Neural Tangent Kernel at Initialization: Linear Width Suffices (Supplementary Material)

A NEURAL TANGENT KERNEL AT INITIALIZATION

A.1 PROOF OF LEMMA 4.1

We start with a result which explicitly shows that the output $\|\alpha^{(l)}\|_2$ is sub-Gaussian for Gaussian random weights. We recall that we are assuming $\phi(0) = 0$, and note that the result can be extended to the general case straightforwardly.

Lemma A.1. *Let $A^{(l)} = [\alpha^{(l)}(\mathbf{x}_i)] \in \mathbb{R}^{n\times m_l}$ be the outputs of layer l. For $g \sim \mathcal{N}(\mathbf{0}_{m_{l-1}}, \sigma^2\mathbb{I}_{m_{l-1}})$, let $\vartheta^{(l)} = \phi(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g) \in \mathbb{R}^n$. Then, $\|\vartheta^{(l)}\|_2$ is a sub-Gaussian random variable with*

$$\left\|\left\|\vartheta^{(l)}\right\|_2\right\|_{\psi_2} = \left\|\left\|\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)\right\|_2\right\|_{\psi_2} \leq c\frac{\sqrt{2}(\sqrt{\log n}+1)\sigma}{\sqrt{m_{l-1}}}\|A^{(l-1)}\|_F$$

for some absolute constant $c > 0$.

Proof. First, note that since ϕ is 1-Lipschitz and $\phi(0) = 0$, we have $\left\|\phi(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g)\right\|_2 \leq \left\|\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right\|_2$. Now,

$$\begin{aligned}
\mathbb{P}\left(\left\|\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right\|_2^2 \geq \epsilon^2\frac{1}{m_{l-1}}\|A^{(l-1)}\|_F^2\right) &= \mathbb{P}\left(\frac{1}{m_{l-1}}\sum_{i=1}^n\langle\alpha_i^{(l-1)}, g\rangle^2 \geq \epsilon^2\frac{1}{m_{l-1}}\sum_{i=1}^n\|\alpha_i^{(l-1)}\|_2^2\right)\\
&\leq \sum_{i=1}^n\mathbb{P}\left(\frac{1}{m_{l-1}}\langle\alpha_i^{(l-1)}, g\rangle^2 \geq \epsilon^2\frac{1}{m_{l-1}}\|\alpha_i^{(l-1)}\|_2^2\right)\\
&= \sum_{i=1}^n\mathbb{P}\left(|\langle\alpha_i^{(l-1)}, g\rangle| \geq \epsilon\|\alpha_i^{(l-1)}\|_2\right)\\
&\leq 2n\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right) .
\end{aligned}$$

In other words, with $\tilde{\epsilon} = \frac{\epsilon}{\sqrt{m_{l-1}}}\|A^{(l-1)}\|_F$, for all $\tilde{\epsilon} > 0$ we have

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right\|_2 \geq \tilde{\epsilon}\right) \leq 2n\exp\left(-\frac{m_{l-1}\tilde{\epsilon}^2}{2\sigma^2\|A^{(l-1)}\|_F^2}\right) .$$

Now, with $\tilde{\epsilon} = \frac{\sigma\sqrt{2\log n}}{\sqrt{m_{l-1}}}\|A^{(l-1)}\|_F + \epsilon$, for all $\epsilon > 0$ we have

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right\|_2 \geq \frac{\sigma\sqrt{2\log n}}{\sqrt{m_{l-1}}}\|A^{(l-1)}\|_F + \epsilon\right) \leq 2n\exp\left(-\log n\right)\exp\left(-\frac{m_{l-1}\epsilon^2}{2\sigma^2\|A^{(l-1)}\|_F^2}\right)$$

$$\Rightarrow \quad \mathbb{P}\left(\left\|\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)\right\|_2 \geq \frac{\sigma\sqrt{2\log n}}{\sqrt{m_{l-1}}}\|A^{(l-1)}\|_F + \epsilon\right) \leq 2\exp\left(-\frac{m_{l-1}\epsilon^2}{2\sigma^2\|A^{(l-1)}\|_F^2}\right) .$$

Then, from Proposition A.1, it follows that $\|\phi(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g)\|_2$ is sub-Gaussian with

$$\left\|\left\|\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)\right\|_2\right\|_{\psi_2} \leq c\frac{\sqrt{2}(\sqrt{\log n}+1)\sigma}{\sqrt{m_{l-1}}}\|A^{(l-1)}\|_F ,$$

for some absolute constant c. This completes the proof. □

Proposition A.1. *Let $a_1, a_2 > 0$. If a non-negative random variable Z satisfies $\mathbb{P}(Z \geq a_1 + \epsilon) \leq 2\exp(-\epsilon^2/a_2^2)$, then $\|Z\|_{\psi_2} \leq c(a_1 + a_2)$, where c is an absolute constant.*

Proof. Note that $Z - a_1 = [Z-a_1]_- + [Z-a_1]_+$. Since Z is non-negative, $|[Z-a_1]_-| \leq a_1$ implying $\|[Z-a_1]_-\|_{\psi_2} \leq c_1 a_1$, where c_1 is an absolute constant. Further, by definition, $[Z-a_1]_+$ is sub-Gaussian with $\|[Z-a_1]_+\|_{\psi_2} \leq c_2 a_2$, where c_2 is an absolute constant. Now, by triangle inequality

$$\begin{aligned}\|Z\|_{\psi_2} &= \|a_1 + [Z-a_1]_- + [Z-a_1]_+\|_{\psi_2} \\ &\leq a_1 + \|[Z-a_1]_-\|_{\psi_2} + \|[Z-a_1]_+\|_{\psi_2} \leq a_1 + c_1 a_1 + c_2 a_2 \\ &\leq c(a_1 + a_2) ,\end{aligned}$$

where $c = \max(1 + c_1, c_2)$. That competes the proof. □

We are now ready to prove Lemma 4.1.

Proof of Lemma 4.1. Let $\vartheta^{(l)} = \phi(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g) \in \mathbb{R}^n$. If $m_l < n$, then $\lambda_{\min}(A^{(l)}(A^{(l)})^\top) = 0$. So, we assume $m_l \geq n$. Let $w_j \in \mathbb{R}^{m_{l-1}}$ denote the j-th row of $W_0^{(l)}$. For a given $t > 0$, let $\hat{A}^{(l)} \in \mathbb{R}^{n\times m_l}$ so that $\hat{A}^{(l)}_{:,j} = \phi(\frac{1}{\sqrt{m_l}}A^{(l-1)}w_j^\top)\mathbb{1}_{[\|\phi(\frac{1}{\sqrt{m_l}}A^{(l-1)}w_j^\top\|_2 \leq t]} \in \mathbb{R}^{m_{l-1}}$ and $\hat{\vartheta}^{(l)} = \phi(\frac{1}{\sqrt{m_l}}A^{(l-1)}g)\mathbb{1}_{[\|\phi(\frac{1}{\sqrt{m_l}}A^{(l-1)}g\|_2 \leq t]} \in \mathbb{R}^n$. Then, we have

$$\begin{aligned}&(i)\ \ \lambda_{\min}(A^{(l)}(A^{(l)})^\top) \geq \lambda_{\min}(\hat{A}^{(l)}(\hat{A}^{(l)})^\top) , \\ &(ii)\ \ \lambda_{\max}(\hat{A}^{(l)}_{:,j}(\hat{A}^{(l)}_{:,j})^\top) \leq t^2 ,\end{aligned}$$

where (i) follows immediately by definition of $\hat{A}^{(l)}$ and (ii) follows since for any unit vector $v \in \mathbb{R}^n$, $v^\top \hat{A}^{(l)}_{:,j}(\hat{A}^{(l)}_{:,j})^\top v = \langle v, \hat{A}^{(l)}_{:,j}\rangle^2 \leq \|\hat{A}^{(l)}_{:,j}\|_2^2 \leq t^2$.

