A DETAILS OF EXPERIMENTAL SETUP

In this section, we present the details on the experimental setup used for the plots depicted in the main body of the paper. As mentioned, the exact width for FCNs have been reported. For WideResNet-16-k we use two block layers, and the initial convolution in the network has a width of 16WF where WF is the reported WF. For instance, $\text{WF} = 16$ means that the first block layer has a width of 256 and the second block layer has a width of 512. For ResNet18, we also used the same approach, multiplying WF by 16. Thus, when $\text{WF} = 4$, the constructed network will have the exact architecture as the classical ResNet18 architecture reported. A WF of 16 means a ResNet18 with each layer being 4 times wider than the original width.

When training the neural networks using SGD, a constant batch size of 128 was used across all different networks and different dataset sizes used for training. The learning rate for all networks was also fixed to 0.1. However, not all networks were trainable with this fixed learning rate, as the gradients would sometimes blow up and give `NaN` training loss, typically for the largest width of each mentioned architecture. In those cases, we decreased the learning rate to 0.01 to train the networks.

Note that to be consistent with the literature on NTKs, techniques like data augmentation have been turned off, but a weight decay of 0.0001 along with a momentum of 0.9 for SGD is used. Data augmentation here plays an important role in the attained test accuracies of the fully trained networks.

B FURTHER RESULTS ON THE APPROXIMATION QUALITY

In this section we'll lay out the proofs of the theorems provided in the main text, mainly Theorems 1, 4 and 5. Towards this, we first define some notation and show a simple recursive formula for computing the tangent kernel that we take advantage of to prove the theorems. Consider a NN $f : \mathbb{R}^D \to \mathbb{R}^O$. We assume the final read-out layer of the NN f is a dense layer with width w. Assuming the NN f has L layers, we define θ_l to be the corresponding parameters of layer $l \in \{1, 2, \dots, L\}$. Furthermore, let's define $g : \mathbb{R}^O \to \mathbb{R}^w$ as the output of the immediate last layer of the NN f, such that $f(x) = \theta_L g(x)$ for some $\theta_L \in \mathbb{R}^{O \times w}$.

As shown by Lee et al. (2019); Yang (2020), the NTK can be reformulated as the layer-wise sum of gradients (when the parameters of each layer θ_l are assumed to be vectorized) of the output with respect to θ_l. Accordingly, we denote eNTK of a NN f as

$$\Theta_f(x_1, x_2) = \sum_{l=1}^{L} \nabla_{\theta_l} f(x_1) \nabla_{\theta_l} f(x_2)^\top. \tag{8}$$

Now, noting that as the final layer of f is a dense layer, we can use the chain rule to write $\nabla_{\theta_l} f(x)$ as $\frac{\partial f}{\partial g(x)} \frac{\partial g(x)}{\partial \theta_l}$ where $\frac{\partial f(x)}{\partial g(x)} = \theta_L$. Thus, we can rewrite (8) as

$$\begin{aligned}
\Theta_f(x_1, x_2) &= \sum_{l=1}^{L-1} \theta_L \nabla_{\theta_l} g(x_1) \, \nabla_{\theta_l} g(x_2)^\top {\theta_L}^\top + \nabla_{\theta_L} f(x_1) \nabla_{\theta_L} f(x_2)^\top \\
&= \theta_L \left(\sum_{l=1}^{L} \nabla_{\theta_l} g(x_1) \, \nabla_{\theta_l} g(x_2)^\top \right) {\theta_L}^\top + g(x_1)^\top g(x_2) \, I_O \\
&= \theta_L \, \Theta_g(x_1, x_2) \, {\theta_L}^\top + g(x_1)^\top g(x_2) \, I_O.
\end{aligned} \tag{9}$$

Applying Equation (9), we can already see that the pNTK of a network f simply calculates a weighted summation of all elements of eNTK into a scalar, since it can be seen as adding a new final dense layer to the network f with the fixed weight vector $\frac{1}{\sqrt{O}} \mathbf{1}_O$ where $\mathbf{1}_O$ is the O-dimensional vector consisting of all 1s.

Before moving on with the approximation proofs, we would like to mention that the proofs in this section rely heavily on concentration inequalities of *sub-exponential* random variables. Thus, we start by providing some background about sub-exponential random variables and the related concentration inequalities that we will use later on.

B.1 BACKGROUND ON SUB-EXPONENTIAL RANDOM VARIABLES

A random variable X with mean μ is called *sub-exponential* (Wainwright, 2019) if there are non-negative parameters (ν, α) such that

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\nu^2\lambda^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{\alpha}.$$

We denote $X \sim SE(\nu, \alpha)$ to show that X is a sub-exponential random variable with parameters (ν, α).

A famous sub-exponential random variable is the product of two standard normal distributions such that the two factors are independent ($X_1 = |z_1||z_2| \sim SE(\nu_p, \alpha_p)$ with mean $2/\pi$) or the same ($X_2 = z^2 \sim SE(2, 4)$ with mean 1.) where $z \sim \mathcal{N}(0, 1)$. We now present a few lemmas regarding sub-exponential random variables that will come in handy in the later subsections of the appendix.

Lemma 6. *If a random variable X is sub-exponential with parameters (ν, α), then the random variable sX where $s \in \mathbb{R}^+$ is also sub-exponential with parameters $(s\nu, s\alpha)$.*

Proof. Consider $X \sim SE(\nu, \alpha)$ and $X' = sX$ with $\mathbb{E}[X'] = s\,\mathbb{E}[X]$, then according to the definition of a sub-exponential random variable

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda(X-\mu)\right)\right] \leq \exp(\frac{\nu^2\lambda^2}{2}) \quad \text{for all } |\lambda| < \frac{1}{\alpha} \\
&\implies \mathbb{E}\left[\exp\left(\frac{\lambda}{s}(sX - s\mu)\right)\right] \leq \exp(\frac{\nu^2 s^2 \frac{\lambda^2}{s^2} 2}{2}) \quad \text{for all } |\frac{\lambda}{s}| < \frac{1}{s\alpha} \\
&\xrightarrow{\lambda' = \frac{\lambda}{s}} \mathbb{E}\left[\exp\left(\lambda'(X' - \mu')\right)\right] \leq \exp(\frac{\nu^2 s^2 \lambda'^2}{2}) \quad \text{for all } |\lambda'| < \frac{1}{s\alpha}
\end{aligned} \tag{10}$$

Defining $\alpha' = s\alpha$ and $\nu' = s\nu$ we recover that $X' \sim SE(s\nu, s\alpha)$.

□

Proposition 7. *If the random variables X_i for $i \in [1 - N]$ for $N \in \mathbb{N}^+$ are all sub-exponential with parameters (ν_i, α_i) and independent, then $\sum_{i=1}^N X_i$ is sub-exponential with parameters $(\sqrt{\sum_{i=1}^N \nu_i^2}, \max_i \alpha_i)$.*

Proof. The proof is a simplification of the discussion prior to equation 2.18 in Wainwright (2019).

