta}^{(t)} = 2\mathcal{T}(\mathcal{K}^{(t)}) + 2\dot{\mathcal{T}}(\mathcal{K}^{(t)}) \odot \Theta^{(t)} \tag{97}$$

Using the fact that $q^{(0)} = 1$ (i.e. the inputs have unit variance), we can compute the diagonal terms $q^{(t)} = e^t$ and $p^{(t)} = te^t$. Letting $q_{ab}^{(t)} = e^t c_{ab}^{(t)}$ and applying the above fractional Taylor expansion to $\mathcal{T}$ and $\dot{\mathcal{T}}$, we have

$$\dot{c}_{ab}^{(t)} = -\frac{2\sqrt{2}}{3\pi}(1 - c_{ab}^{(t)})^{\frac{3}{2}} + O((1 - c_{ab}^{(t)})^{\frac{5}{2}}) \tag{98}$$

Ignoring the higher order term and set $y(t) = (1 - c_{ab}^{(t)})$, we have

$$\dot{y} = \frac{2\sqrt{2}}{3\pi}y^{\frac{3}{2}}\,. \tag{99}$$

Solving this gives $y(t) = \frac{9\pi^2}{2}t^{-2}$ (note that $y(\infty) = 0$), which implies

$$q_{ab}^{(t)} = (1 - \frac{9\pi^2}{2}t^{-2} + o(t^{-2}))e^t\,. \tag{100}$$

Applying this estimate to Equation 97 gives

$$p_{ab}^{(t)} = (\frac{1}{4}t + \mathcal{O}(1))e^t\,. \tag{101}$$

Thus the limiting condition number of the NTK is $m/3 + 1$. This is the same as the above non-residual Relu case although the entries of $\mathcal{K}^{(t)}$ and $\Theta^{(t)}$ blow up exponentially with t.

B.3. Residual Relu + Layer Norm

As we saw above, all the entries of $\mathcal{K}^{(l)}$ and $\Theta^{(l)}$ of a residual Relu network blow up exponentially, so do its gradients. In what follows, we show that normalization could help to avoid this issue. We consider the following "continuum" residual network with "layer norm"

$$x^{(t+dt)} = \frac{1}{\sqrt{1+dt}}\left(x^{(t)} + (dt)^{1/2} W\phi(x^{(t)})\right) \tag{102}$$

We also set $\sigma_\omega^2 = 2$ (i.e. $\mathbb{E}[WW^T] = 2\mathbf{Id}$). The normalization term $\frac{1}{\sqrt{1+dt}}$ makes sure $x^{(t+dt)}$ has unit norm and removes the exponentially factor e^t in both NNGP and NTK. To ses this, note that

$$\mathcal{K}^{(t+dt)} = \frac{1}{1+dt}\left(\mathcal{K}^{(t)} + 2dt\mathcal{T}(\mathcal{K}^{(t)})\right) \tag{103}$$

$$\Theta^{(t+dt)} = \frac{1}{1+dt}\left(\Theta^{(t)} + dt\mathcal{K}^{(t)} + 2dt\dot{\mathcal{T}}(\mathcal{K}^{(t)})\Theta^{(t)}\right) \tag{104}$$

Taking the limit $dt \to 0$ gives

$$\dot{\mathcal{K}}^{(t)} = -\mathcal{K}^{(t)} + 2\mathcal{T}(\mathcal{K}^{(t)}) \tag{105}$$

$$\dot{\Theta}^{(t)} = 2\mathcal{T}(\mathcal{K}^{(t)}) + 2\dot{\mathcal{T}}(\mathcal{K}^{(t)}) \odot \Theta^{(t)} \tag{106}$$

Using the fact that $q^{(0)} = 1$ (i.e. the inputs have unit variance) and the mapping $2\mathcal{T}$ is norm preserving, we see that $q^{(t)} = 1$ because

$$\dot{q}^{(t)} = -q^{(t)} + 2\mathcal{T}(q^{(t)}) = 0. \tag{107}$$

This implies $p^{(t)} = t$ (note that $\dot{p}^{(t)} = q^{(t)} = 1$ and we assume the initial value $p^{(0)} = 0$.) The off-diagonal terms can be computed similarly and

$$q_{ab}^{(t)} = 1 - \frac{9\pi^2}{2}t^{-2} + o(t^{-2}) \tag{108}$$

$$p_{ab}^{(t)} = \frac{1}{4}t + \mathcal{O}(1)\,. \tag{109}$$

Thus the condition number of the NTK is $m/3 + 1$. This is the same as the non-residual Relu case discussed above.

C. Asymptotic of $P(\Theta^{(l)})$

To keep the notation simple, we denote $X_d = X_{\text{train}}$, $Y_d = Y_{\text{train}}$, $\Theta_{td} = \Theta_{\text{test, train}}$, $\Theta_{dd} = \Theta_{\text{train, train}}$. Recall that

$$P(\Theta^{(l)})Y_d = \left(\Theta_{td}^{(l)}\left(\Theta_{dd}^{(l)}\right)^{-1}\right)Y_d \tag{110}$$

We split our calculation into three parts.

C.1. Chaotic phase

In this case the diagonal $p^{(l)}$ diverges exponentially and the off-diagonals $p_{ab}^{(l)}$ converges to a bounded constant p_{ab}^*. We further assume the input labels are centered in the sense Y_d contains the same number of positive (+1) and negative (-1) labels[4]. We expand $\Theta^{(l)}$ about its “fixed point”

$$P(\Theta^{(l)})Y_d = \Theta^{(l)}_{td}\left(\Theta^{(l)}_{dd}\right)^{-1}Y_d \tag{111}$$

$$= \left(\Theta^*_{td} + \mathcal{O}(\delta^{(l)}_{ab})\right)\left(p^{(l)}\mathbf{Id} + p^*_{ab}(\mathbf{1}\mathbf{1}^T - \mathbf{Id}) + \mathcal{O}(\delta^{(l)}_{ab})\right)^{-1}Y_d \tag{112}$$

$$= (p^{(l)})^{-1}\left(\Theta^*_{td} + \mathcal{O}(\delta^{(l)}_{ab})\right)\left(\mathbf{Id} - \frac{p^*_{ab}}{p^{(l)}}(\mathbf{1}\mathbf{1}^T - \mathbf{Id}) + \mathcal{O}(\delta^{(l)}_{ab}/p^{(l)})\right)Y_d \tag{113}$$

$$= (p^{(l)})^{-1}\left(\Theta^*_{td} + \mathcal{O}(\delta^{(l)}_{ab})\right)\left(\mathbf{Id} - \frac{p^*_{ab}}{p^{(l)}}(\mathbf{1}\mathbf{1}^T - \mathbf{Id}) + \mathcal{O}(\delta^{(l)}_{ab}/p^{(l)})\right)Y_d \tag{114}$$

$$= (p^{(l)})^{-1}\left(\mathcal{O}(\delta^{(l)}_{ab}) + \mathcal{O}(\delta^{(l)}_{ab}/p^{(l)})\right)Y_d \tag{115}$$