From Lemma A.1, we know that $\|\|\vartheta^{(l)}\|_2\|_{\psi_2} \leq c_1\frac{\sqrt{2}(\sqrt{\log n}+1)\sigma}{\sqrt{m}}\|A^{(l-1)}\|_F$. Recall that for any subGaussian random variable Z, $\mathbb{P}(Z \geq t) \leq \exp(1 - ct^2/\|Z\|_{\psi_2}^2)$ for some absolute constant c. For our analysis with $\|\vartheta^{(l)}\|_2$, for a suitable constant $a > 0$ we will use

$$t = \frac{\sqrt{2}(\sqrt{\log n}+1)\sigma\|A^{(l-1)}\|_F}{\sqrt{cm_{l-1}}}\sqrt{\max\left(1, \log\frac{2a(\sqrt{\log n}+1)^2\sigma^2\|A^{(l-1)}\|_F^2}{c\lambda_l m_{l-1}}\right)} . \tag{1}$$

Let

$$\begin{aligned}G_l &:= \mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_{m_{l-1}},\sigma^2\mathbb{I}_{m_{l-1}})}\left[\vartheta^{(l)}(\vartheta^{(l)})^\top\right] = \mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_{m_{l-1}},\sigma^2\mathbb{I}_{m_{l-1}})}\left[\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)^\top\right] , \\ \hat{G}_l &:= \mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_{m_{l-1}},\sigma^2\mathbb{I}_{m_{l-1}})}\left[\hat{\vartheta}^{(l)}(\hat{\vartheta}^{(l)})^\top\right] \\ &= \mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_{m_{l-1}},\sigma^2\mathbb{I}_{m_{l-1}})}\left[\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)^\top\mathbb{1}_{[\|\phi(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g)\|_2 \leq t]}\right] .\end{aligned}$$

Note that $\lambda_l = \lambda_{\min}(G_l)$.

By Matrix Chernoff bound, for any $\epsilon \in [0,1)$, we have

$$\mathbb{P}\left(\lambda_{\min}(\hat{A}^{(l)}(\hat{A}^{(l)})^\top) \le (1-\epsilon)\lambda_{\min}(\mathbb{E}_{W_0^{(l)}}[\hat{A}^{(l)}(\hat{A}^{(l)})^\top])\right) \le n\left(\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right)^{\lambda_{\min}\left(\mathbb{E}_{W_0^{(l)}}[\hat{A}^{(l)}(\hat{A}^{(l)})^\top]\right)/t^2} .$$

For $\epsilon = 1/2$, with $c_3 = \frac{1}{2}(1-\log 2)$, we have

$$\mathbb{P}\left(\lambda_{\min}(\hat{A}^{(l)}(\hat{A}^{(l)})^\top) \le \frac{m_l}{2}\lambda_{\min}(\hat{G}_l)\right) \le \exp\left(-\frac{c_3 m_l}{t^2}\lambda_{\min}(\hat{G}) + \log n\right) ,$$

where we used the fact that $\mathbb{E}_{W_0^{(l)}}[\hat{A}^{(l)}(\hat{A}^{(l)})^\top] = m_l \mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_{m_{l-1}},\sigma^2\mathbb{I}_{m_{l-1}})}[\hat{\vartheta}^{(l)}(\hat{\vartheta}^{(l)})^\top] = m_l\hat{G}_l$.

With $m_l \ge \frac{t^2}{c_3\lambda_{\min}(\hat{G}_l)}\log\frac{n}{\bar{\delta}}$, with probability at least $(1-\bar{\delta})$ we have

$$\lambda_{\min}(\hat{A}^{(l)}(\hat{A}^{(l)})^\top) \ge \frac{m_l}{2}\lambda_{\min}(\hat{G}_l) . \tag{2}$$

Now, note that

$$\begin{aligned}
\|\hat{G}_l - G_l\|_2 &\le \mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_{m_{l-1}},\sigma^2\mathbb{I}_{m_{l-1}})}\Bigg\|\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)^\top \mathbb{1}_{[\|\phi(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g)\|_2\le t]} \\
&\qquad - \phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)^\top\Bigg\|_2 \\
&= \mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_{m_{l-1}},\sigma^2\mathbb{I}_{m_{l-1}})}\left[\left\|\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)\right\|_2^2 \mathbb{1}_{[\|\phi(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g)\|_2 > t]}\right] \\
&\overset{(a)}{=} \int_{s=0}^{\infty}\mathbb{P}\left(\left\|\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)\right\|_2 \mathbb{1}_{[\|\phi(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g)\|_2>t]} > \sqrt{s}\right) ds \\
&= \int_{s=0}^{\infty}\mathbb{P}\left(\left\|\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)\right\|_2 > t\right)\mathbb{P}\left(\left\|\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)\right\|_2 > \sqrt{s}\right) ds \\
&\overset{(b)}{\le} \exp(2)\exp\left(-c\frac{m_{l-1}t^2}{2(\sqrt{\log n}+1)^2\sigma^2\|A^{(l-1)}\|_F^2}\right)\int_{s=0}^{\infty}\exp\left(-c\frac{m_{l-1}s}{2(\sqrt{\log n}+1)^2\sigma^2\|A^{(l-1)}\|_F^2}\right) ds \\
&\overset{(c)}{=} \exp(2)\exp\left(-\max\left(1,\log\frac{2a(\sqrt{\log n}+1)^2\sigma^2\|A^{(l-1)}\|_F^2}{c\lambda_l m_{l-1}}\right)\right)\frac{2(\sqrt{\log n}+1)^2\sigma^2\|A^{(l-1)}\|_F^2}{cm_{l-1}} ,
\end{aligned}$$

where (a) follows since for any non-negative random variable Z, $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \ge s)ds$, (b) follows from Lemma A.1, and (c) follows from our choice of t in (1) and since for $b > 0$, $\int_0^\infty \exp(-s/b)ds = b$. To simplify further, we consider the following two exhaustive cases:

Case 1. Assume

$$\frac{2a(\sqrt{\log n}+1)^2\sigma^2\|A^{(l-1)}\|_F^2}{c\lambda_l m_{l-1}} \le \exp(1) \quad \Rightarrow \quad \frac{2(\sqrt{\log n}+1)^2\sigma^2\|A^{(l-1)}\|_F^2}{cm_{l-1}} \le \frac{\lambda_l}{a}\exp(1) .$$

Then,

$$\|\hat{G}_l - G_l\|_2 \le \exp(2)\exp(-1)\frac{2(\sqrt{\log n}+1)^2\sigma^2\|A^{(l-1)}\|_F^2}{cm_{l-1}} \le \exp(2)\exp(-1)\exp(1)\frac{\lambda_l}{a} = \frac{\exp(2)}{a}\lambda_l \overset{(a)}{\le} \frac{\lambda_l}{2} ,$$

where (a) follows if $a \ge 2\exp(2)$.

Case 2. On the other hand, assume

$$\frac{2a(\sqrt{\log n}+1)^2\sigma^2\|A^{(l-1)}\|_F^2}{cm_{l-1}\lambda_l} \ge \exp(1) .$$

Then,

$$\|\hat{G}_l - G_l\|_2 \leq \exp(2)\frac{c\lambda_l m_{l-1}}{2a(\sqrt{\log n}+1)^2\sigma^2\|A^{(l-1)}\|_F^2}\frac{2(\sqrt{\log n}+1)^2\sigma^2\|A^{(l-1)}\|_F^2}{cm_{l-1}} = \frac{\exp(2)}{a}\lambda_l \overset{(a)}{\leq} \frac{\lambda_l}{2},$$

where (a) follows if $a \geq 2\exp(2)$. Thus, choosing $a = 15$ in (1) ensures $\|\hat{G}_l - G_l\|_2 \leq \frac{\lambda_l}{2}$. As a result,

$$\lambda_{\min}(\hat{G}_l) \geq \lambda_{\min}(G_l) - \|\hat{G}_l - G_l\|_2 \geq \lambda_l/2. \tag{3}$$

Then, for $m_l \geq \frac{2t^2}{c_3\lambda_l}\log\frac{n}{\bar{\delta}}$, with probability at least $(1-\bar{\delta})$ we have

$$\lambda_{\min}(A^{(l)}(A^{(l)})^\top) \geq \lambda_{\min}(\hat{A}^{(l)}(\hat{A}^{(l)})^\top) \overset{(a)}{\geq} \frac{m_l}{2}\lambda_{\min}(\hat{G}) \overset{(b)}{\geq} \frac{m_l}{4}\lambda_l .$$

where (a) follows from (2) and (b) from (3). Finally, note that we have used $m_l \geq n$ and $m_l \geq \frac{t^2}{c_3\lambda_{\min}(\hat{G})}\log\frac{n}{\delta}$ in the above analysis. Then, with $v = \frac{2(\sqrt{\log n}+1)^2\sigma^2\|A^{(l-1)}\|_F^2}{c_3\lambda_l m_{l-1}}$, the choice of t in (1), and $a = 15$, and noting that $\lambda_{\min}(\hat{G}) \geq \lambda_l/2$, the analysis holds if we have

$$m_l \geq \max\left(n, c_2 v\max(1,\log(15v))\log\frac{n}{\bar{\delta}}\right) .$$

for some constant $c_2 > 0$. Choosing $\bar{\delta} = \frac{\delta}{L}$ completes the proof. □

A.2 PROOF OF LEMMA 4.2

Proof of Lemma 4.2. We do the proof by induction. Let $\mathbf{x} \in \{\mathbf{x}_i, i \in [n]\}$. For $l = 0$, $\|\alpha^{(0)}(\mathbf{x}_i)\|_2^2 = \|x_i\|_2^2 = c_{\phi,\sigma_0}d$. For $l = 0$ and $m_0 = d$, so the result is satisfied at $l = 0$ almost surely.