Proposition 8. *For a random variable $X \sim SE(\nu, \alpha)$, the following concentration inequality holds:*

$$\Pr\left(|X - \mu| \geq t\right) \leq 2\exp\left(-\min\left(\frac{t^2}{2\nu^2}, \frac{t}{2\alpha}\right)\right)$$

Proof. The proof directly follows from applying a scalar multiplication to the result derived in Equation 2.18 in Wainwright (2019).

Corollary 9. *For a random variable $X \sim SE(\nu, \alpha)$ the following inequality holds with probability at least $1 - \delta$:*

$$|X - \mu| < \max\left(\nu\sqrt{2\log\frac{2}{\delta}}, 2\alpha\log\frac{2}{\delta}\right).$$

B.2 PSEUDO-NTK RELATIVELY CONVERGES TO ENTK AS WIDTH GROWS

Let's denote a neural network with L dense hidden layers whose width is n as:

$$\begin{aligned}
f^0(x) &= x \\
f^{l+1}(x) &= \phi(W^{(l+1)} f^l(x)) \\
f(x) &= f^L(x) = W^{(L)} f^{L-1}(x)
\end{aligned} \tag{11}$$

such that ϕ is a differentiable coordinate-wise activation function.

Setting A (ReLU-MLP)**.** We assume the following assumptions hold in our setting:

- We assume W^l for $l \in 1, \ldots, L$ is initialized according to the *NTK parameterization*, meaning that each scalar parameter is distributed according to $\mathcal{N}(0, 1/n)$.
- We assume the width of all hidden layers are identical (and equal to n). The proof extends naturally to the case of non-equal widths as long as $n'/n \in (0, \infty)$ for each consecutive pair of layers.
- We assume ϕ is a ReLU-like (GeLU, PReLU, or any other similar function) activation function that is at most 1-Lipschitz.
- We assume the training data $\mathcal{X}$ is finite and contained in a compact set and there are no overlapping datapoints.

Note that we can recursively define the eNTK of f^{l+1} using the eNTK of f^l as

$$\begin{aligned}
\Theta^{(l+1)}(x_1, x_2) &= \sum_{i=1}^{l} \frac{\partial f^{l+1}(x_1)}{\partial W^i} \frac{\partial f^{l+1}(x_2)}{\partial W^i}^\top + \overbrace{\frac{\partial f^{l+1}(x_1)}{\partial W^{l+1}} \frac{\partial f^{l+1}(x_2)}{\partial W^{l+1}}^\top}^{K_D^{l+1}(x_1,x_2)} \\
&= \sum_{i=1}^{l} \frac{\partial \phi(W^{(l+1)} f^l(x_1))}{\partial W^i} \frac{\partial \phi(W^{(l+1)} f^l(x_2))}{\partial W^i}^\top + K_D^{l+1}(x_1, x_2) \\
&= \sum_{i=1}^{l} \frac{\partial \phi(W^{(l+1)} f^l(x_1))}{\partial f^l(x_1)} \frac{\partial f^l(x_1)}{\partial W^i} \frac{\partial f^l(x_2)}{\partial W^i}^\top \frac{\partial \phi(W^{(l+1)} f^l(x_2))}{\partial f^l(x_2)}^\top + K_D^{l+1}(x_1, x_2) \\
&= \frac{\partial \phi(W^{(l+1)} f^l(x_1))}{\partial f^l(x_1)} \left[\sum_{i=1}^{l} \frac{\partial f^l(x_1)}{\partial W^i} \frac{\partial f^l(x_2)}{\partial W^i}^\top \right] \frac{\partial \phi(W^{(l+1)} f^l(x_2))}{\partial f^l(x_2)}^\top + K_D^{l+1}(x_1, x_2) \\
&= \frac{\partial \phi(W^{(l+1)} f^l(x_1))}{\partial f^l(x_1)} \Theta^{(l)}(x_1, x_2) \frac{\partial \phi(W^{(l+1)} f^l(x_2))}{\partial f^l(x_2)}^\top + K_D^{l+1}(x_1, x_2)
\end{aligned} \tag{12}$$

where

$$\frac{\partial \phi(W^{(l+1)} f^l(x))}{\partial f^l(x)} = W^{(l+1)} \odot \left[\dot{\phi}(W^{(l+1)} f^l(x)) \right]_{1 \times n} \tag{13}$$

and $K_D^{l+1}(x_1, x_2) = f^l(x_1)^\top f^l(x_2) I_n$ is a diagonal matrix. We can think of the last layer as following the same equations with ϕ the identity function, so that $\phi'(x) = 1$. Furthermore, note that using the same approach we can show that pNTK of the layer l can be derived as

$$\hat{\Theta}^{(l+1)}(x_1, x_2) = \frac{\mathbf{1}_n}{\sqrt{n}} \Theta^{(l+1)}(x_1, x_2) \frac{\mathbf{1}_n^\top}{\sqrt{n}} \tag{14}$$

where $\mathbf{1}_n$ is the vector of 1s with size n.

We now sketch the proof idea first and then move onto rigorously proving each part of the sketch. First, note that using Equation (12) we can recursively calculate the eNTK of a general MLP. We take advantage of this recursive definition and derive bounds for the magnitude of the elements of the eNTK on a layer-to-layer basis recursively. To do so, we first show that the eNTK of the first layer of the NN, $\Theta^{(1)}(x_1, x_2)$, is in general a diagonal matrix. Then, we present a series of lemmas that bound the elements of the eNTK of layer $l + 1$ based on the magnitude (bounds) of the eNTK of layer l. Finally, based on the derived bounds on the magnitude of elements of the eNTK of a NN with l layers and Equation (14), we prove that the Frobenius norm of the pNTK relatively converges to the Frobenius norm of the corresponding eNTK with high probability over random initialization.

Before moving on, it's useful to first show a simple inequality on the elements of a tangent kernel based on the Lipschitz-ness of the activation function; this will help us further in deriving the

aforementioned bounds. Define $V^{(l)}(x) = W^{(l)} \odot \left[\dot{\phi}(W^{(l)} f^{l-1}(x))\right]_{1\times n}$. We can write each entry of $\Theta^{(l+1)}(x_1, x_2)$ as

$$\Theta^{(l+1)}(x_1,x_2)_{ij} = \sum_{a=1}^{n}\sum_{b=1}^{n} V^{(l+1)}(x_1)_{ia} V^{(l+1)}(x_2)_{jb} \Theta^{(l)}(x_1,x_2)_{ab} + f^l(x_1)^\top f^l(x_2)\mathcal{I}(i=j)$$
$$\leq_{\text{in abs}} \sum_{a=1}^{n}\sum_{b=1}^{n} |{W^{(l)}}_{ia}||{W^{(l)}}_{jb}||\Theta^{(l)}(x_1,x_2)_{ab}| + f^l(x_1)^\top f^l(x_2)\mathcal{I}(i=j) \quad (15)$$

where $\mathcal{I}$ denotes the 0-1 indicator function, and the inequality follows from the activation function ϕ being 1-Lipschitz.