In the last equation, we have used the fact $\mathbf{1}\mathbf{1}^TY_d = \mathbf{0}$ and $\Theta^*_{td}Y_d = \mathbf{0}$ since Y_d is balanced. Therefore

$$P(\Theta^{(l)})Y_d = \mathcal{O}((p^{(l)})^{-1}\delta^{(l)}_{ab}) = \mathcal{O}(l(\chi_{c^*}/\chi_1)^l)\,. \tag{116}$$

Remark 1. *Without centering the labels Y_d and normalizing each input in X_d to have the same variance, we will get a χ_1^l decay for $P(\Theta^{(l)})Y_d$ instead of $l(\chi_{c^*}/\chi_1)^l$.*

C.2. Critical line

Note that in this phase, both the diagonals and the off-diagonals diverge linearly. In this case

$$\lim_{l\to\infty}\frac{1}{lq^*}\Theta^{(l)}_{td} = \frac{1}{3}\mathbf{1}_t\mathbf{1}_d^T \qquad \lim_{l\to\infty}\frac{1}{lq^*}\Theta^{(l)}_{dd} = B \equiv \frac{2}{3}\mathbf{Id} + \frac{1}{3}\mathbf{1}_d\mathbf{1}_d^T \tag{117}$$

Here we use $\mathbf{1}_d$ to denote the all ‘1’ (column) vector with length equal to the number of training points in X_d and $\mathbf{1}_t$ is defined similarly. Note that the constant matrix B is invertible. By Equation 77

$$P(\Theta^{(l)}) = \frac{1}{3}\left(\frac{3}{lq^*}\Theta^{(l)}_{td}\right)\left(\frac{1}{lq^*}\Theta^{(l)}_{dd}\right)^{-1} \tag{118}$$

$$= \frac{1}{3}\left(\mathbf{1}_t\mathbf{1}_d^T + \mathcal{O}(1/lq^*)\right)\left(B + \mathcal{O}(1/lq^*)\right)^{-1} \tag{119}$$

$$= \frac{1}{3}\left(\mathbf{1}_t\mathbf{1}_d^T + \mathcal{O}(1/lq^*)\right)\left(B^{-1} + \mathcal{O}(1/lq^*)\right) \tag{120}$$

$$= \frac{1}{3}\mathbf{1}_t\mathbf{1}_d^TB^{-1} + \mathcal{O}(1/lq^*) \tag{121}$$

The term $\mathbf{1}_t\mathbf{1}_d^TB^{-1}$ is independent of the inputs and $\mathbf{1}_t\mathbf{1}_d^TB^{-1}Y_d = 0$ when Y_d is centered. Thus

$$P(\Theta^{(l)})Y_d = \mathcal{O}(1/lq^*) \tag{122}$$

C.3. Ordered Phase

In the ordered phase, we have that $\Theta^{(l)}_{dd} = p^*\mathbf{1}_d\mathbf{1}_d^T + l\chi_1^l\boldsymbol{A}^{(l)}_{dd}$ where $\boldsymbol{A}^{(l)}_{dd}$, a symmetric matrix, represents the data-dependent piece of $\Theta^{(l)}_{dd}$. By Lemma 2, $\boldsymbol{A}^{(l)}_{dd} \to \boldsymbol{A}_{dd}$ as $l \to \infty$. To simply the notation, in the calculation below we will replace $\boldsymbol{A}^{(l)}_{dd}$ by $\boldsymbol{A}_{dd}$. We also assume $\boldsymbol{A}_{dd}$ is invertible. To compute the mean predictor, $P(\Theta^{(l)})$, asymptotically we begin by computing $(\Theta^{(l)}_{dd})^{-1}$ via the Woodbury identity,

$$(\Theta^{(l)}_{dd})^{-1} = \left(p^*\mathbf{1}_d\mathbf{1}_d^T + l\chi_1^l\boldsymbol{A}_{dd}\right)^{-1} \tag{123}$$

$$= l^{-1}\chi_1^{-l}\left[\boldsymbol{A}_{dd}^{-1} - \boldsymbol{A}_{dd}^{-1}\mathbf{1}_d\left(\frac{1}{p^*} + \frac{\mathbf{1}_m^T\boldsymbol{A}_{dd}^{-1}\mathbf{1}_d}{l\chi_1^l}\right)^{-1}\mathbf{1}_d^T\boldsymbol{A}_{dd}^{-1}l^{-1}\chi_1^{-1}\right] \tag{124}$$

$$= l^{-1}\chi_1^{-l}\left[\boldsymbol{A}_{dd}^{-1} - \hat{p}\boldsymbol{A}_{dd}^{-1}\mathbf{1}_d\mathbf{1}_d^T\boldsymbol{A}_{dd}^{-1}\right] \tag{125}$$

$$= l^{-1}\chi_1^{-l}\left[\boldsymbol{A}_{dd}^{-1} - \hat{p}\boldsymbol{a}\boldsymbol{a}^T\right] \tag{126}$$

[4] When the number of classes is greater than two, we require Y_d to have mean zero along the batch dimension for each class.