Assume that the result holds for a certain l, so that for any $i \in [n]$

$$c_{\phi,\sigma_0}\left(1 - \frac{h_C(l)}{2h_C(L)}\right)m_l \leq \min_{i\in[n]}\|\alpha^l(\mathbf{x})\|_2^2 \leq \max_{i\in[n]}\|\alpha^{(l)}(\mathbf{x})\|_2^2 \leq c_{\phi,\sigma_0}\left(1 + \frac{h_C(l)}{2h_C(L)}\right)m_l$$
$$\Rightarrow \quad \max_{i\in[n]}\left|\frac{\|\alpha^{(l)}(\mathbf{x})\|_2^2}{c_{\phi,\sigma_0}m_l} - 1\right| \leq \frac{h_C(l)}{2h_C(L)} . \tag{4}$$

We condition on $\{W_0^{(l')}, l' \in [l]\}$, and focus on layer $\alpha^{(l+1)}$ with random weights $W_0^{(l+1)} = [w_{1,:}; \cdots; w_{m_{l+1},:}] \in \mathbb{R}^{m_{l+1}\times m_l}$. Note that $\|\alpha^{(l+1)}(\mathbf{x})\|_2^2 = \sum_{j=1}^{m_{l+1}}(\alpha_j^{(l+1)}(\mathbf{x}))^2$. Since ϕ is 1-Lipschitz and $\phi(0) = 0$, $|\phi(a)| \leq |a|$, we have

$$\begin{aligned}
\mathbb{E}_{W_0^{(l+1)}}\|\alpha^{(l+1)}(\mathbf{x})\|_2^2 &= \sum_{j=1}^{m_{l+1}}\mathbb{E}_{w_{0,j,:}}[(\alpha_j^{(l+1)}(\mathbf{x}))^2] = m_{l+1}\mathbb{E}_{g\sim\mathcal{N}\left(\mathbf{0}_{m_l},\frac{\sigma_0^2}{c_{\phi,\sigma_0}}\mathbb{I}_{m_l}\right)}\left[\phi^2\left(\frac{1}{\sqrt{m_l}}\langle g, \alpha^{(l)}(\mathbf{x})\rangle\right)\right] \\
&= m_{l+1}\mathbb{E}_{g\sim\mathcal{N}\left(\mathbf{0}_{m_l},\frac{\sigma_0^2}{c_{\phi,\sigma_0}}\mathbb{I}_{m_l}\right)}\left[\phi^2\left(\frac{\|\alpha^{(l)}(\mathbf{x})\|_2}{\sqrt{m_l}}\left\langle g, \frac{\alpha^{(l)}(\mathbf{x})}{\|\alpha^{(l)}(\mathbf{x})\|_2}\right\rangle\right)\right] \\
&= m_{l+1}\mathbb{E}_{z\sim\mathcal{N}\left(0,\frac{\sigma_0^2}{c_{\phi,\sigma_0}}\right)}\left[\phi^2\left(\frac{\|\alpha^{(l)}(\mathbf{x})\|_2}{\sqrt{m_l}}z\right)\right] \qquad (5) \\
&= m_{l+1}\mathbb{E}_{z\sim\mathcal{N}(0,\sigma_0^2)}\left[\phi^2\left(\frac{\|\alpha^{(l)}(\mathbf{x})\|_2}{\sqrt{c_{\phi,\sigma_0}m_l}}z\right)\right] . \qquad (6)
\end{aligned}$$

Now, from Proposition A.2 with $\beta = \frac{\|\alpha^{(l)}(\mathbf{x})\|_2}{\sqrt{c_{\phi,\sigma_0}m_l}}$, we have

$$\begin{aligned}
&c_{\phi,\sigma_0} - \sigma_0^2\left|\frac{\|\alpha^l(\mathbf{x})\|_2^2}{c_{\phi,\sigma_0}m_l} - 1\right| \leq \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_0^2)}\left[\phi^2\left(\frac{\|\alpha^{(l)}(\mathbf{x})\|_2}{\sqrt{c_{\phi,\sigma_0}m_l}}z\right)\right] \leq c_{\phi,\sigma_0} + \sigma_0^2\left|\frac{\|\alpha^l(\mathbf{x})\|_2^2}{c_{\phi,\sigma_0}m_l} - 1\right| \\
\overset{(a)}{\Rightarrow} \quad &c_{\phi,\sigma_0}\left(1 - \frac{\vartheta_0^2 h_C(l)}{2h_C(L)}\right) \leq \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_0^2)}\left[\phi^2\left(\frac{\|\alpha^{(l)}(\mathbf{x})\|_2}{\sqrt{c_{\phi,\sigma_0}m_l}}z\right)\right] \leq c_{\phi,\sigma_0}\left(1 + \frac{\vartheta_0^2 h_C(l)}{2h_C(L)}\right) , \qquad (7)
\end{aligned}$$

where (a) follows from (4).

Combining (6) and (7), we have

$$c_{\phi,\sigma_0}\left(1-\frac{\vartheta_0^2 h_C(l)}{2h_C(L)}\right)m_{l+1} \le \mathbb{E}_{W_0^{(l+1)}}\|\alpha^{(l+1)}(\mathbf{x})\|_2^2 \le c_{\phi,\sigma_0}\left(1+\frac{\vartheta_0^2 h_C(l)}{2h_C(L)}\right)m_{l+1}\,. \tag{8}$$

Let $\kappa := \|(\alpha_j^{(l+1)}(\mathbf{x}))^2\|_{\psi_1}$. Then, we have $\kappa = \|(\alpha_j^{(l+1)}(\mathbf{x}))\|_{\psi_2}^2 \overset{(a)}{\le} \frac{4\bar{c}\sigma_0^2}{m_l c_{\phi,\sigma_0}}\|\alpha^{(l)}(\mathbf{x})\|_2^2 \overset{(b)}{\le} 6\bar{c}\sigma_0^2 = \tilde{O}(1)$, where (a) follows from a similar procedure to Lemma A.1 where $\bar{c} > 0$ is some absolute constant, and (b) follows from (4). Conditioned on $\{W_0^{(l')}, l' \in [L]\}$, since $\alpha_j^{(l+1)}(\mathbf{x})$, $j \in [m_{l+1}]$, are independent, from Bernstein's inequality we have

$$\mathbb{P}\left(\left|\sum_{j=1}^{m_{l+1}}(\alpha_j^{(l+1)}(\mathbf{x}))^2 - \mathbb{E}_{W_0^{(l+1)}}\|\alpha^{(l+1)}(\mathbf{x})\|_2^2\right| \ge t\right) \le 2\exp\left[-c\min\left(\frac{t^2}{m_{l+1}\kappa^2}, \frac{t}{\kappa}\right)\right]$$

for some absolute constant $c > 0$. Then, by union bound

$$\mathbb{P}\left(\max_{i\in[n]}\left|\sum_{j=1}^{m_{l+1}}(\alpha_j^{(l+1)}(\mathbf{x}_i))^2 - \mathbb{E}_{W_0^{(l+1)}}\|\alpha^{(l+1)}(\mathbf{x})\|_2^2\right| \ge t\right) \le 2n\exp\left[-c\min\left(\frac{t^2}{m_{l+1}\kappa^2}, \frac{t}{\kappa}\right)\right]$$