Lemma 10 (Diagonality of the first layer's tangent kernel)**.** *For a NN under Setting A the corresponding eNTK of the first layer $\Theta^{(1)}(x_1, x_2)$ is diagonal. Moreover, for any $\delta > 0$, there is a corresponding constant $C^{(1)} > 0$ such that for each diagonal element $\Theta^{(1)}(x_1, x_2)_{ii}$ we have that*

$$|\Theta^{(1)}(x_1,x_2)_{ii}| < C^{(1)}$$

with probability at least $1 - \delta$.

Proof. Consider the one layer NN $f^1(x) = \phi(W^{(1)}x)$. For this case, we have:

$$\Theta^{(1)}(x_1,x_2)_{ij} = \begin{cases} \sum_{a=1}^{D} x_{1a}\dot{\phi}(W_i x_1) x_{2a}\dot{\phi}(W_i x_2) & \text{if } i=j \\ 0 & \text{if } i\neq j \end{cases} \quad (16)$$

and thus, based on the fact that the activation function ϕ is 1-Lipschitz we can conclude that

$$|\Theta^{(1)}(x_1,x_2)_{ij}| \leq \begin{cases} |x_1^\top x_2| & \text{if } i=j \\ 0 & \text{if } i\neq j \end{cases}. \quad (17)$$

Thus, the tangent kernel of the first layer is a diagonal matrix whose entries are independent of the width of the first layer (n), and can be bounded by a positive constant with high probability. □

Next, we present a series of lemmas that will help us derive the bounds on the elements of the tangent kernel of layer $l + 1$ based on the bounds of the tangent kernel of layer l.

Lemma 11. *Consider a NN under Setting A with depth $\geq l + 1$. Assume there is a constant $C^{(l)} > 0$ such that $|\Theta^{(l)}(x_1, x_2)_{ii}| < C^{(l)}$ with probability at least $1 - \delta_{in}$ and every non-diagonal element of $\Theta^{(l)}(x_1, x_2)$ is zero. Then for any small $\delta > 0$ there are corresponding constants $C_1^{(l+1)}, C_2^{(l+1)}, n^{l+1} > 0$ such that for any $n > n^{l+1}$*

$$|\Theta^{(l+1)}(x_1,x_2)_{ij}| \leq \begin{cases} C_1^{(l+1)} n & \text{if } i=j \\ C_2^{(l+1)}/\sqrt{n} & \text{if } i\neq j \end{cases} \quad (18)$$

with probability at least $1 - \delta$.

Proof. Based on Equation (15) we can expand the elements of $\Theta^{(l+1)}(x_1, x_2)$ as

$$|\Theta^{(l+1)}(x_1,x_2)_{ij}| \leq \begin{cases} \sum_{a=1}^{n}\sum_{b=1}^{n} |{W^{(l+1)}}_{ia}||{W^{(l+1)}}_{ib}||\Theta^{(l)}(x_1,x_2)_{ab}| + f^l(x_1)^\top f^l(x_2) & \text{if } i=j \\ \sum_{a=1}^{n}\sum_{b=1}^{n} |{W^{(2)}}_{ia}||{W^{(l+1)}}_{jb}||\Theta^{(l)}(x_1,x_2)_{ab}| & \text{if } i\neq j \end{cases}$$
$$= \begin{cases} \sum_{a=1}^{n} |{W^{(l+1)}}_{ia}||{W^{(l+1)}}_{ia}||\Theta^{(l)}(x_1,x_2)_{aa}| + f^l(x_1)^\top f^l(x_2) & \text{if } i=j \\ \sum_{a=1}^{n} |{W^{(l+1)}}_{ia}||{W^{(l+1)}}_{ja}||\Theta^{(l)}(x_1,x_2)_{aa}| & \text{if } i\neq j \end{cases} \quad (19)$$

Using the bound provided for the elements of $\Theta^{(l)}(x_1, x_2)$, replacing each element of the weight matrix as $\frac{1}{n}z$ where $z \sim \mathcal{N}(0,1)$ and and applying Lemma 18 we can further show that for the diagonal elements of $\Theta^{(l+1)}(x_1, x_2)$ we have that

$$\begin{aligned}|\Theta^{(l+1)}(x_1,x_2)_{ii}| \leq \frac{C^{(l)}}{n}\sum_{a=1}^{n} z_{ia}^2 + G^{(l)} \sim \frac{C^{(l)}}{n}\sum_{a=1}^{n} SE(2,4) + G^{(l)}n \\ = \frac{C^{(l)}}{\sqrt{n}} SE\left(2, \frac{4\sqrt{n}}{n}\right) + G^{(l)}n\end{aligned} \tag{20}$$

with probability at least $(1-\delta_{in})(1-\delta_g)$. Note that in the above equation we have utilized the fact that the product of the same standard normal distribution with itself is a sub-exponential random variable with $SE(2,4)$ parameters in conjunction with Proposition 7. Applying Corollary 9 we can see that

$$|\Theta^{(l+1)}(x_1,x_2)_{ii}| \leq \frac{C^{(l)}}{\sqrt{n}}\left(\sqrt{4\log\frac{2}{\delta}} + 1\right) + G^{(l)}n \tag{21}$$

with probability at least $(1-\delta)(1-\delta_g)$ where δ depends on δ_{in}. As we see, the right hand side term is dominating this inequality and thus, we can claim that there is a $n^{l+1} > 0$ such that for all $n > n^{l+1}$; $|\Theta^{(l+1)}(x_1,x_2)_{ii}| < G^{(l)}n$ as desired.

For the non-diagonal case we can show that

$$\begin{aligned}|\Theta^{(l+1)}(x_1,x_2)_{ij}| \leq \frac{C^{(l)}}{n}\sum_{a=1}^{n}|z_{ia}||z_{ja}| \sim \frac{C^{(l)}}{n}\sum_{a=1}^{n} SE\left(\nu_p, \alpha_p\right) \\ = \frac{C^{(l)}}{\sqrt{n}} SE\left(\nu_p, \frac{\alpha_p\sqrt{n}}{n}\right)\end{aligned} \tag{22}$$

with probability at least $1-\delta_{in}$. Again, applying Corollary 9 shows us that

$$|\Theta^{(l+1)}(x_1,x_2)_{ij}| \leq \frac{C^{(l)}}{\sqrt{n}}\left(\sqrt{2\nu_p\log\frac{2}{\delta}} + \frac{2}{\pi}\right) \tag{23}$$

with probability at least $(1-\delta)(1-\delta_{in})$ as desired. □

Lemma 12. *Consider a NN under Setting A with depth $\geq l+1$. Assume there are constants $C_1^{(l)}, C_2^{(l)} > 0$ such that $|\Theta^{(l)}(x_1,x_2)_{ii}| < C_1^{(l)}n$ and $|\Theta^{(l)}(x_1,x_2)_{ij}| < C_2^{(l)}/\sqrt{n}$ with probability at least $1-\delta_{in}$. Then for any arbitrary small $\delta > 0$ there are constants $C_1^{(l+1)}, C_2^{(l+1)}, n^{l+1} > 0$ such that for any $n > n^{l+1}$*

$$|\Theta^{(l+1)}(x_1,x_2)_{ij}| \leq \begin{cases} C_1^{(l+1)}n & \text{if } i = j \\ C_2^{(l+1)}\sqrt{n} & \text{if } i \neq j \end{cases} \tag{24}$$

with probability at least $1-\delta$.