where we have set

$$\boldsymbol{a} = \boldsymbol{A}_{dd}^{-1}\mathbf{1}_d \quad \text{and} \quad \hat{p} = \frac{p^*}{l\chi_1^l + p^*\mathbf{1}_d^T\boldsymbol{A}_{dd}^{-1}\mathbf{1}_d} = \frac{p^*}{l\chi_1^l + p^*\mathbf{1}_d^T\boldsymbol{a}} \tag{127}$$

and $\boldsymbol{a}_i = \frac{1}{m}\sum_j \boldsymbol{A}_{ij}^{-1}$. Noting that $\Theta_{td}^{(l)} = p^*\mathbf{1}_t\mathbf{1}_d^T + l\chi_1^l\boldsymbol{A}_{td}$ we can compute the mean predictor,

$$\begin{aligned} P(\Theta^{(l)}) = \Theta_{td}^l(\Theta_{dd}^l)^{-1} &= (p^*\mathbf{1}_t\mathbf{1}_d^T + l\chi_1^l\boldsymbol{A}_{td})l^{-1}\chi_1^{-l}\left[\boldsymbol{A}_{dd}^{-1} - \hat{p}\boldsymbol{a}\boldsymbol{a}^T\right] & (128)\\ &= \boldsymbol{A}_{td}\boldsymbol{A}_{dd}^{-1} - \hat{p}\boldsymbol{A}_{td}\boldsymbol{a}\boldsymbol{a}^T + l^{-1}\chi_1^{-l}p^*(\mathbf{1}_t\mathbf{1}_d^T\boldsymbol{A}_{dd}^{-1} - \hat{p}\mathbf{1}_t\mathbf{1}_d^T\boldsymbol{a}\boldsymbol{a}^T) & (129)\\ &= \boldsymbol{A}_{td}\boldsymbol{A}_{dd}^{-1} - \hat{p}\boldsymbol{A}_{td}\boldsymbol{a}\boldsymbol{a}^T + l^{-1}\chi_1^{-l}p^*(1 - \hat{p}\mathbf{1}_d^T\boldsymbol{a})\mathbf{1}_t\boldsymbol{a}^T & (130)\\ &= \boldsymbol{A}_{td}\boldsymbol{A}_{dd}^{-1} - \hat{p}\boldsymbol{A}_{td}\boldsymbol{a}\boldsymbol{a}^T + \hat{p}\mathbf{1}_t\boldsymbol{a}^T & (131)\end{aligned}$$

Note that there is no divergence in $P(\Theta^{(l)})$ as $l \to \infty$ and the limit is well-defined. The term $\hat{p}\mathbf{1}_t\boldsymbol{a}^T$ is independent from the input data.

$$\lim_{l\to\infty} P(\Theta^{(l)})Y_{\text{train}} = (\boldsymbol{A}_{td}\boldsymbol{A}_{dd}^{-1} - \hat{p}\boldsymbol{A}_{td}\boldsymbol{a}\boldsymbol{a}^T + \hat{p}\mathbf{1}_t\boldsymbol{a}^T)Y_{\text{train}} \equiv (\boldsymbol{A}_{td}\boldsymbol{A}_{dd}^{-1} + \hat{\boldsymbol{A}})Y_{\text{train}} \tag{132}$$

We therefore see that even in the infinite-depth limit the mean predictor retains its data-dependence and we expect these networks to be able generalize indefinitely.

D. Dropout

In this section, we investigate the effect of adding a dropout layer to the penultimate layer. Let $0 < \rho \leq 1$ and $\gamma_j^{(L)}(x)$ be iid random variables

$$\gamma_j^{(L)}(x) = \begin{cases} 1, & \text{with probability} \quad \rho \\ 0, & \text{with probability} \quad 1-\rho. \end{cases} \tag{133}$$

For $0 \leq l \leq L-1$,

$$z_i^{(l+1)}(x) = \frac{\sigma_w}{\sqrt{N^{(l)}}}\sum_j W_{ij}^{(l+1)}\phi(z_j^{(l)}(x)) + \sigma_b b_i^{(l+1)} \tag{134}$$

and for the output layer,

$$z_i^{(L+1)}(x) = \frac{\sigma_w}{\rho\sqrt{N^{(L)}}}\sum_{j=1}^{N^{(L)}} W_{ij}^{(L+1)}\phi(z_j^{(L)}(x))\gamma_j^{(L)}(x) + \sigma_b b_i^{(L+1)} \tag{135}$$

where $W_{ij}^{(l)}$ and $b_i^{(l)}$ are iid Gaussians $\mathcal{N}(0,1)$. Since no dropout is applied in the first L layers, the NNGP kernel $\mathcal{K}^{(l)}$ and $\Theta^{(l)}$ can be computed using Equation 20 and Equation 8. Let $\mathcal{K}_\rho^{(L+1)}$ and $\Theta_\rho^{(L+1)}$ denote the NNGP and NTK of the $(L+1)$-th layer. Note that when $\rho = 1$, $\mathcal{K}_1^{(L+1)} = \mathcal{K}^{(L+1)}$ and $\Theta_1^{(L+1)} = \Theta^{(L+1)}$. We will compute the correction induced by $\rho < 1$. The fact

$$\mathbb{E}[\gamma_j^{(L)}(x)\gamma_i^{(L)}(x')] = \begin{cases} \rho^2, & \text{if} \quad (j,x) \neq (i,x') \\ \rho, & \text{if} \quad (j,x) = (i,x') \end{cases} \tag{136}$$

implies that the NNGP kernel $\mathcal{K}_\rho^{(L+1)}$ (Schoenholz et al., 2017) is

$$\mathcal{K}_\rho^{(L+1)}(x,x') \equiv \mathbb{E}[z_i^{(L+1)}(x)z_i^{(L+1)}(x')] = \begin{cases} \sigma_w^2\mathcal{T}(\mathcal{K}^{(L)}(x,x')) + \sigma_b^2, & \text{if} \quad x \neq x' \\ \\ \frac{1}{\rho}\sigma_w^2\mathcal{T}(\mathcal{K}^{(L)}(x,x)) + \sigma_b^2 & \text{if} \quad x = x'\,. \end{cases} \tag{137}$$

Now we compute the NTK $\Theta_\rho^{(L+1)}$, which is a sum of two terms

$$\Theta_\rho^{(L+1)}(x,x') = \mathbb{E}\left[\frac{\partial z_i^{(L+1)}(x)}{\partial\theta^{(L+1)}}\left(\frac{\partial z_i^{(L+1)}(x')}{\partial\theta^{(L+1)}}\right)^T\right] + \mathbb{E}\left[\frac{\partial z_i^{(L+1)}(x)}{\partial\theta^{(\le L)}}\left(\frac{\partial z_i^{(L+1)}(x')}{\partial\theta^{(\le L)}}\right)^T\right]. \tag{138}$$