Choosing $t = \frac{c_{\phi,\sigma_0}}{2h_C(L)}\frac{\kappa}{\min(c,\sqrt{c})}(m_{l+1})^{3/4}(\log m_{l+1} + \log n) \le \frac{c_{\phi,\sigma_0}}{2h_C(L)}m_{l+1}$, we have

$$\begin{aligned}
&\mathbb{E}_{W_0^{(l+1)}}\|\alpha^{(l+1)}(\mathbf{x})\|_2^2 - \frac{c_{\phi,\sigma_0}}{2h_C(L)}m_{l+1} \le \min_{i\in[n]}\|\alpha^{(l+1)}(\mathbf{x}_i)\|_2^2 \le \max_{i\in[n]}\|\alpha^{(l+1)}(\mathbf{x}_i)\|_2^2 \le E_{W_0^{(l+1)}}\|\alpha^{(l+1)}(\mathbf{x})\|_2^2 + \frac{c_{\phi,\sigma_0}}{2h_C(L)}m_{l+1}\\
&\overset{(a)}{\Rightarrow}\quad c_{\phi,\sigma_0}\left(1-\frac{1+\vartheta_0^2 h_C(l)}{2h_C(L)}\right)m_{l+1} \le \min_{i\in[n]}\|\alpha^{(l+1)}(\mathbf{x}_i)\|_2^2 \le \max_{i\in[n]}\|\alpha^{(l+1)}(\mathbf{x}_i)\|_2^2 \le c_{\phi,\sigma_0}\left(1-\frac{1+\vartheta_0^2 h_C(l)}{2h_C(L)}\right)m_{l+1}\\
&\overset{(b)}{\Rightarrow}\quad c_{\phi,\sigma_0}\left(1-\frac{h_C(l+1)}{2h_C(L)}\right)m_{l+1} \le \min_{i\in[n]}\|\alpha^{(l+1)}(\mathbf{x}_i)\|_2^2 \le \max_{i\in[n]}\|\alpha^{(l+1)}(\mathbf{x}_i)\|_2^2 \le c_{\phi,\sigma_0}\left(1-\frac{h_C(l+1)}{2h_C(L)}\right)m_{l+1}\,,
\end{aligned}$$

where (a) follows from (8) and (b) follows since $h_C(l+1) = 1 + \vartheta_0^2 h_C(l)$, with probability at least

$$1 - 2n\exp\left[-\min\left(\frac{c_{\phi,\sigma_0}^2 m_{l+1}^{1/2}}{8h_C(L)^2}, \frac{c_{\phi,\sigma_0}m_{l+1}^{3/4}}{4h_C(L)}\right)2(\log m_{l+1} + \log n)\right] \ge 1 - \frac{2}{m_l^2}\,.$$

Applying union bound over all layers completes the proof. □

Proposition A.2. *Let $c_{\phi,\sigma_0} := E_{z\sim\mathcal{N}(0,\sigma_0^2)}[\phi^2(z)]$. Then,*

$$c_{\phi,\sigma_0} - \sigma_0^2|\beta^2 - 1| \le \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_0^2)}[\phi^2(\beta z)] \le c_{\phi,\sigma_0} + \sigma_0^2|\beta^2 - 1|\,.$$

Proof. We have

$$\begin{aligned}
\left|\mathbb{E}_{z\sim\mathcal{N}(0,\sigma_0^2)}[\phi^2(\beta z)] - E_{z\sim\mathcal{N}(0,\sigma_0^2)}[\phi^2(z)]\right| &\le \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_0^2)}[|\phi^2(\beta z) - \phi^2(z)|]\\
&= \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_0^2)}[|\phi(\beta z) - \phi(z)||\phi(\beta z) + \phi(z)|]\\
&\overset{(a)}{\le} |\beta - 1|\mathbb{E}_{z\sim\mathcal{N}(0,\sigma_0^2)}[|z|(|\phi(\beta z)| + |\phi(z)|)]\\
&\overset{(b)}{\le} |\beta - 1|\mathbb{E}_{z\sim\mathcal{N}(0,\sigma_0^2)}[|z|(\beta+1)|z|)]\\
&\le |\beta^2 - 1|\mathbb{E}_{z\sim\mathcal{N}(0,\sigma_0^2)}[|z|^2]\\
&= \sigma_0^2|\beta^2 - 1|\,,
\end{aligned}$$

where (a) and (b) follows from the 1-Lipschitzness of ϕ. As a result, we have

$$c_{\phi,\sigma_0} - \sigma_0^2|\beta^2 - 1| \le \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\phi^2(\beta z)] \le c_{\phi,\sigma_0} + \sigma_0^2|\beta^2 - 1|\,.$$

This completes the proof. □

A.3 PROOF OF LEMMA 4.3

We start with a specific consequence of the Schur product theorem (Oymak and Soltanolkotabi, 2020, Lemma 6.5) applied to r-th order Hadamard product of positive definitive matrices.

Proposition A.3. *Let $B = AA^\top$ where $A \in \mathbb{R}^{n\times p}$. Let $b_0 = \min_{i\in[n]} B_{ii}$. Then, for any $r \geq 1$ $\lambda_{\min}((AA^\top)^{\odot r}) \geq b_0^{r-1}\lambda_{\min}(AA^\top)$.*

Proof. Recall that for PSD matrices P, Q, it holds that $\lambda_{\min}(P \odot Q) \geq \min_{i\in[n]} Q_{ii} \cdot \lambda_{\min}(P)$. Further, note that $B_{ii} \geq 0$ and $b_0 \geq 0$ by construction. Then,

$$\lambda_{\min}((AA^\top)^{\odot r}) = \lambda_{\min}(B^{\odot(r-1)} \odot B) \geq \min_{i\in[n]}(B^{\odot(r-1)})_{ii} \cdot \lambda_{\min}(B) \leq b_0^{r-1}\lambda_{\min}(AA^\top)\,.$$

That completes the proof. □

Now, we are ready to prove Lemma 4.3.

Proof of Lemma 4.3. For convenience, let

$$\lambda_{l+1} := \lambda_{\min}\left(\mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_{m_l},\sigma^2\mathbb{I}_{m_l})}\left[\phi\left(\frac{1}{\sqrt{m_l}}A^{(l)}g\right)\phi\left(\frac{1}{\sqrt{m_l}}(A^{(l)}g)^\top\right)\right]\right)$$

Let $U_l \in \mathbb{R}^{n\times m_l}$ have ith row $U_{l,i:} = \frac{\alpha^{(l)}(\mathbf{x}_i)}{\|\alpha^{(l)}(\mathbf{x}_i)\|_2}$, so that U_l is a row normalized version of $A^{(l)}$. Let $C_l = \mathrm{diag}(c_{l,i})$ where $c_{l,i} = \frac{\|\alpha^{(l)}(\mathbf{x}_i)\|_2}{\sqrt{m_l}}$. Note that $\frac{1}{\sqrt{m_l}}A^{(l)} = C_lU_l$. Further, from Lemma 4.2, $\min_{i,l} c_{l,i} \geq \sqrt{\frac{c_{\phi,\sigma_0}}{2}}$ and $\max_{i,l} c_{l,i} \leq \sqrt{\frac{3c_{\phi,\sigma_0}}{2}}$ with probability at least $1 - 2n\sum_{l=1}^L \frac{1}{m_l}$. Let $M_r^{(l)}(\phi) = \mathrm{diag}\left(\mu_r^{[c_i^2\sigma^2]}(\phi)\right)$, and let $(\mu_{r,0}^{(l)})^2 = \min_{i\in[n]}\left(\mu_r^{[c_i^2\sigma^2]}(\phi)\right)^2$. Then, for any integer $r > 0$, we have