Proof. Using the same expansion that we utilized in the proof for the previous lemma, for the diagonal elements of $\Theta^{(l+1)}(x_1, x_2)$ we have:

$$\begin{aligned}|\Theta^{(l+1)}(x_1,x_2)_{ii}| &\leq \sum_{a=1}^{n}\sum_{b=1}^{n}|W^{(l+1)}{}_{ia}||W^{(l+1)}{}_{ib}||\Theta^{(l)}(x_1,x_2)_{ab}| + G^{(l)}n \\ &= \underbrace{\frac{1}{n}\sum_{a=1}^{n} z_{ia}^2|\Theta^{(l)}(x_1,x_2)_{aa}|}_{\Theta_1^{(l+1)}(x_1,x_2)_{ii}} + \underbrace{\frac{1}{n}\sum_{a=1}^{n}\sum_{b=1,b\neq a}^{n}|z_{ia}||z_{ib}||\Theta^{(l)}(x_1,x_2)_{ab}|}_{\Theta_2^{(l+1)}(x_1,x_2)_{ii}} + G^{(l)}n\end{aligned} \tag{25}$$

where

$$|\Theta_1^{(l+1)}(x_1,x_2)_{ii}| \leq \frac{C_1^{(l)} n}{n} \sum_{a=1}^{n} z_{ia}^2 \sim C_1^{(l)} \sqrt{n} SE\left(2, \frac{4\sqrt{n}}{n}\right) \tag{26}$$

and

$$\begin{aligned} |\Theta_2^{(l+1)}(x_1,x_2)_{ii}| &\leq \frac{C_2^{(l)}}{n\sqrt{n}} \sum_{a=1}^{n} \sum_{b=1, b\neq a}^{n} |z_{ia}||z_{ib}| \\ &\sim \frac{C_2^{(l)}}{n\sqrt{n}} \sum_{a=1}^{n^2} SE\left(\nu_p, \alpha_p\right) = \frac{C_2^{(l)}}{\sqrt{n}} SE\left(\nu_p, \frac{\alpha_p}{n}\right) \end{aligned} \tag{27}$$

each with probability at least $1-\delta_{in}$. As shown before, both of these terms are dominated by the $G^{(l)}n$ term in the inequality for diagonal elements and thus, we can claim that there is a $n_1^{l+1} > 0$ such that for all $n > n_1^{l+1}$; $|\Theta^{(l+1)}(x_1,x_2)_{ii}| < G^{(l)}n$ as desired.

Next, for the non-diagonal elements of $\Theta^{(l+1)}(x_1,x_2)$ we have:

$$\begin{aligned} |\Theta^{(l+1)}(x_1,x_2)_{ij}| &= \sum_{a=1}^{n}\sum_{b=1}^{n} |W^{(l+1)}{}_{ia}||W^{(l+1)}{}_{jb}||\Theta^{(l)}(x_1,x_2)_{ab}| \\ &= \underbrace{\frac{1}{n}\sum_{a=1}^{n}|z_{ia}||z_{ja}||\Theta^{(l)}(x_1,x_2)_{aa}|}_{\Theta_1^{(l+1)}(x_1,x_2)_{ij}} + \underbrace{\frac{1}{n}\sum_{a=1}^{n}\sum_{b=1,b\neq a}^{n}|z_{ia}||z_{jb}||\Theta^{(l)}(x_1,x_2)_{ab}|}_{\Theta_2^{(l+1)}(x_1,x_2)_{ij}} \end{aligned} \tag{28}$$

where

$$|\Theta_1^{(l+1)}(x_1,x_2)_{ij}| \leq \frac{C_1^{(l)} n}{n} \sum_{a=1}^{n} |z_{ia}||z_{ja}| \sim C_1^{(l)} \sqrt{n} SE\left(\nu_p, \frac{\alpha\sqrt{n}}{n}\right) \tag{29}$$

and

$$\begin{aligned} |\Theta_2^{(l+1)}(x_1,x_2)_{ij}| &\leq \frac{C_2^{(l)}}{n\sqrt{n}} \sum_{a=1}^{n} \sum_{b=1, b\neq a}^{n} |z_{ia}||z_{jb}| \\ &\sim \frac{C_2^{(l)}}{n\sqrt{n}} \sum_{a=1}^{n^2} SE\left(\nu_p, \alpha_p\right) = \frac{C_2^{(l)}}{\sqrt{n}} SE\left(\nu_p, \frac{\alpha_p}{n}\right) \end{aligned} \tag{30}$$

each with probability at least $1-\delta_{in}$. As $|\Theta_2^{(l+1)}(x_1,x_2)_{ij}|$ is dominated by $|\Theta_1^{(l+1)}(x_1,x_2)_{ij}|$ we can claim that according to Corollary 9, there exists $n_2^{l+1} > 0$ such that for $n > n_2^{l+1}$ we have that

$$|\Theta^{(l+1)}(x_1,x_2)_{ij}| \leq C^{(l)}\sqrt{n}\left(\sqrt{2\nu_p \log\frac{2}{\delta}} + \frac{2}{\pi}\right) \tag{31}$$

with probability at least $(1-\delta)(1-\delta_{in})$. Thus, the lemma's claim holds with probability at least $(1-\delta)(1-\delta_{in})(1-\delta_g)$ and $n > \max(n_1^{l+1}, n_2^{l+1})$ as desired. □

Lemma 13. *Consider a NN under Setting A with depth $\geq l+1$. Assume there are constants $C_1^{(l)}, C_2^{(l)} > 0$ such that $|\Theta^{(l)}(x_1,x_2)_{ii}| < C_1^{(l)} n$ and $|\Theta^{(l)}(x_1,x_2)_{ij}| < C_2^{(l)}\sqrt{n}$ with probability at*

least $1-\delta_{in}$. *Then for any arbitrary small* $\delta > 0$ *there are constants* $C_1^{(l+1)}, C_2^{(l+1)}, n^{l+1} > 0$ *such that for any* $n > n^{l+1}$

$$|\Theta^{(l+1)}(x_1,x_2)_{ij}| \leq \begin{cases} C_1^{(l+1)} n & \text{if } i=j \\ C_2^{(l+1)}\sqrt{n} & \text{if } i \neq j \end{cases} \tag{32}$$

with probability at least $1-\delta$. *In other words, the magnitude of elements of the tangent kernel in the recursive definition will not grow.*