Here $\theta^{(L+1)}$ denote the parameters in the $(L+1)$ layer, namely, $W_{ij}^{(L+1)}$ and $b_i^{(L+1)}$ and $\theta^{(\le L)}$ the remaining parameters. Note that the first term in Equation 138 is equal to $\mathcal{K}_\rho^{(L+1)}(x,x')$. Using the chain rule, the second term is equal to

$$\frac{\sigma_\omega^2}{\rho^2 N^{(L)}}\mathbb{E}\left[\sum_{j,k=1}^{N^{(L)}} W_{ij}^{(L+1)}W_{ik}^{(L+1)}\dot\phi(z_j^{(L)}(x))\gamma_j^{(L)}(x)\dot\phi(z_k^{(L)}(x'))\gamma_j^{(L)}(x')\frac{\partial z_j^{(L)}(x)}{\partial\theta^{(\le L)}}\left(\frac{\partial z_k^{(L)}(x')}{\partial\theta^{(\le L)}}\right)^T\right] \tag{139}$$

$$=\frac{\sigma_\omega^2}{\rho^2 N^{(L)}}\mathbb{E}\left[\sum_{j}^{N^{(L)}} \dot\phi(z_j^{(L)}(x))\gamma_j^{(L)}(x)\dot\phi(z_j^{(L)}(x'))\gamma_j^{(L)}(x')\frac{\partial z_j^{(L)}(x)}{\partial\theta^{(\le L)}}\left(\frac{\partial z_j^{(L)}(x')}{\partial\theta^{(\le L)}}\right)^T\right] \tag{140}$$

$$=\frac{\sigma_\omega^2}{\rho^2}\mathbb{E}\left[\gamma_j^{(L)}(x)\gamma_j^{(L)}(x')\right]\mathbb{E}[\dot\phi(z_j^{(L)}(x))\dot\phi(z_j^{(L)}(x'))]\mathbb{E}\left[\frac{\partial z_j^{(L)}(x)}{\partial\theta^{(\le L)}}\left(\frac{\partial z_j^{(L)}(x')}{\partial\theta^{(\le L)}}\right)^T\right] \tag{141}$$

$$=\begin{cases}\sigma_\omega^2\dot{\mathcal{T}}(\mathcal{K}^{(L)}(x,x'))\Theta^{(L)}(x,x') & \text{if} \quad x\neq x' \\ \\ \frac{1}{\rho}\sigma_\omega^2\dot{\mathcal{T}}(\mathcal{K}^{(L)}(x,x))\Theta^{(L)}(x,x) & \text{if} \quad x = x'\,.\end{cases} \tag{142}$$

In sum, we see that dropout only modifies the diagonal terms

$$\begin{cases}\Theta_\rho^{(L+1)}(x,x') = \Theta^{(L+1)}(x,x') \\ \\ \Theta_\rho^{(L+1)}(x,x) = \frac{1}{\rho}\Theta^{(L+1)}(x,x) + (1-1/\rho)\sigma_b^2\end{cases} \tag{143}$$

In the ordered phase, we see

$$\lim_{L\to\infty}\Theta_\rho^{(L)}(x,x') = p^*, \qquad \lim_{L\to\infty}\Theta_\rho^{(L)}(x,x) = \frac{1}{\rho}p^* + (1-\frac{1}{\rho})\sigma_b^2 \tag{144}$$

and the condition number

$$\lim_{L\to\infty}\kappa_\rho^{(L)} = \frac{(m-1)p^* + \frac{1}{\rho}p^* + (1-\frac{1}{\rho})\sigma_b^2}{(\frac{1}{\rho}-1)(p^*-\sigma_b^2)} = \frac{mp^*}{(\frac{1}{\rho}-1)(p^*-\sigma_b^2)} + 1 \tag{145}$$

In Fig 4, we plot the evolution of $\kappa_\rho^{(L)}$ for $\rho = 0.8, 0.95, 0.99$ and 1, confirming Equation 145.

E. Convolutions

In this section, we compute the evolution of $\Theta^{(l)}$ for CNNs.

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-or higher-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 $\mu_i^{(l)}$, respectively, with outgoing (incoming) channel index i (j) and filter relative spatial location $\beta\in[\pm k]\equiv\{-k,\dots,0,\dots,k\}$.[5] As

[5] 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.

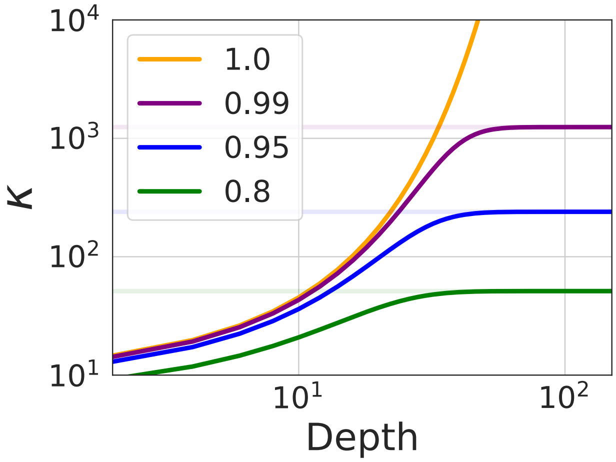


Figure 4. **Dropout improves conditioning of the NTK.** In the ordered phase, the condition number $\kappa^{(l)}$ explodes exponentially (yellow) as $l \to \infty$. However, a dropout layer could significantly improves the conditioning, making $\kappa^{(l)}$ converge to a finite constant (horizontal lines) Equation 145.

above, we will assume a Gaussian prior on both the filter weights and biases,

$$W^{(l)}_{ij,\beta} = \frac{\sigma_\omega}{\sqrt{(2k+1)N^{(l)}}}\omega^{(l)}_{ij,\beta} \qquad b^{(l)}_i = \sigma_b \mu^{(l)}_i, \qquad \omega^{(l)}_{ij,\beta}, \quad \mu^{(l)}_i \sim \mathcal{N}(0,1) \tag{146}$$

As above, σ^2_ω and σ^2_b 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 $x \in \mathcal{X} \subseteq \mathbb{R}^{N^{(0)}d}$, with input channel count $N^{(0)} \in \mathbb{N}$, number of pixels $d \in \mathbb{N}$, where

$$y^{(l)}_{i,\alpha}(x) \equiv \begin{cases} x_{i,\alpha} & l = 0 \\ \phi\left(z^{(l-1)}_{i,\alpha}(x)\right) & l > 0 \end{cases}, \qquad z^{(l)}_{i,\alpha}(x) \equiv \sum_{j=1}^{N^{(l)}} \sum_{\beta=-k}^{k} W^{(l)}_{ij,\beta} y^{(l)}_{j,\alpha+\beta}(x) + b^{(l)}_i. \tag{147}$$