$$\begin{aligned}
\lambda_{l+1} &= \lambda_{\min}\left(\mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_{m_l},\sigma^2\mathbb{I}_{m_l})}\left[\phi\left(C_lU_lg\right)\phi\left(C_lU_lg\right)^\top\right]\right)\\
&\overset{(a)}{\geq} \sigma^{6r}\left(\frac{c_{\phi,\sigma_0}}{2}\right)^{3r}\lambda_{\min}\left((M_r^{(l)}(\phi)(U_l)^{\star r})(M_r^{(l)}(\phi)(U_l)^{\star r})^\top\right)\\
&\overset{(b)}{\geq} (\mu_{r,0}^{(l)})^2\sigma^{6r}\left(\frac{c_{\phi,\sigma_0}}{2}\right)^{3r}\lambda_{\min}\left(((U_l)^{\star r})((U_l)^{\star r})^\top\right)\\
&\geq (\mu_{r,0}^{(l)})^2\sigma^{6r}\left(\frac{c_{\phi,\sigma_0}}{2}\right)^{3r}\lambda_{\min}\left((U_lU_l^\top)^{\odot r}\right)\\
&\overset{(c)}{\geq} (\mu_{r,0}^{(l)})^2\sigma^{6r}\left(\frac{c_{\phi,\sigma_0}}{2}\right)^{3r}\lambda_{\min}\left(U_lU_l^\top\right)\\
&= (\mu_{r,0}^{(l)})^2\sigma^{6r}\left(\frac{c_{\phi,\sigma_0}}{2}\right)^{3r}\frac{1}{m_l}\lambda_{\min}\left(C_l^{-1}A^{(l)}(A^{(l)})^\top C_l^{-1}\right)\\
&\geq (\mu_{r,0}^{(l)})^2\sigma^{6r}\left(\frac{c_{\phi,\sigma_0}}{2}\right)^{3r}\frac{2}{3c_{\phi,\sigma_0}m_l}\lambda_{\min}\left(A^{(l)}(A^{(l)})^\top\right)\\
&\overset{(d)}{\geq} (\mu_{r,0}^{(l)})^2\sigma^{6r}\left(\frac{c_{\phi,\sigma_0}}{2}\right)^{3r}\frac{1}{6c_{\phi,\sigma_0}}\lambda_l\,,
\end{aligned}$$

where (a) follows from Lemma A.3,

$$\begin{aligned}
\mathbb{E}_{\tilde{g}\sim\mathcal{N}(\mathbf{0}_{m_l},\sigma^2\mathbb{I}_{m_l})}\left[\phi(C_lU_lg)\phi(C_lU_lg)^\top\right] &\succeq \sum_{r'=0}^{\infty}\sigma^{6r'}(\min_i c_{l,i})^{6r'}(M_r^{(l)}(\phi)(U_l)^{\star r'})(M_r^{(l)}(\phi)(U_l)^{\star r'})^\top\\
&\succeq \sigma^{6r}\left(\frac{c_{\phi,\sigma_0}}{2}\right)^{3r}(M_r^{(l)}(\phi)(U_l)^{\star r})(M_r^{(l)}(\phi)(U_l)^{\star r})^\top
\end{aligned}$$

for any $r > 0$; (b) follows since for a diagonal matrix M with $\mu_0^2 = \min_{i\in[n]} M_{ii}^2$ and a compatible matrix U,

$$\begin{aligned}
\inf_{v:\|v\|_2=1} v^\top(MU)(MU)^\top v = \inf_{v:\|v\|_2=1} v^\top M(UU^\top)M^\top v &\geq \inf_{v:\|v\|_2=1} v^\top MM^\top v \inf_{w:\|w\|_2=1} w^\top UU^\top w\\
&\geq \mu_0^2\lambda_{\min}(UU^\top)\,;
\end{aligned}$$

(c) follows from Proposition A.3; and (d) follows from Lemma 4.1.

Proceeding recursively, using $\sigma^2 = \nu_0^2 = \frac{\sigma_0^2}{c_{\phi,\sigma_0}}$, we have

$$\begin{aligned}
\lambda_{l+1} &\geq \frac{(\mu_{r,0}^{(l)})^2}{6c_{\phi,\sigma_0}} \sigma^{6r} \left(\frac{c_{\phi,\sigma_0}}{2}\right)^{3r} \lambda_l \\
&\geq \frac{(\mu_{r,0}^{(l)})^2}{6c_{\phi,\sigma_0}} \left(\frac{\sigma_0^2}{2}\right)^{3r} \lambda_l \\
&\geq \left(\frac{(\mu_{r,0}^{(l)})^2}{6c_{\phi,\sigma_0}}\right)^2 \left(\frac{\sigma_0^2}{2}\right)^{6r} \lambda_{l-1} \\
&\geq \left(\frac{(\mu_{r,0}^{(l)})^2}{6c_{\phi,\sigma_0}}\right)^l \left(\frac{\sigma_0^2}{2}\right)^{3rl} \lambda_1
\end{aligned}$$

That completes the proof. □

A.4 BACKGROUND ON HERMITE POLYNOMIALS AND HERMITE SERIES EXPANSIONS

Let $L^2(\mathbb{R}, w(\mathbf{x}))$ denote the set of all functions $f : \mathbb{R} \to \mathbb{R}$ such that

$$\int_{-\infty}^{\infty} f^2(\mathbf{x}) w(\mathbf{x}) dx < \infty \,. \tag{9}$$

Probabilist's and Physicist's Hermite Polynomials. The normalized *probabilist's Hermite polynomials* are given by:

$$H_r(\mathbf{x}) = \frac{(-1)^r}{\sqrt{r!}} e^{\frac{x^2}{2}} \frac{d^r}{dx^r} e^{-\frac{x^2}{2}} \,. \tag{10}$$

The polynomials are orthogonal with respect to the weight function $w(\mathbf{x}) = e^{-\frac{x^2}{2}}$ in the sense that

$$\int_{-\infty}^{\infty} H_r(\mathbf{x}) H_{r'}(\mathbf{x}) w(\mathbf{x}) dx = \sqrt{2\pi} \delta_{rr'} \,, \tag{11}$$

where $\delta_{rr'} = 1$, if $r = r'$, and 0 otherwise, i.e., the Kronecker delta. The corresponding unnormalized probabilist's Hermite polynomials are given by $\bar{H}_r(\mathbf{x}) = \sqrt{r!} H_r(\mathbf{x})$.

The normalized physicist's Hermite polynomials are respectively

$$\tilde{H}_r(\mathbf{x}) = \frac{(-1)^r}{\sqrt{r!}} e^{x^2} \frac{d^r}{dx^r} e^{-x^2} \,. \tag{12}$$

The polynomials are orthogonal with respect to the weight function $\tilde{w}(\mathbf{x}) = e^{-x^2}$ in the sense that

$$\int_{-\infty}^{\infty} \tilde{H}_r(\mathbf{x}) \tilde{H}_{r'}(\mathbf{x}) \tilde{w}(\mathbf{x}) dx = \sqrt{2\pi} 2^r \delta_{rr'} \,, \tag{13}$$

where $\delta_{rr'}$ is the Kronecker delta. The corresponding unnormalized physicist's Hermite polynomials are given by $\bar{\tilde{H}}_r(\mathbf{x}) = \sqrt{r!} \tilde{H}_r(\mathbf{x})$.

Generalized Hermite Polynomials. Our analysis of potentially inhomogeneous activation functions will need the substantially more flexible notion of normalized *generalized Hermite polynomials* $H_r^{[q]}(\mathbf{x})$, for a given $q > 0$, which are orthogonal with respect to $w^{[q]}(\mathbf{x}) = \frac{1}{\sqrt{2\pi q}} e^{-x^2/2a}$, and are given by

$$H_r^{[q]}(\mathbf{x}) = \frac{(-1)^r}{\sqrt{r!}} e^{\frac{x^2}{2q}} \frac{d^r}{dx^r} e^{-\frac{x^2}{2q}} \,. \tag{14}$$

It is easy to see that $H_r^{[1]}(\mathbf{x}) = H_r(\mathbf{x})$, the probabilist's Hermite polynomial in (10), and $H_r^{[\frac{1}{2}]} = \tilde{H}_r(\mathbf{x})$, the physicist's Hermite polynomial in (12). Furthermore, the generalized Hermite polynomials can be written as scaled versions of probabilist's Hermite polynomials as

$$H_r^{[q]}(\mathbf{x}) = a^{\frac{r}{2}} H_r\left(\frac{x}{\sqrt{q}}\right) . \tag{15}$$

Hermite Series. The polynomials $\{H_r(\mathbf{x})\}_{r=0}^{\infty}$ form an orthonormal basis for $L^2\left(\mathbb{R}, \frac{e^{-x^2/2}}{\sqrt{2\pi}}\right)$ which is a Hilbert space with inner product

$$\langle \phi_1, \phi_2 \rangle = \int_{-\infty}^{\infty} \phi_1(\mathbf{x})\phi_2(\mathbf{x}) \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx . \tag{16}$$

Thus, any function in $L^2\left(\mathbb{R}, \frac{e^{-x^2/2}}{\sqrt{2\pi}}\right)$ can be represented as a Hermite series expansion

$$\phi(\mathbf{x}) = \sum_{r=0}^{\infty} \mu_r(\phi) H_r(\mathbf{x}) , \tag{17}$$

where $\mu_r(\phi)$ is the r-th Hermite coefficient given by

$$\mu_r(\phi) = \int_{-\infty}^{\infty} \phi(z) H_r(z) \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz . \tag{18}$$

Note that $\phi \in L^2\left(\mathbb{R}, \frac{e^{-x^2/2}}{\sqrt{2\pi}}\right)$ if and only if $\|\phi\|^2 = \langle \phi, \phi \rangle = \sum_{r=0}^{\infty} \mu_r^2(\phi) < \infty$.