Proof. The proof for the bound on diagonal elements is available in the proof of the previous lemma and follows the exact same structure and obtains the same bounds. Thus, we avoid repeating it here. The bound for non-diagonal elements however slightly changes due to the change in the input tangent kernel and thus needs to be proven. For the non-diagonal elements of $\Theta^{(l+1)}(x_1,x_2)$ we have:

$$\begin{aligned} |\Theta^{(l+1)}(x_1,x_2)_{ij}| &= \sum_{a=1}^{n}\sum_{b=1}^{n} |W^{(l+1)}{}_{ia}||W^{(l+1)}{}_{jb}||\Theta^{(l)}(x_1,x_2)_{ab}| \\ &= \underbrace{\frac{1}{n}\sum_{a=1}^{n}|z_{ia}||z_{ja}|\Theta^{(l)}(x_1,x_2)_{aa}}_{\Theta_1^{(l+1)}(x_1,x_2)_{ij}} + \underbrace{\frac{1}{n}\sum_{a=1}^{n}\sum_{b=1,b\neq a}^{n}|z_{ia}||z_{jb}||\Theta^{(l)}(x_1,x_2)_{ab}|}_{\Theta_2^{(l+1)}(x_1,x_2)_{ij}} \end{aligned} \tag{33}$$

where

$$|\Theta_1^{(l+1)}(x_1,x_2)_{ij}| \leq \frac{C_1^{(l)} n}{n}\sum_{a=1}^{n}|z_{ia}||z_{ja}| \sim C_1^{(l)}\sqrt{n} SE\left(\nu_p, \frac{\alpha_p\sqrt{n}}{n}\right) \tag{34}$$

and

$$\begin{aligned} |\Theta_2^{(l+1)}(x_1,x_2)_{ij}| &\leq \frac{C_2^{(l)}\sqrt{n}}{n}\sum_{a=1}^{n}\sum_{b=1,b\neq a}^{n}|z_{ia}||z_{jb}| \\ &\sim \frac{C_2^{(l)}\sqrt{n}}{n}\sum_{a=1}^{n^2} SE\left(\nu_p,\alpha_p\right) = C_2^{(l)}\sqrt{n} SE\left(\nu_p,\frac{\alpha_p}{n}\right) \end{aligned} \tag{35}$$

each with probability at least $1-\delta_{in}$. Thus, we can claim that according to Corollary 9 that there exists $n^{l+1} > 0$ such that for $n > n^{l+1}$ we have that

$$|\Theta^{(l+1)}(x_1,x_2)_{ij}| \leq C^{(l)}\sqrt{n}\left(\sqrt{2\nu_p \log\frac{2}{\delta}} + \frac{2}{\pi}\right) \tag{36}$$

with probability at least $(1-\delta)(1-\delta_{in})^2$. Thus, the lemma's claim holds with probability at least $(1-\delta)(1-\delta_{in})^2(1-\delta_g)$ and $n > n^{l+1}$) as desired.

□

An alert reader already can notice that connecting the previous four lemmas would result in an upper bound for the diagonal and non-diagonal elements of the eNTK of the NN f at initialization.

Lemma 14. *Consider a NN* f *under Setting A. For every arbitrary small* $\delta > 0$, *the corresponding eNTK of* f *on the arbitrary datapoints* x_1 *and* x_2 *satisfies*

$$|\Theta(x_1,x_2)_{ij}| \leq \begin{cases} C_1 n & \text{if } i=j \\ C_2\sqrt{n} & \text{if } i \neq j \end{cases} \tag{37}$$

with probability at least $1-\delta$ *over random initialization for some* $C_1, C_2, n_0 > 0$ *where* $n > n_0$.

Proof. The proof is straightforward, starting with Lemma 10 and applying Lemma 11, Lemma 12 and Lemma 13 consecutively one can show derive the mentioned bound. □

Lemma 15. *Consider a NN f under Setting A. For every arbitrary small $\delta > 0$ and the arbitrary datapoints x_1 and x_2, it holds that*

$$\|\Theta(x_1, x_2) - \hat{\Theta}(x_1, x_2)\|_F \leq \mathcal{O}(\sqrt{n}) \tag{38}$$

with probability at least $1 - \delta$ over random initialization. In other words, the Frobenius norm of the difference between eNTK and pNTK evaluated on two datapoints are bounded by $\mathcal{O}(\sqrt{n})$.

Proof. We note by $D(x_1, x_2) = \Theta^{(L)}(x_1, x_2) - \hat{\Theta}^{(L)}(x_1, x_2) \otimes I_O$. Using the expansion provided in Equation (14) we can write the

$$|D(x_1, x_2)_{ij}| \leq \frac{1}{n} \begin{cases} \sum_{a=1}^{n}\sum_{b=1}^{n} |z_{ia}||z_{ib}||\Theta^{(L-1)}(x_1, x_2)_{ab}| - \frac{1}{O}\sum_{c=1}^{O}\sum_{d=1}^{O}\sum_{a=1}^{n}\sum_{b=1}^{n} |z_{ca}||z_{db}||\Theta^{(L-1)}(x_1, x_2)_{ab}| & \text{if } i = j \\ \sum_{a=1}^{w}\sum_{b=1}^{w} |z_{ia}||z_{jb}||\Theta^{(L-1)}(x_1, x_2)_{ab}| & \text{if } i \neq j \,. \end{cases}$$

Applying Lemma 14 we can assume there are constants $C_1^{(L-1)}, C_2^{(L-1)} > 0$ such that $|\Theta^{(L-1)}(x_1, x_2)_{aa}| < C_1^{(L-1)} n$ and $|\Theta^{(L-1)}(x_1, x_2)_{ab}| < C_2^{(L-1)}\sqrt{n}$ with probability at least $1 - \delta$. Thus we can write the diagonal elements of $D(x_1, x_2)$ as

$$\begin{aligned} |D(x_1, x_2)_{ii}| \leq & \underbrace{C_1^{(L-1)} \sum_{a=1}^{n} \left(z_{ia}^2 - \frac{1}{O}\sum_{c=1}^{O} z_{ca}^2 \right)}_{D_1(x_1,x_2)_{ii}} \underbrace{- \frac{C_1^{(L-1)}}{O} \sum_{c=1}^{O}\sum_{d=1, d\neq c}^{O} |z_{ca}||z_{da}|}_{D_2(x_1,x_2)_{ii}} \\ & + \underbrace{\frac{C_2^{(L-1)}}{\sqrt{n}} \sum_{a=1}^{n}\sum_{b=1, b\neq a}^{n} |z_{ia}||z_{ib}|}_{D_3(x_1,x_2)_{ii}} \underbrace{- \frac{C_2^{(L-1)}}{O\sqrt{n}} \sum_{a=1}^{n}\sum_{b=1, b\neq a}^{n}\sum_{c=1}^{O}\sum_{d=1}^{O} |z_{ca}||z_{db}|}_{D_4(x_1,x_2)_{ii}} . \end{aligned}$$