$\phi : \mathbb{R} \to \mathbb{R}$ 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 $\min\{N^1 \dots, N^{(l-1)}\} \to \infty$, 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., 2019b; Xiao et al., 2018)

$$\mathcal{K}^{(l+1)} = (\sigma^2_\omega \mathcal{A} + \sigma^2_b) \circ \mathcal{T}(\mathcal{K}^{(l)}) = \left((\sigma^2_\omega \mathcal{A} + \sigma^2_b) \circ \mathcal{T}\right)^{l+1}(\mathcal{K}^0) \tag{148}$$

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 acting on $\mathcal{X}d \times \mathcal{X}d$ PSD matrices

$$[\mathcal{T}(\mathcal{K})]_{\alpha,\alpha'}(x,x') \equiv \mathbb{E}_{u \sim \mathcal{N}(0,\mathcal{K})}\left[\phi\left(u_\alpha(x)\right)\phi\left(u_{\alpha'}(x')\right)\right] \tag{149}$$

$$[\mathcal{A}(\mathcal{K})]_{\alpha,\alpha'}(x,x') \equiv \frac{1}{2k+1}\sum_\beta [\mathcal{K}]_{\alpha+\beta,\alpha'+\beta}(x,x'). \tag{150}$$

E.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 $\hat{\Theta}^{(l)}$

$$\hat{\Theta}^{(l)}_{\alpha,\alpha'}(x,x') = \nabla_{\theta^{\leq l}} z^{(l)}_{i,\alpha}(x)\left(\nabla_{\theta^{\leq l}} z^{l}_{i,\alpha'}(x')\right)^T \tag{151}$$

where $\theta^{\leq l}$ denotes all of the parameters in layers at-or-below the l'th layer. It does not matter which channel index i is used because as the number of channels approach infinity, this kernel will also 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 (Yang, 2019; Arora et al., 2019),

$$\Theta^{(l+1)} = \mathcal{K}^{(l+1)} + \mathcal{A} \circ (\sigma_\omega^2 \dot{\mathcal{T}}(\mathcal{K}^{(l)}) \odot \Theta^{(l)}), \tag{152}$$

where $\dot{\mathcal{T}}$ is given by Equation 149 with ϕ replaced by its derivative ϕ'. 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 diagonal entries of $\mathcal{K}^{(l)}$ and $\Theta^{(l)}$ are exactly the same as the fully-connected setting, which are q^* and $p^{(l)} = lq^*$, respectively. We only need to consider the off-diagonal terms. Letting $l \to \infty$ in Equation 152 we see that all the off-diagonal terms also converge p^*_{ab}. Note that $\mathcal{A}$ does not mix terms from different diagonals and it suffices to handle 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 $\Theta^{(l)}$ to the fixed points. Linearizing Equation 148 and Equation 152 gives

$$\epsilon^{(l+1)}_{ab} \approx \chi_{c^*} \mathcal{A} \epsilon^{(l)}_{ab} \tag{153}$$

$$\delta^{(l+1)}_{ab} \approx \chi_{c^*} \mathcal{A}(\epsilon^{(l+1)}_{ab} + \frac{\chi_{c^*,2}}{\chi_{c^*}} p^*_{ab} \epsilon^{(l)}_{ab} + \delta^{(l)}_{ab}) \,. \tag{154}$$

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

$$\epsilon^{(l+1)}_{ab,\alpha} \approx (\rho_\alpha \chi_{c^*})^{(l+1)} \epsilon^{(0)}_{ab,\alpha} \tag{155}$$

$$\delta^{(l+1)}_{ab,\alpha} \approx \rho_\alpha \chi_{c^*} (\epsilon^{(l+1)}_{ab,\alpha} + \frac{\chi_{c,2}}{\chi_{c^*}} p^*_{ab} \epsilon^{(l)}_{ab,\alpha} + \delta^{(l)}_{ab,\alpha}) \tag{156}$$

which gives

$$\epsilon^{(l)}_{ab,\alpha} \approx (\rho_\alpha \chi_{c^*})^l \epsilon^{(0)}_{ab,\alpha} \ , \tag{157}$$

$$\delta^{(l)}_{ab,\alpha} \approx (\rho_\alpha \chi_{c^*})^l \left[\delta^{(0)}_{ab,\alpha} + l \left(1 + \frac{\chi_{c,2}}{\chi_{c^*}} p^*_{ab} \right) \epsilon^{(0)}_{ab,\alpha} \right] \tag{158}$$

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

E.2. The effect of pooling and flattening of CNNs

With the bulk of the theory in hand, we now turn our attention to CNN-F and CNN-P. We have shown 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 $\mathcal{K}^{(l)}_{\alpha,\alpha'}(x, x')$ and $\Theta^{(l)}_{\alpha,\alpha'}(x, x')$, 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 logits layer. The former strategy “flattens” an image of size (d, N) into a vector in $\mathbb{R}^{dN}$ 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 the trace and the mean of the pixel-to-pixel covariance on the NNGP/NTK, respectively, i.e.,

$$\Theta^{(l)}_{\text{flatten}}(x, x') = \frac{1}{d} \sum_{\alpha \in [d]} \Theta^{(l)}_{\alpha,\alpha}(x, x') \,, \tag{159}$$

$$\Theta^{(l)}_{\text{pool}}(x, x') = \frac{1}{d^2} \sum_{\alpha,\alpha' \in [d]} \Theta^{(l)}_{\alpha,\alpha'}(x, x') \,, \tag{160}$$

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. $\mathcal{K}^{(l)}_{\text{flatten}}$, $\mathcal{K}^{(l)}_{\text{pool}}$ and $\mathcal{K}^{(l)}_{\text{fc}}$ 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 FCN are of the same order.) Applying flattening and pooling, the NTKs become

$$\Theta^{(l)}_{\text{flatten}}(x,x') = \frac{1}{d}\sum_{\alpha}\Theta^{(l)}_{\alpha,\alpha}(x,x') = \mathbf{1}_{x=x'}p^{(l)} + \mathbf{1}_{x\neq x'}p^{(l)}_{ab}\,, \tag{161}$$

$$\Theta^{(l)}_{\text{pool}}(x,x') = \frac{1}{d^2}\sum_{\alpha,\alpha'}\Theta^{(l)}_{\alpha,\alpha'}(x,x') = \frac{1}{d}\mathbf{1}_{x=x'}(p^{(l)} - p^{(l)}_{ab}) + p^{(l)}_{ab}\,, \tag{162}$$

respectively. As we can see, $\Theta^{(l)}_{\text{flatten}}$ is essentially the same as its FCN 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 $\Theta^{(l)}_{\text{flatten}}$ and $\Theta^{(l)}_{\text{fc}}$ are essentially the same for large l; see Figure 1 (a - c).