For our analysis with inhomogeneous activation functions, we will need to use Hermite series expansions with generalized Hermite polynomials. The polynomials $\{H_r^{[q]}(\mathbf{x})\}_{r=0}^{\infty}$ form an orthonormal basis for $L^2\left(\mathbb{R}, \frac{e^{-x^2/2a}}{\sqrt{2\pi a}}\right)$ which is a Hilbert space with inner product

$$\langle \phi_1, \phi_2 \rangle = \int_{-\infty}^{\infty} \phi_1(\mathbf{x})\phi_2(\mathbf{x}) \frac{e^{-x^2/2q}}{\sqrt{2\pi q}} dx . \tag{19}$$

Any function in $L^2\left(\mathbb{R}, \frac{e^{-x^2/2q}}{\sqrt{2\pi q}}\right)$ can be represented as a Hermite series expansion:

$$\phi(\mathbf{x}) = \sum_{r=0}^{\infty} \mu_r^{[q]}(\phi) H_r^{[q]}(\mathbf{x}) , \tag{20}$$

where $\mu_r^{[q]}(\phi)$ is the r-th Hermite coefficient given by

$$\mu_r^{[q]}(\phi) = \int_{-\infty}^{\infty} \phi(z) H_r^{[q]}(z) \frac{e^{-z^2/2q}}{\sqrt{2\pi q}} dz . \tag{21}$$

Note that $\phi \in L^2\left(\mathbb{R}, \frac{e^{-x^2/2q}}{\sqrt{2\pi q}}\right)$ if and only if $\|\phi\|^2 = \langle \phi, \phi \rangle = \sum_{r=0}^{\infty} (\mu_r^{[q]}(\phi))^2 < \infty$.

A.5 EXPECTATION OF PRODUCT OF HERMITE POLYNOMIALS

Our NTK analysis for general activation functions, including inhomogeneous functions, depends on the following key result on expectation of product of Hermite polynomials. The equivalent prior analysis in (Oymak and Soltanolkotabi, 2020; Nguyen and Mondelli, 2020; Nguyen et al., 2021b) only works for homogeneous functions, and uses basic Hermite polynomials. Our general analysis instead uses generalized Hermite polynomials.

Lemma A.2. *Let $\mathbf{u}_x, \mathbf{u}_y \in \mathbb{R}^d$ be unit vectors, and let $c_x, c_y \in \mathbb{R}_{++}$ be positive constants. Then, for $r, r' = 0, 1, \ldots$ and $\delta_{rr'}$ denoting the Kronecker delta, we have*

$$\mathbb{E}_{\tilde{\mathbf{g}}\sim\mathcal{N}(\mathbf{0}_d,\sigma^2\mathbb{I}_d)}\left[H_r^{[c_x^2\sigma^2]}(c_x\langle\tilde{\mathbf{g}},\mathbf{u}_x\rangle)H_{r'}^{[c_y^2\sigma^2]}(c_y\langle\tilde{\mathbf{g}},\mathbf{u}_y\rangle)\right] = \sigma^{6r}c_x^{3r}c_y^{3r}\langle\mathbf{u}_x,\mathbf{u}_y\rangle^r\delta_{rr'}\ . \tag{22}$$

Proof. Let $\mathbf{g}\sim\mathcal{N}(\mathbf{0},\mathbb{I}_d)$ so that $\sigma\mathbf{g}$ is identically distributed as $\tilde{\mathbf{g}}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbb{I}_d)$, and consider any $s,t,\in\mathbb{R}$. Then,

$$\begin{aligned}
\mathbb{E}_{\tilde{\mathbf{g}}\sim\mathcal{N}(\mathbf{0}_d,\sigma^2\mathbb{I}_d)}\left[\exp(sc_x\langle\tilde{\mathbf{g}},\mathbf{u}_x\rangle + tc_y\langle\tilde{\mathbf{g}},\mathbf{u}_y\rangle)\right] &= \mathbb{E}_{\mathbf{g}\sim\mathcal{N}(\mathbf{0}_d,\mathbb{I}_d)}\left[\exp(s\sigma c_x\langle\mathbf{g},\mathbf{u}_x\rangle + t\sigma c_y\langle\mathbf{g},\mathbf{u}_y\rangle)\right]\\
&= \prod_{j=1}^d \mathbb{E}_{\mathbf{g}\sim\mathcal{N}(\mathbf{0}_d,\mathbb{I}_d)}\left[\exp(s\sigma c_x g_j u_{x,j} + t\sigma c_y g_j u_{y,j})\right]\\
&= \prod_{j=1}^d \exp\left(\frac{\sigma^2(sc_x u_{x,j} + tc_y u_{y,j})^2}{2}\right)\\
&= \exp\left(\frac{s^2\sigma^2c_x^2}{2}\|\mathbf{u}_x\|_2^2 + \frac{t^2\sigma^2c_y}{2}\|\mathbf{u}_y\|_2^2 + st\sigma^2c_xc_y\langle\mathbf{u}_x,\mathbf{u}_y\rangle\right)\ ,
\end{aligned}$$

so that, since $\|\mathbf{u}_x\|_2^2 = \|\mathbf{u}_y\|_2^2 = 1$, we have

$$\mathbb{E}_{\mathbf{g}\sim\mathcal{N}(\mathbf{0}_d,\mathbb{I}_d)}\left[\exp\left(s\sigma c_x\langle\mathbf{g},\mathbf{u}_x\rangle - \frac{s^2\sigma^2c_x^2}{2}\right)\exp\left(t\sigma c_y\langle\mathbf{g},\mathbf{u}_y\rangle - \frac{t^2\sigma^2c_y^2}{2}\right)\right] = \exp\left(st\sigma^2c_xc_y\langle\mathbf{u}_x,\mathbf{u}_y\rangle\right)\ . \tag{23}$$

We consider the functions $f, h : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f(s) = \exp\left(s\sigma c_x\langle\mathbf{g},\mathbf{u}_x\rangle - \frac{s^2\sigma^2c_x^2}{2}\right)\ , \qquad h(t) = \exp\left(t\sigma c_y\langle\mathbf{g},\mathbf{u}_y\rangle - \frac{t^2\sigma^2c_y^2}{2}\right)\ . \tag{24}$$

Consider the Taylor expansion of $f(s)$ with respect to $f(0)$ given by

$$f(s) = \sum_{r=0}^\infty f_r(0)\frac{s^r}{\sqrt{r!}}\ , \qquad \text{where} \qquad f_r(0) = \frac{1}{\sqrt{r!}}\frac{d^r}{ds^r}e^{s\sigma c_x\langle\mathbf{g},\mathbf{u}_x\rangle - \frac{s^2\sigma^2c_x^2}{2}}\Big|_{s=0}\ . \tag{25}$$

With $z = \langle\mathbf{g},\mathbf{u}_x\rangle$ and $\tilde{z} = \frac{z}{\sigma c_x}$, we have

$$\begin{aligned}
f_r(0) &= \frac{1}{\sqrt{r!}}\frac{d^r}{ds^r}e^{s\sigma c_x z - \frac{s^2\sigma^2c_x^2}{2}}\Big|_{s=0}\\
&= \frac{1}{\sqrt{r!}}\,e^{\frac{z^2}{2}}\frac{d^r}{ds^r}e^{-\frac{1}{2}(z-s\sigma c_x)^2}\Big|_{s=0}\\
&= \frac{1}{\sqrt{r!}}\,e^{\frac{z^2}{2}}\frac{d^r}{ds^r}e^{-\frac{\sigma^2c_x^2}{2}\left(\frac{z}{\sigma c_x}-s\right)^2}\Big|_{s=0}\\
&= \frac{1}{\sqrt{r!}}\,e^{\frac{\sigma^2c_x^2\tilde{z}^2}{2}}\frac{d^r}{ds^r}e^{-\frac{\sigma^2c_x^2}{2}(\tilde{z}-s)^2}\Big|_{s=0}\\
&\overset{(a)}{=} \frac{(-1)^r}{\sqrt{r!}}\,e^{\frac{\sigma^2c_x^2\tilde{z}^2}{2}}\frac{d^r}{d\tilde{z}^r}e^{-\frac{\sigma^2c_x^2}{2}(\tilde{z}-s)^2}\Big|_{s=0}\\
&= \frac{(-1)^r}{\sqrt{r!}}e^{\frac{\sigma^2c_x^2\tilde{z}^2}{2}}\frac{d^r}{d\tilde{z}^r}e^{-\frac{\sigma^2c_x^2\tilde{z}^2}{2}}\\
&\overset{(b)}{=} H_r^{\left[\frac{1}{\sigma^2c_x^2}\right]}(\tilde{z})\ ,
\end{aligned}$$

where (a) follows by the transport or advection equation $\frac{d}{ds}\psi(\tilde{z}-s) = (-1)\frac{d}{d\tilde{z}}\psi(\tilde{z}-s)$ and the equality of mixed partial derivatives for any sufficiently smooth function ψ, and (b) follows by definition of generalized Hermite polynomials in (14).