We would like to bound each of $D_k(x_1, x_2)_{ii}$ for $k \in \{1, 2, 3, 4\}$ and then find a bound for the diagonal elements using the combination of them. Starting with $D_1(x_1, x_2)_{ii}$:

$$\begin{aligned} |D_1(x_1, x_2)_{ii}| &\leq C_1^{(L-1)} \sum_{a=1}^{n} \left(z_{ia}^2 - \frac{1}{O}\sum_{c=1}^{O} z_{ca}^2 \right) \\ &= C_1^{(L-1)} \sum_{a=1}^{n} \left(z_{ia}^2 - \frac{1}{O} z_{ia}^2 - \frac{1}{O}\sum_{c=1, c\neq i}^{O} z_{ca}^2 \right) \\ &\sim C_1^{(L-1)} \sum_{a=1}^{n} \left(SE\left(\frac{2(O-1)}{O}, \frac{4(O-1)}{O} \right) - SE\left(\sqrt{\frac{(O-1)2^2}{O^2}}, \frac{4}{O} \right) \right) \\ &= C_1^{(L-1)} \sum_{a=1}^{n} SE\left(\sqrt{\frac{4(O-1)^2}{O^2} + \frac{4(O-1)}{O^2}}, \max\left(\frac{4(O-1)}{O}, \frac{4}{O} \right) \right) \\ &= C_1^{(L-1)} \sqrt{n} SE\left(\sqrt{1 - \frac{1}{O}}, \frac{1}{\sqrt{n}}\left(1 - \frac{1}{O} \right) \right) \end{aligned} \tag{39}$$

Thus, using Corollary 9, we can claim

$$|D_1(x_1, x_2)_{ii}| < C_1^{(L-1)} \sqrt{n} \left(\sqrt{2\left(1 - \frac{1}{O} \right) \log \frac{8}{\delta}} + 1 \right) \tag{40}$$

with probability at least $1-\delta/4$.

For the other terms, we can simply the analysis through noting that they are all a form of weighted summation of independent sub-exponential random variables of the same distribution. For such a summation with a weight of a and b summation terms, we have

$$X = a\sum_{i=1}^{b}|z_i||z_i'| \sim a\sum_{i=1}^{b} SE\left(\nu_p, \alpha_p\right) = aSE\left(\nu_p\sqrt{b}, \alpha_p\right). \tag{41}$$

Thus, applying Corollary 9, we get that

$$|X| < 2a\left(\max\left(\nu_p\sqrt{b\log\frac{8}{\delta}}, \alpha_p\log\frac{8}{\delta}\right) + |\mathbb{E}[X]|\right) \tag{42}$$

with probability at least $1-\delta/4$. Accordingly we can claim

$$|D_2(x_1,x_2)_{ii}| < 2C_1^{(L-1)}\left(\max\left(\nu_p\sqrt{\left(1-\frac{1}{O^2}\right)\log\frac{8}{\delta}}, \alpha_p\log\frac{8}{\delta}\right) + \frac{2}{\pi}\right), \tag{43}$$

$$|D_3(x_1,x_2)_{ii}| < 2C_2^{(L-1)}\sqrt{n}\left(\max\left(\nu_p\sqrt{\left(1-\frac{1}{n^2}\right)\log\frac{8}{\delta}}, \alpha_p\log\frac{8}{\delta}\right) + \frac{2}{\pi}\right), \tag{44}$$

$$|D_4(x_1,x_2)_{ii}| < 2C_2^{(L-1)}\sqrt{n}\left(\max\left(\nu_p\sqrt{\left(1-\frac{1}{O^2}-\frac{1}{n^2}+\frac{1}{O^2n^2}\right)\log\frac{8}{\delta}}, \alpha_p\log\frac{8}{\delta}\right) + \frac{2}{\pi}\right), \tag{45}$$

all independently and with probability at least $1-\delta/4$. Moreover, one can easily apply the same technique and see that $D(x_1,x_2)_{ij}$ for $i \neq j$ follows a similar bound to the one of Equation (44).

Thus, loosening the off-diagonal terms for simplicity, applying a union bound on the previous three inequalities yields

$$|D(x_1,x_2)_{ij}| < 8\left(C_1^{(L-1)} + C_2^{(L-1)}\right)\sqrt{n}\left(\max\left(\nu_p\sqrt{\log\frac{8}{\delta}}, \alpha_p\log\frac{8}{\delta}\right) + \frac{2}{\pi}\right) \tag{46}$$

with probability at least $1-\delta$.

Finally, as $\|D(x_1,x_2)\|_F = \sqrt{\sum_{i,j} D(x_1,x_2)_{ij}^2}$, if each entry's absolute value is less than $t > 0$ then the Frobenius norm is less than tO. Thus we can combine a bound on each of the O^2 entries to see that

$$\Pr\left(\|D(x_1,x_2)\|_F \leq 8O\left(C_1^{(L-1)} + C_2^{(L-1)} + 1 + \frac{2}{\pi}\right)\sqrt{n}\max\left(2\sqrt{\log\frac{8O^2}{\delta}}, 4\log\frac{8O^2}{\delta}\right)\right) \geq 1-\delta \tag{47}$$

as desired.

□

Lemma 16. *Consider a NN f under Setting A. For every arbitrary small $\delta > 0$ and the arbitrary datapoints x_1 and x_2, it holds that*

$$\|\Theta(x_1,x_2)\|_F \geq \Omega(n) \tag{48}$$

with probability at least $1-\delta$ over random initialization. In other words, the Frobenius norm of the eNTK evaluated on two datapoints is lower bounded by $\Omega(n)$.

Proof. Considering that the dot product of post-activations appear in the diagonal elements of the eNTK in conjunction with Lemma 18, this is straightforward. Note that this bound also applies to the maximum eigenvalue of the eNTK matrix since the maximum eigenvalue is bigger (or equal to) than the sum of elements of the matrix divided by the number of columns. □

We are finally ready to present the proof of Theorem 1.

Theorem 17. *Consider a NN f under Setting A. For every arbitrary small $\delta > 0$ and the arbitrary datapoints x_1 and x_2, there exists n_0 such that*

$$\frac{\|\Theta^{(L)}(x_1, x_2) - \hat{\Theta}^{(L)}(x_1, x_2) \otimes I_O\|_F}{\|\Theta^{(L)}(x_1, x_2)\|_F} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \tag{49}$$

with probability at least $1 - \delta$ for $n > n_0$.