However, pooling alters the NTK/NNGP spectrum in an interesting way. Noticeably, the contribution from $p^{(l)}$ is discounted by a factor of d. On the critical line, asymptotically, the on- and off-diagonal terms are

$$\Theta^{(l)}_{\text{pool}}(x,x) = \frac{2+d}{3d}lq^* + \mathcal{O}(1) \tag{163}$$

$$\Theta^{(l)}_{\text{pool}}(x,x') = \frac{1}{3}lq^* + \mathcal{O}(1) \tag{164}$$

This implies

$$\lambda^{(l)}_{\text{max}} = (md+2)q^*l/(3d) + \mathcal{O}(1) \tag{165}$$

$$\lambda^{(l)}_{\text{bulk}} = 2q^*l/(3d) + \mathcal{O}(1) \tag{166}$$

$$\kappa^{(l)} = \frac{md+2}{2} + md\mathcal{O}(l^{-1}) \tag{167}$$

Here we use blue color to indicate the changes of such quantities against their $\Theta^{(l)}_{\text{flatten}}$ counterpart. Alternatively, one can consider $\Theta^{(l)}_{\text{flatten}}$ as a special version (with $d=1$) of $\Theta^{(l)}_{\text{pool}}$. Thus pooling decreases $\lambda^{(l)}_{\text{bulk}}$ 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, i.e. $p^{(l)} = \mathcal{O}(\chi_1^l/d)$. This improves $P(\Theta^{(l)})$ by a factor of 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 $\lambda^{(l)}_{\text{max}}$ much but does decrease $\lambda^{(l)}_{\text{bulk}}$ by a factor of d and the condition number $\kappa^{(l)}$ grows approximately like $dl\chi_1^{-l}$, d times bigger than its flattening and fully-connected network counterparts. 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. These statements are supported empirically in Figure 2.

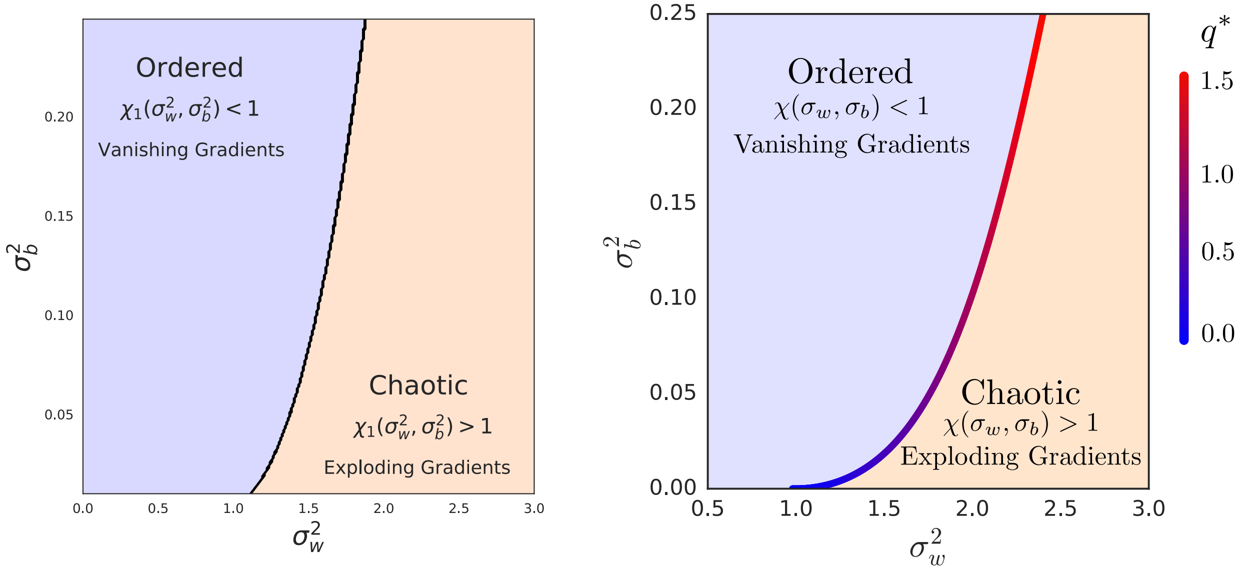
F. Figure Zoo

F.1. Phase Diagrams: Figure 5.

We plot the phase diagrams for the Erf function and the tanh function (adopted from (Pennington et al., 2018)).

F.2. SGD on FCN on Larger Dataset: Figure 6.

We report the training and test accuracy of FCN trained on a subset (16k training points) of CIFAR-10 using SGD with 20 × 20 different (σ_ω^2, l) configurations.



(a) Phase Diagram for Erf.

(b) Phase Diagram for $\tanh$.

Figure 5. Phase Diagram for $\tanh$ and Erf (right).

F.3. NNGP vs NTK prediction: Figure 7.

Here we compare the test performance of the NNGP and NTK with different (σ_ω^2, l) configurations. In the chaotic phase, the generalizable depth-scale of the NNGP is captured by $\xi_{c^*} = -1/\log(\chi_{c^*})$. In contrast, the generalizble depth-scale of the NTK is captured by $\xi_* = -1/(\log(\chi_{c^*}) - \log(\chi_1))$. Since $\chi_1 > 1$ in the chaotic phase, $\xi_{c^*} > \xi_*$. Thus for larger depth, the NNGP kernel performs better than the NTK. Corrections due to an additional average pooling layer is plotted in the third column of Figure .7

F.4. Densely Sweeping Over σ_b^2: Figure 8

We demonstrate that our prediction for the generalizable depth-scales for the NTK (ξ_*) and NNGP (ξ_c) are robust across a variety of hyperparameters. We densely sweep over 9 different values of $\sigma_b^2 \in [0.2, 1.8]$. For each σ_b^2 we compute the NTK/NNGP test accuracy for 20 * 50 different configurations of (l, σ_ω^2) with $l \in [1, 100]$ and $\sigma_\omega^2 \in [0.1^2, 4.9^2]$. The training set is a 8k subset of CIFAR-10.