Now, note that

$$
\begin{aligned}
H_r^{\left[\frac{1}{\sigma^2 c_x^2}\right]}(\tilde{z}) &\overset{(a)}{=} \frac{1}{\sigma^r c_x^r} H_r(\sigma c_x \tilde{z}) \\
&= \frac{1}{\sigma^r c_x^r} H_r(z) \\
&= \frac{1}{\sigma^r c_x^r} H_r\left(\langle \mathbf{g}, \mathbf{u}_x \rangle\right) \\
&= \frac{1}{\sigma^r c_x^r} H_r\left(\frac{c_x}{\sigma c_x}\langle \tilde{\mathbf{g}}, \mathbf{u}_x \rangle\right) \\
&\overset{(b)}{=} \frac{1}{\sigma^{2r} c_x^r} H_r^{[c_x^2\sigma^2]}\left(c_x \langle \tilde{\mathbf{g}}, \mathbf{u}_x \rangle\right) \ .
\end{aligned}
$$

where (a) and (b) follow from (15). Thus,

$$
f(s) = \sum_{r=0}^{\infty} f_r(0) \frac{s^r}{\sqrt{r!}} \ , \qquad \text{where} \qquad f_r(0) = \frac{1}{\sigma^{2r} c_x^{2r}} H_r^{[c_x^2\sigma^2]}\left(c_x \langle \tilde{\mathbf{g}}, \mathbf{u}_x \rangle\right) \ . \tag{26}
$$

Similarly, considering the Taylor expansion of $h(t)$ with respect to $h(0)$, we have

$$
h(t) = \sum_{r'=0}^{\infty} h_{r'}(0) \frac{t^{r'}}{\sqrt{r'!}} \ , \qquad \text{where} \qquad h_{r'}(0) = \frac{1}{\sigma^{2r'} c_y^{2r'}} H_{r'}^{[c_y^2\sigma^2]}\left(c_y \langle \tilde{\mathbf{g}}, \mathbf{u}_y \rangle\right) \ . \tag{27}
$$

Then, from (23), we have

$$
\begin{aligned}
&\mathbb{E}_{\tilde{\mathbf{g}} \sim \mathcal{N}(\mathbf{0}_d, \sigma^2 \mathbb{I}_d)} \left[\left(\sum_{r=0}^{\infty} \frac{1}{\sigma^{2r} c_x^{2r}} H_r^{[c_x^2\sigma^2]}(c_x \langle \tilde{\mathbf{g}}, \mathbf{u}_x \rangle) \frac{s^r}{\sqrt{r!}} \right) \left(\sum_{r'=0}^{\infty} \frac{1}{\sigma^{2r'} c_y^{2r'}} H_{r'}^{[c_y^2\sigma^2]}\left(c_y \langle \tilde{\mathbf{g}}, \mathbf{u}_y \rangle\right) \frac{t^{r'}}{\sqrt{r'!}} \right) \right] \\
&\quad = \sum_{r=0}^{\infty} \frac{\sigma^{2r} c_x^r c_y^r \langle \mathbf{u}_x, \mathbf{u}_y \rangle^r}{r!} s^r t^r \ .
\end{aligned}
$$

Since the equality holds for arbitrary $s, t \in R$, equating coefficients of $s^r t^{r'}$ on both sides, we have

$$
\mathbb{E}_{\tilde{\mathbf{g}} \sim \mathcal{N}(\mathbf{0}_d, \sigma^2 \mathbb{I}_d)} \left[H_r^{[c_x^2\sigma^2]}(c_x \langle \tilde{\mathbf{g}}, \mathbf{u}_x \rangle) H_{r'}^{[c_y^2\sigma^2]}(c_y \langle \tilde{\mathbf{g}}, \mathbf{u}_y \rangle) \right] = \sigma^{6r} c_x^{3r} c_y^{3r} \langle \mathbf{u}_x, \mathbf{u}_y \rangle^r \delta_{rr'} \ , \tag{28}
$$

where $\delta_{rr'}$ is the Kronecker delta. That completes the proof. □

The following result is an important consequence of Lemma A.2.

Lemma A.3. *Let ϕ be an inhomogeneous activation function. Let $\mathbf{u}_x, \mathbf{u}_y \in \mathbb{R}^d$ be unit vectors and c_x, c_y be positive constants. Then, we have*

$$
\mathbb{E}_{\tilde{\mathbf{g}} \sim \mathcal{N}(\mathbf{0}_d, \sigma^2 \mathbb{I}_d)} \left[\phi(c_x \langle \tilde{\mathbf{g}}, \mathbf{u}_x \rangle) \phi(c_y \langle \tilde{\mathbf{g}}, \mathbf{u}_y \rangle) \right] = \sum_{r=0}^{\infty} \mu_r^{[c_x^2\sigma^2]}(\phi) \mu_r^{[c_y^2\sigma^2]}(\phi) \sigma^{6r} c_x^{3r} c_y^{3r} \langle \mathbf{u}_x, \mathbf{u}_y \rangle^r \ . \tag{29}
$$

Further, let $U = [\mathbf{u}_1, \cdots, \mathbf{u}_n]^\top \in \mathbb{R}^{n \times m}$ be such that $\|\mathbf{u}_i\|_2 = 1, i \in [n]$. Let $C = \mathrm{diag}(c_i) \in \mathbb{R}^{n \times n}, c_i > 0$, and $c_0 = \min_{i \in [n]} c_i > 0$. Let $M_r(\phi) = \mathrm{diag}\left(\mu_r^{[c_i^2\sigma^2]}(\phi)\right)$. Then,

$$
\mathbb{E}_{\tilde{\mathbf{g}} \sim \mathcal{N}(\mathbf{0}_d, \sigma^2 \mathbb{I}_d)} \left[\phi(CU\tilde{\mathbf{g}}) \phi(CU\tilde{\mathbf{g}})^\top \right] \succeq \sum_{r=0}^{\infty} \sigma^{6r} c_0^{6r} (M_r(\phi) U^{\star r}) (M_r(\phi) U^{\star r})^\top \ , \tag{30}
$$

where $U^{\star r} \in \mathbb{R}^{n \times m^r}$ is such that the ith row $U_{i,:}^{\star r} = (\mathbf{u}_i^{\odot r})^\top$, i.e., r times Kronecker product of $\mathbf{u}_i$ with itself.