Proof. The proof is straightforward from applying Lemma 15 and **??**. □

Lemma 18. *Consider a NN under Setting A with $L \geq 2$ and ReLU activation function. The dot product of two post-activations ${f^{(l)}(x_1)}^\top f^{(l)}(x_2)$ grows linearly with the width of the network with high probability over random initialization.*

Proof. We already know that when the MLP is parameterized according to He et al. (2015) (or also NTK-parameterization), the pre-activations of every layer will have bounded variance and be distributed with mean zero. Thus, applying the Jensen's inequality on each coordinate on post-activations of layer l we have that

$$\mathbb{E}\left[f_i^{(l)}(x)\right] = \mathbb{E}\left[\phi\left(W_i^{(l)} f^{(l-1)}(x)\right)\right] \geq \phi\left(\mathbb{E}\left[W_i^{(l)} f^{(l-1)}(x)\right]\right) \tag{50}$$

where the right hand side of the above inequality is zero in the case of ReLU activation. Now note that $\mathbb{E}\left[f_i^{(l)}(x)\right]$ will be zero only and only if $f^{(l-1)}(x) = 0$. Assuming this is not the case (as it is only the case with zero probability under Setting A) we have that $\mu = \mathbb{E}\left[\phi\left(W_i^{(l)} f^{(l-1)}(x)\right)\right] > 0$. Thus, we have that

$$\mathbb{E}\left[\sum_{i=1}^{n} f_i^{(l)}(x_1) f_i^{(l)}(x_2)\right] = \sum_{i=1}^{n} \mathbb{E}\left[f_i^{(l)}(x_1)\right] \mathbb{E}\left[f_i^{(l)}(x_2)\right] = n\mu \tag{51}$$

with bounded variance. Accordingly, one can come up with constants $G_1^{(l)}, G_2^{(l)}, n_0 > 0$ for any arbitrary small $\delta > 0$ such that for $n > n_0$ and $L \geq 2$ (depth of the network), $G_1^{(l)} n < {f^{(l)}(x_1)}^\top f^{(l)}(x_2) < G_2^{(l)} n$ with probability at least $1 - \delta$.

□

B.3 Pseudo-NTK's maximum eigenvalue converges to eNTK's maximum eigenvalue as width grows

In this subsection, we present a formal proof for Theorem 4.

Proof. Note that, as both pNTK and eNTK are symmetric PSD matrices, their maximum eigenvalues are equal to their spectral norm. Furthermore, the spectral norm of a matrix is upper-bounded by its Frobenius norm. Now, note that according to the triangle inequality, we have

$$\begin{aligned} \|\Theta(x_1, x_2)\| &= \|\hat{\Theta}(x_1, x_2) \otimes I_O + \Big(\Theta(x_1, x_2) - \hat{\Theta}(x_1, x_2) \otimes I_O\Big)\| \\ &\leq \|\hat{\Theta}(x_1, x_2) \otimes I_O\| + \|\Theta(x_1, x_2) - \hat{\Theta}(x_1, x_2) \otimes I_O\| \end{aligned} \tag{52}$$

Thus

$$\|\Theta(x_1, x_2)\| - \|\hat{\Theta}(x_1, x_2) \otimes I_O\| \leq \|\Theta(x_1, x_2) - \hat{\Theta}(x_1, x_2) \otimes I_O\|. \tag{53}$$

which according to (47) together with the fact that for any matrix A, $\lambda_{\max}(A \otimes I) = \lambda_{\max}(A)$ implies that with probability at least $1 - \delta$,

$$\left|\lambda_{\max}\left(\Theta(x_1, x_2)\right) - \lambda_{\max}\left(\hat{\Theta}(x_1, x_2)\right)\right| \leq \\ 8O\left(C_1^{(L-1)} + C_2^{(L-1)}\right)\sqrt{n}\max\left(\sqrt{\log\frac{8O^2}{\delta}}, \sqrt{2}\log\frac{8O^2}{\delta}\right). \tag{54}$$

Moreover, as mentioned in the proof of Lemma 16, combining the previous inequality with the fact that $\lambda_{\max}\left(\Theta(x_1, x_2)\right) \geq \Omega(n)$ with high probability shows that there exists δ' and n_0 such that

$$\left|\frac{\lambda_{\max}\left(\Theta(x_1, x_2)\right) - \lambda_{\max}\left(\hat{\Theta}(x_1, x_2)\right)}{\lambda_{\max}\left(\Theta(x_1, x_2)\right)}\right| \leq \mathcal{O}(1/\sqrt{n}) \tag{55}$$

with probability $1 - \delta'$ over random initialization for $n > n_0$ as desired.

□

B.4 Kernel regression using pNTK vs Kernel regression using eNTK

In this subsection we provide a formal proof for Theorem 5.

Proof. We start by proving a simpler version of a theorem, and then show a correspondence that expands the result of the simpler proof to the original Theorem. Assuming $|\mathcal{X}| = |\mathcal{Y}| = N$ (training data), we define

$$h(x) = \Theta(x_1, \mathcal{X})\Theta(\mathcal{X}, \mathcal{X})^{-1}\mathcal{Y} \text{ and } \hat{h}(x) = \left(\hat{\Theta}(x_1, \mathcal{X}) \otimes I_O\right)\left(\hat{\Theta}(\mathcal{X}, \mathcal{X}) \otimes I_O\right)^{-1}\mathcal{Y}. \tag{56}$$

Note that as the result of kernel regression (without any regularization) does not change with scaling the kernel with a fixed scalar, we can use a weighted version of the kernels mentioned in the previous equation without loss of generality. Accordingly, we define

$$\alpha = \left(\frac{1}{n}\Theta(\mathcal{X}, \mathcal{X})\right)^{-1}\mathcal{Y} \text{ and } \hat{\alpha} = \left(\frac{1}{n}\hat{\Theta}(\mathcal{X}, \mathcal{X}) \otimes I_O\right)^{-1}\mathcal{Y}. \tag{57}$$

Using the fact that $\hat{M}^{-1} - M^{-1} = -\hat{M}^{-1}(\hat{M} - M)M^{-1}$ and $(A \otimes I)^{-1} = A^{-1} \otimes I$ we can show that

$$\hat{\alpha} - \alpha = -\hat{\Theta}(\mathcal{X}, \mathcal{X})^{-1} \otimes I_O\left(\frac{1}{n}\hat{\Theta}(\mathcal{X}, \mathcal{X}) \otimes I_O - \frac{1}{n}\Theta(\mathcal{X}, \mathcal{X})\right)^{-1}\Theta(x_1, x_2)\mathcal{Y} \tag{58}$$

Assume $\lambda = \min\left(\lambda_{\min}(\Theta(\mathcal{X}, \mathcal{X})), \lambda_{\min}(\hat{\Theta}(\mathcal{X}, \mathcal{X}))\right)$. Then

$$\|\hat{\alpha} - \alpha\| \leq \frac{1}{\lambda^2}\|\frac{1}{n}\hat{\Theta}(\mathcal{X}, \mathcal{X}) \otimes I_O - \frac{1}{n}\Theta(\mathcal{X}, \mathcal{X})\|\|\mathcal{Y}\| \tag{59}$$