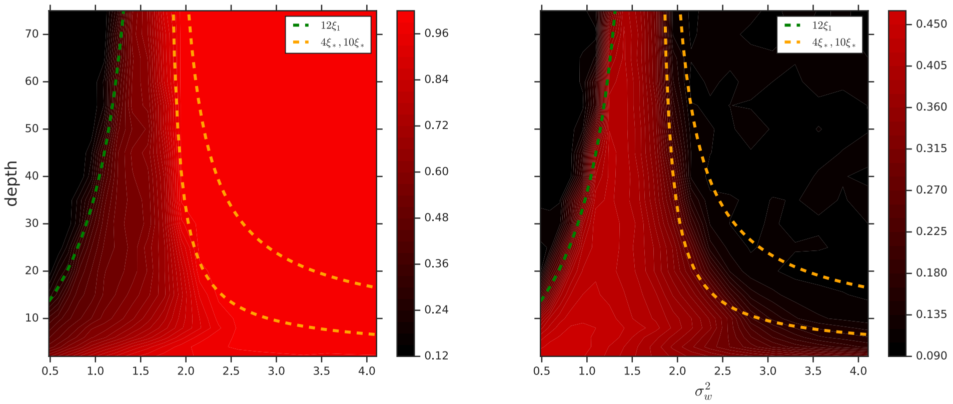


Figure 6. Training and Test Accuracy for FCN for different (σ_ω^2, l) configurations.

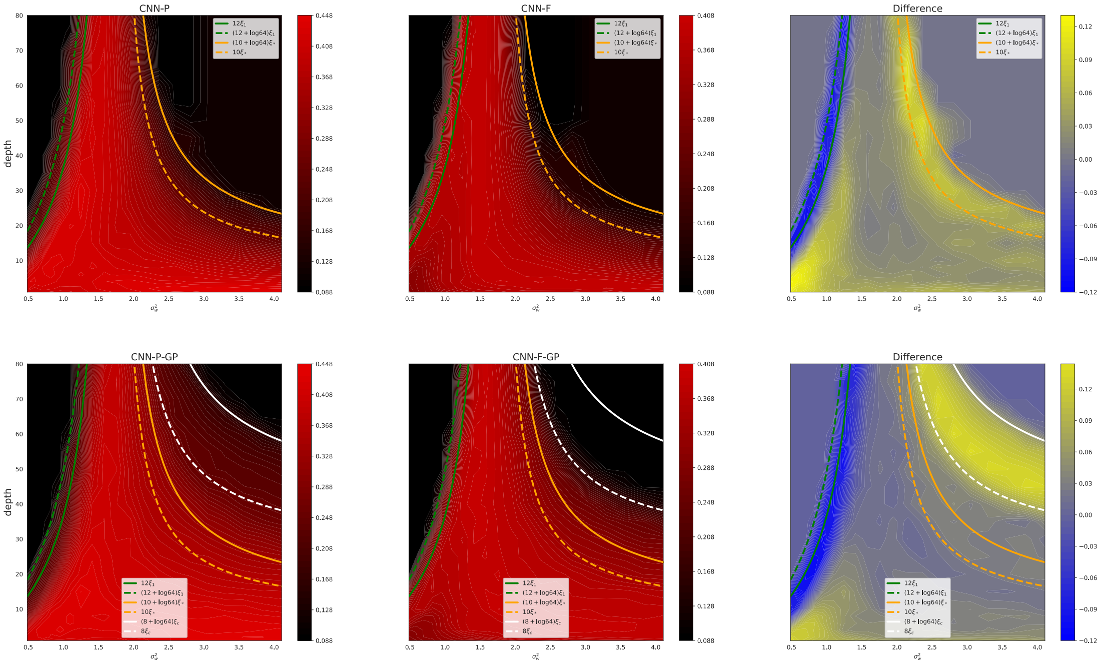


Figure 7. Test Accuracy for NTK (top) and NNGP prediction for different (σ_ω^2, l) configurations. First/second column: CNN with/without pooling. Last column: difference between the first and second columns.

F.5. Densely Sweeping Over the Regularization Strength σ: Figure 9

Similar to the above setup, we fixed $\sigma_b^2 = 1.6$ and densely vary $\sigma \in \{0, 10^{-6}, \ldots, 10^0\}$.