Proof. Consider the generalized Hermite series expansion of $\phi(\cdot)$ in terms of the generalized Hermite functions $H_n^{[\sigma^2]}$:

$$\phi(c_x\langle\tilde{\mathbf{g}},\mathbf{u}_x\rangle)=\sum_{r=0}^{\infty}\mu_r^{[c_x^2\sigma^2]}(\phi)H_r^{[c_x^2\sigma^2]}(c_x\langle\tilde{\mathbf{g}},\mathbf{u}_x\rangle)\;. \tag{31}$$

Then, we have

$$\begin{aligned}
&\mathbb{E}_{\tilde{\mathbf{g}}\sim\mathcal{N}(\mathbf{0}_d,\sigma^2\mathbb{I}_d)}\left[\phi(c_x\langle\tilde{\mathbf{g}},\mathbf{u}_x\rangle)\phi(c_y\langle\tilde{\mathbf{g}},\mathbf{u}_y\rangle)\right]\\
&\quad=\mathbb{E}_{\tilde{\mathbf{g}}\sim\mathcal{N}(\mathbf{0}_d,\sigma^2\mathbb{I}_d)}\left[\left(\sum_{r=0}^{\infty}\mu_r^{[c_x^2\sigma^2]}(\phi)H_r^{[c_x^2\sigma^2]}(c\langle\tilde{\mathbf{g}},\mathbf{u}_x\rangle)\right)\left(\sum_{r'=0}^{\infty}\mu_{r'}^{[c_y^2\sigma^2]}(\phi)H_{r'}^{[c_y^2\sigma^2]}(c\langle\tilde{\mathbf{g}},\mathbf{u}_y\rangle)\right)\right]\\
&\quad=\sum_{r,r'=0}^{\infty}\mu_r^{[c_x^2\sigma^2]}(\phi)\mu_{r'}^{[c_y^2\sigma^2]}(\phi)\mathbb{E}_{\tilde{\mathbf{g}}\sim\mathcal{N}(\mathbf{0}_d,\sigma^2\mathbb{I}_d)}\left[H_r^{[c_x^2\sigma^2]}(c_x\langle\tilde{\mathbf{g}},\mathbf{u}_x\rangle)H_{r'}^{[c_y^2\sigma^2]}(c_y\langle\tilde{\mathbf{g}},\mathbf{u}_y\rangle)\right]\\
&\quad\overset{(a)}{=}\sum_{r=0}^{\infty}\mu_r^{[c_x^2\sigma^2]}(\phi)\mu_r^{[c_y^2\sigma^2]}(\phi)\sigma^{6r}c_x^{3r}c_y^{3r}\langle\mathbf{u}_x,\mathbf{u}_y\rangle^r\;.
\end{aligned}$$

where (a) follows from Lemma A.2.

The matrix case result follows by noting that for any $i,j\in[n]$, $c_i^{3r}c_j^{3r}\geq c_0^{6r}$, $\langle\mathbf{u}_i,\mathbf{u}_j\rangle^r=\langle\mathbf{u}_i^{\otimes r},\mathbf{u}_j^{\otimes r}\rangle$, and $\mu_r^{[c_i^2\sigma^2]}(\phi)$ form the diagonal elements of $M_r(\phi)$. That completes the proof. □

A.6 ADDITIONAL REMARKS ON λ_1

The main result in Theorem 4.1 establishes $\lambda_{\min}(K_{\text{ntk}}(\cdot;\theta_0))\geq c_0\lambda_1$ where

$$\lambda_1=\lambda_{\min}\left(\mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_d,\sigma^2\mathbb{I}_d)}\left[\phi\left(\frac{1}{\sqrt{d}}Xg\right)\phi\left(\frac{1}{\sqrt{d}}Xg\right)^\top\right]\right)\;,$$

where $\sigma^2=\nu_0^2=\frac{\sigma_0^2}{c_{\phi,\sigma_0}}$. We made some high level informal remarks on why and when $\lambda_1>0$ in Remark 4.2. Such results have been studied in the recent literature (Du et al., 2019; Zou et al., 2020; Allen-Zhu et al., 2019; Oymak and Soltanolkotabi, 2020; Nguyen et al., 2021b). We provide additional details on the topic.

Related to assumptions in Du et al. (2019), the simplest analysis comes from assuming $\lambda_{\min}(XX^\top)=c_{\phi,\sigma_0}d\lambda_0>0$ for some positive constant λ_0, where the scaling is simply because $\|\mathbf{x}_i\|_2^2=c_{\phi,\sigma_0}d$. With $\bar{X}:=\frac{1}{\sqrt{dc_{\phi,\sigma_0}}}X$ so that rows of $\bar{X}$ satisfy $\|\bar{\mathbf{x}}_i\|_2=1$ and $\lambda_{\min}(\bar{X}\bar{X}^\top)=\lambda_0>0$. Let $C_0=\text{diag}(c_{0,i})$ where $c_{0,i}=\sqrt{c_{\phi,\sigma_0}}$. Note that $\frac{1}{\sqrt{d}}X=C_0\bar{X}$. Let $M_r^{(0)}(\phi)=\mu_r^{[\sigma_0^2]}(\phi)\,\text{diag}(1)$, and let $(\mu_{r,0}^{(0)})^2=\left(\mu_r^{[\sigma_0^2]}(\phi)\right)^2$. From Lemma A.3, for any integer $r>0$, we have

$$\begin{aligned}
\lambda_1&=\lambda_{\min}\left(\mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_d,\sigma^2\mathbb{I}_d)}\left[\phi\left(\frac{1}{\sqrt{d}}Xg\right)\phi\left(\frac{1}{\sqrt{d}}Xg\right)^\top\right]\right)\\
&\geq\sigma^{6r}\left(c_{\phi,\sigma_0}\right)^{3r}\lambda_{\min}\left((M_r^{(0)}(\phi)(\bar{X})^{\star r})(M_r^{(0)}(\phi)(\bar{X})^{\star r})^\top\right)\\
&\geq(\mu_{r,0}^{(0)})^2\sigma^{6r}\left(c_{\phi,\sigma_0}\right)^{3r}\lambda_{\min}\left((\bar{X})^{\star r})(\bar{X})^{\star r})^\top\right)\\
&=(\mu_{r,0}^{(l)})^2\sigma^{6r}\left(c_{\phi,\sigma_0}\right)^{3r}\lambda_{\min}\left(\bar{X}\bar{X}^\top)^{\odot r}\right)\\
&\geq(\mu_{r,0}^{(l)})^2\sigma^{6r}\left(c_{\phi,\sigma_0}\right)^{3r}\lambda_{\min}\left(\bar{X}\bar{X}^\top\right)\\
&\geq(\mu_{r,0}^{(l)})^2\sigma^{6r}\left(c_{\phi,\sigma_0}\right)^{3r}\lambda_0\;,
\end{aligned}$$

which gives the desired result.

λ_1 can also be lower bounded by making assumptions on the activation function ϕ, e.g., the separability and/or the

distribution of x_i (**?**Allen-Zhu et al., 2019; Oymak and Soltanolkotabi, 2020; Nguyen et al., 2021b). For any unit vector v,

$$\begin{aligned}\lambda_1(v) &:= v^\top \mathbb{E}_{g\sim\mathcal{N}\left(\mathbf{0}_d, \frac{\sigma_0^2}{c_{\phi,\sigma_0}}\mathbb{I}_d\right)}[\phi(\sqrt{c_{\phi,\sigma_0}}\bar{X}g)\phi(\sqrt{c_{\phi,\sigma_0}}\bar{X}g)^\top]v \\ &= v^\top \mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_d,\sigma_0^2\mathbb{I}_d)}[\phi(\bar{X}g)\phi(\bar{X}g)^\top]v \\ &= \mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_d,\sigma_0^2\mathbb{I}_d)}[\|\phi(\bar{X}g)^\top v\|_2^2]\,.\end{aligned}$$

Note that with $\tilde{g} = \bar{X}g$, it suffices to show $\mathbb{E}_{\tilde{g}}[\langle\phi(\tilde{g}), v\rangle^2] = \mathbb{E}_{Z=\langle\phi(\tilde{g}),v\rangle}[Z^2] \geq \chi_0 > 0$, for some uniform positive constant χ_0 since $\lambda_1 = \inf_v \lambda_1(v)$. For any $c > 0$, by Markov's inequality, we have

$$\begin{aligned}&\mathbb{P}(\|\phi(\bar{X}g)^\top v\|_2 \geq c) = \mathbb{P}(\|\phi(\bar{X}g)^\top v\|_2^2 \geq c^2) \leq \frac{\mathbb{E}[\|\phi(\bar{X}g)^\top v\|_2^2]}{c^2} \\ \Rightarrow \quad &\mathbb{E}[\|\phi(\bar{X}g)^\top v\|_2^2 \geq c^2\mathbb{P}(\|\phi(\bar{X}g)^\top v\|_2 \geq c)\,.\end{aligned}$$

Thus, the problem boils down to lower bounding $\mathbb{P}(\|\phi(\bar{X}g)^\top v\|_2 \geq c)$ for a suitable choice of c, or, more conveniently $\mathbb{P}(\|\phi(\bar{X}g)^\top v\|_2 \geq c\|v\|_\infty)$ and using $\|v\|_\infty \geq \frac{1}{\sqrt{n}}$. Proceeding further rigorously needs specific assumptions on the activation ϕ, as has been done in recent related work (Oymak and Soltanolkotabi, 2020; Allen-Zhu et al., 2019; **?**).