Plugging into the formula for kernel regression, we get that

$$\begin{aligned}\hat{h}(x) - h(x) &= \left(\frac{1}{n}\hat{\Theta}(x, \mathcal{X}) \otimes I_O\right)\hat{\alpha} - \frac{1}{n}\Theta(x, \mathcal{X})\alpha \\ &= \left(\frac{1}{n}\hat{\Theta}(x, \mathcal{X}) \otimes I_O - \frac{1}{n}\Theta(x, \mathcal{X})\right)\hat{\alpha} + \frac{1}{n}\Theta(x, \mathcal{X})(\hat{\alpha} - \alpha)\end{aligned} \tag{60}$$

Thus

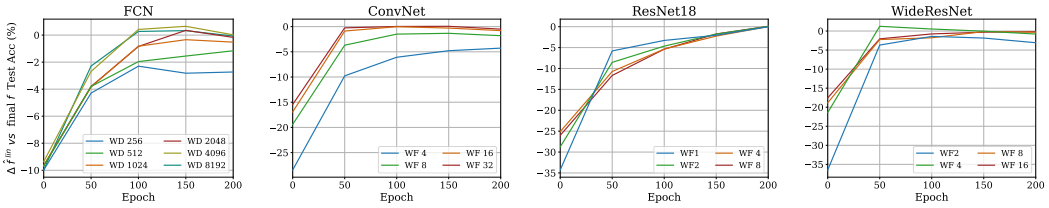

Figure 12: Evaluating the **difference in test accuracy of kernel regression using pNTK as in (6) vs the final model** f throughout SGD training on the full CIFAR-10 dataset. How much worse would it be to "give up" on SGD at this point and train $\hat{f}^{lin}$ with the current representation?

$$\begin{aligned}\|\hat{h}(x) - h(x)\| &\leq \|\frac{1}{n}\hat{\Theta}(x,\mathcal{X}) \otimes I_O - \frac{1}{n}\Theta_f(x,\mathcal{X})\|\|\hat{\alpha}\| + \|\frac{1}{n}\Theta(x,\mathcal{X})\|\|\hat{\alpha} - \alpha\| \\ &\leq \frac{1}{\lambda}\|\frac{1}{n}\hat{\Theta}(x,\mathcal{X}) \otimes I_O - \frac{1}{n}\Theta(x,\mathcal{X})\|\|\mathcal{Y}\| \\ &\quad + \frac{1}{\lambda^2}\|\frac{1}{n}\Theta(x,\mathcal{X})\|\|\frac{1}{n}\hat{\Theta}(\mathcal{X},\mathcal{X}) \otimes I_O - \frac{1}{n}\Theta(\mathcal{X},\mathcal{X})\|\|\mathcal{Y}\|.\end{aligned} \tag{61}$$

Now, note that as for a block matrix A of A_{ij} blocks we have that $\|A\| \leq \sum_{i,j}\|A_{ij}\|$ it follows that for any matrix valued kernel K

$$\|K(\mathcal{X},\mathcal{X})\| \leq \sum_{x_1,x_2\in\mathcal{X}} \|K(x_1,x_2)\|. \tag{62}$$

Using this fact, we can rewrite the bound as

$$\begin{aligned}\|\hat{h}(x) - h(x)\| &\leq \frac{N}{\lambda}\|\frac{1}{n}\hat{\Theta}(x,x_1^*) \otimes I_O - \frac{1}{n}\Theta(x,x_1^*)\|\|\mathcal{Y}\| \\ &\quad + \frac{N^2}{\lambda^2}\|\frac{1}{n}\Theta(x,\mathcal{X})\|\|\frac{1}{n}\hat{\Theta}(x_2^*,x_3^*) \otimes I_O - \frac{1}{n}\Theta(x_2^*,x_3^*)\|\|\mathcal{Y}\|\end{aligned} \tag{63}$$

for some particular $x_1^*, x_2^*, x_3^* \in \mathcal{X}$. Using (47), we can see with probability at least $1-\delta$ that

$$\left\|\hat{h}(x) - h(x)\right\| \leq \frac{8NO\alpha}{\lambda\sqrt{n}} \max\left(\sqrt{\log\frac{8O^2}{\delta}}, \sqrt{2}\log\frac{8O^2}{\delta}\right)\|\mathcal{Y}\|\left(1 + \frac{N}{\lambda}\|\frac{1}{n}\Theta(x,\mathcal{X})\|\right). \tag{64}$$

To show the correspondence between $\hat{h}(x)$ and $\hat{f}^{lin}(x)$, as in (6), note that

$$\begin{aligned}\hat{h}(x) &= \left(\hat{\Theta}(x,\mathcal{X}) \otimes I_O\right)\left(\hat{\Theta}(\mathcal{X},\mathcal{X})^{-1} \otimes I_O\right)\mathcal{Y} \\ &= \left(\hat{\Theta}(x,\mathcal{X})\hat{\Theta}(\mathcal{X},\mathcal{X})^{-1} \otimes I_O\right)\mathcal{Y} \\ &= \text{vec}\left(I_O\mathcal{Y}_v\hat{\Theta}(x,\mathcal{X})\hat{\Theta}(\mathcal{X},\mathcal{X})^{-1}\right)\end{aligned} \tag{65}$$

where $\mathcal{Y}_v = \text{vec}^{-1}(\mathcal{Y})$ is the result of inverse of the vectorization operation, converting the $NO \times 1$ vector to a $O \times N$ matrix. Thus, $\hat{h}(x) = \hat{\Theta}(x,\mathcal{X})\hat{\Theta}(\mathcal{X},\mathcal{X})^{-1}\mathcal{Y}'$ where $\mathcal{Y}'$ is the $N \times O$ matrix derived from reshaping the $NO \times 1$ vector $\mathcal{Y}$. The proof is complete. □

C More details on kernel Regression using pNTK on full CIFAR-10 dataset

In this section we provide another figure comparing the accuracy of $\hat{f}^{lin}(x)$ with parameters derived at epoch $E \in \{0, 50, 100, 150, 200\}$ of training the NN with SGD. On the y-axis, the reported number is $f^{lin}(x) - f^*(x)$ where f^* denotes the final model obtained after training f for 200

epochs. As seen in Figure 12 the architecture of the model has a significant impact on how good the linearization predicts the final accuracy of the fully-trained model. However, as proven in Theorem 1 in conjunction with the linearization approximations provided in Lee et al. (2019), as width grows, this approximation becomes more accurate. One unexplored fact regarding this experiment is that fact that lineraization with trained parameters significantly outperforms linearization at initialization, which is intuitive but not rigorously investigated yet.