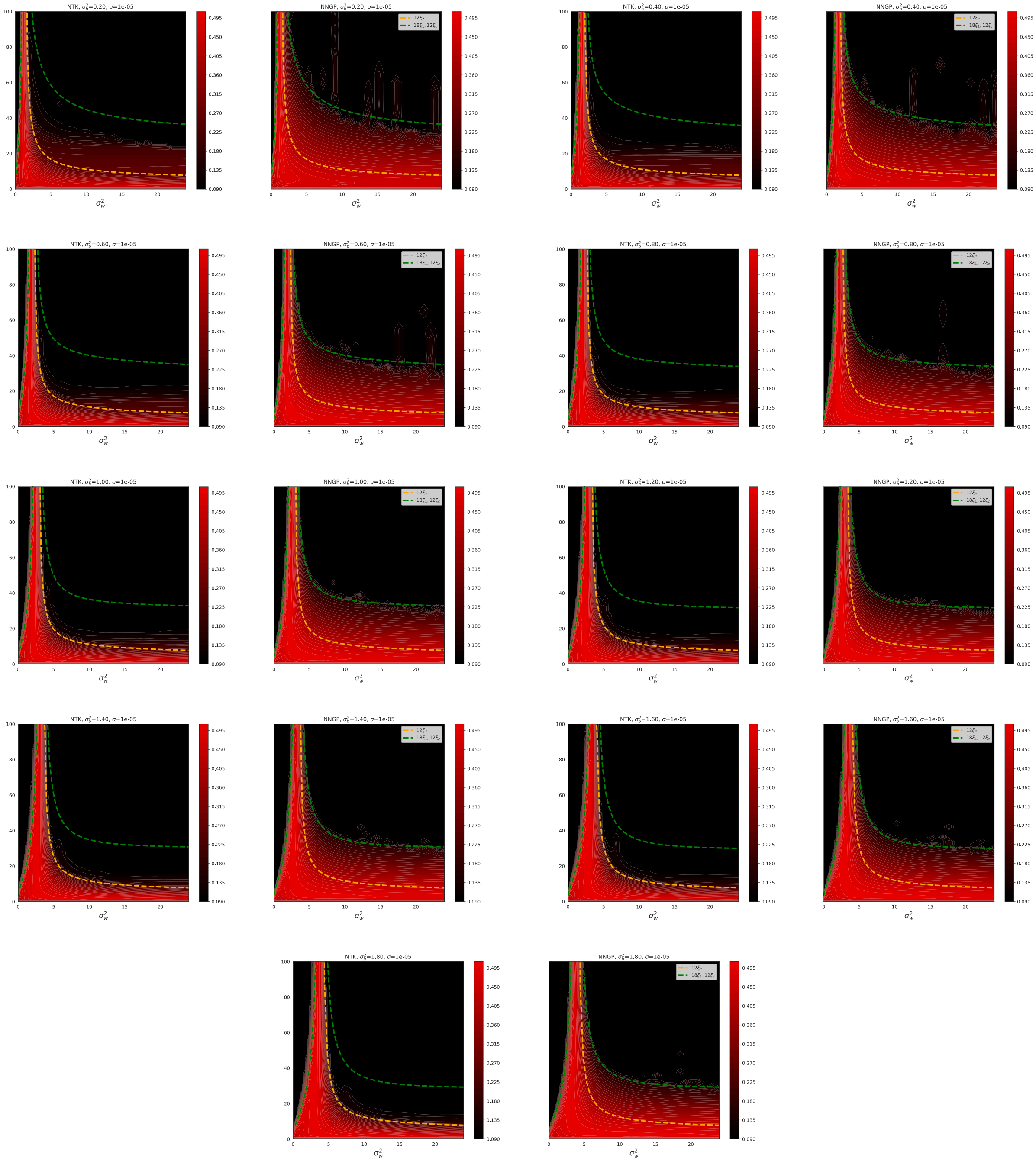


Figure 8. **Generalization metrics for NTK/NNGP vs Test Accuracy vs σ_b^2.**

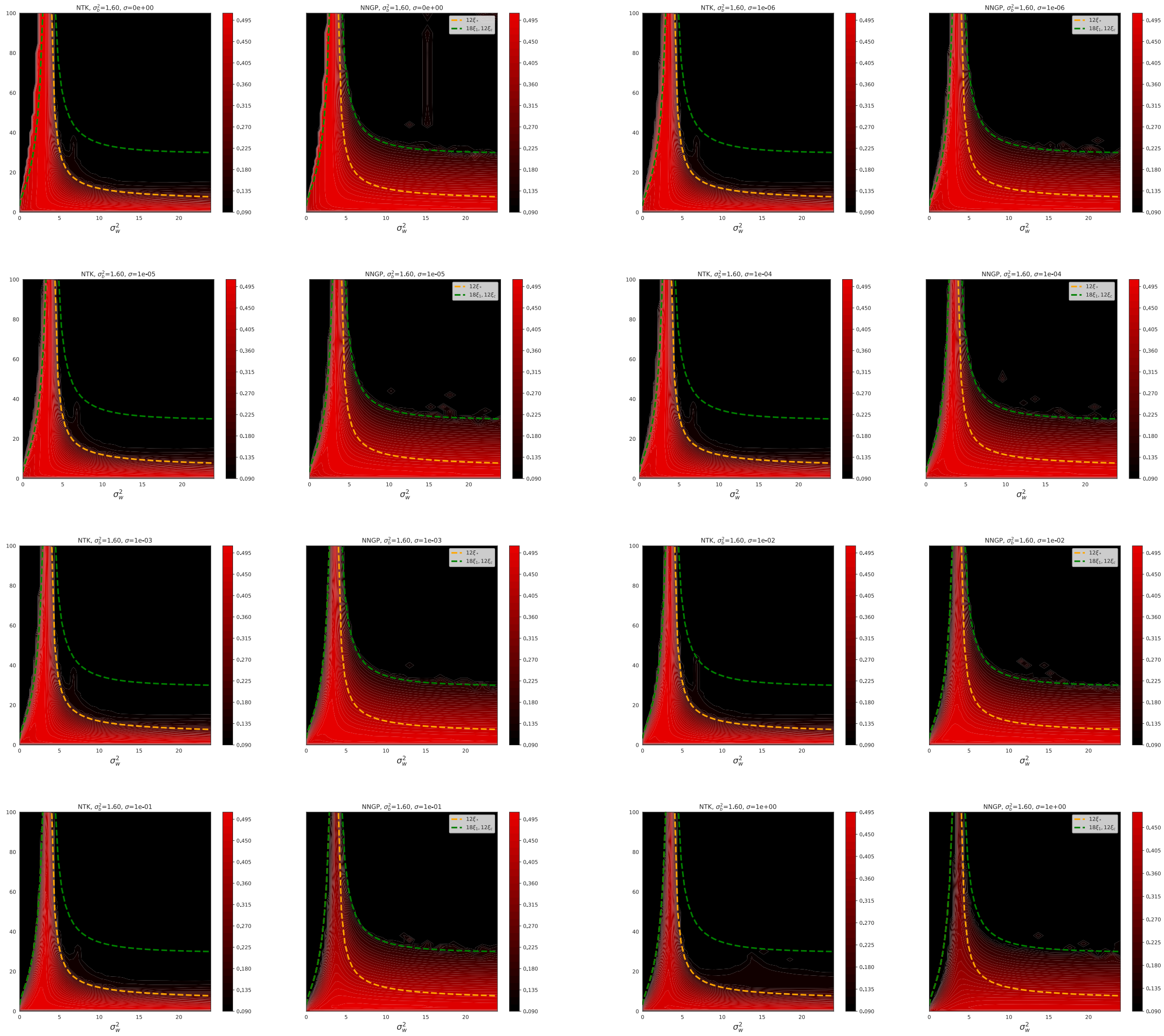


Figure 9. **Generalization metrics for NTK/NNGP vs Test Accuracy vs σ.**