Frequency Bias in Neural Networks for Input of Non-Uniform Density

Supplementary Material

Ronen Basri [1] **Meirav Galun** [1] **Amnon Geifman** [1] **David Jacobs** [2] **Yoni Kasten** [1] **Shira Kritchman** [1]

A. Eigenfunctions of NTK for a two layer-network for data drawn from a piecewise constant distribution

Lemma 1. *Let $p(x)$ be a piecewise constant density function on $\mathbb{S}^1$. Then the eigenfunctions in Eq. (9) in the paper satisfy the following ordinary differential equation*

$$f''(x) = -\frac{p(x)}{\pi\lambda} f(x). \tag{1}$$

Proof. Combining Eqs. (9) and (10) in the paper we have

$$\int_{x-\pi}^{x+\pi} (1+\cos(z-x))(\pi - |z-x|) f(z) p(z) dz = 4\pi\lambda f(x) \tag{2}$$

Below we take six derivatives of (2) with respect to x. We use parenthesized superscripts $f^{(n)}(x)$ to denote the n^{th} derivative of f at x. First derivative

$$\begin{aligned} 4\pi\lambda f^{(1)}(x) &= -\int_{x-\pi}^{x} \left(1+\cos(z-x) - (\pi+z-x)\sin(z-x)\right) f(z)p(z)dz \\ &\quad + \int_{x}^{x+\pi} \left(1+\cos(z-x) + (\pi-z+x)\sin(z-x)\right) f(z)p(z)dz \end{aligned}$$

Second derivative

$$\begin{aligned} 4\pi\lambda f^{(2)}(x) + 4f(x)p(x) &= -\int_{x-\pi}^{x} \left(2\sin(z-x) + (\pi+z-x)\cos(z-x)\right) f(z)p(z)dz \\ &\quad + \int_{x}^{x+\pi} \left(2\sin(z-x) - (\pi-z+x)\cos(z-x)\right) f(z)p(z)dz \end{aligned}$$

Adding this to (2)

$$\begin{aligned} 4\pi\lambda f^{(2)}(x) + 4f(x)p(x) + 4\pi\lambda f(x) &= \int_{x-\pi}^{x} \left(\pi+z-x-2\sin(z-x)\right) f(z)p(z)dz \\ &\quad + \int_{x}^{x+\pi} \left(\pi-z+x+2\sin(z-x)\right) f(z)p(z)dz \end{aligned} \tag{3}$$

Third derivative

$$\begin{aligned} &4\pi\lambda f^{(3)}(x) + 4\pi\lambda f^{(1)}(x) + 4f^{(1)}(x)p(x) + 4f(x)p^{(1)}(x) = \\ &\int_{x-\pi}^{x} \left(2\cos(z-x) - 1\right) f(z)p(z)dz - \int_{x}^{x+\pi} \left(2\ \cos(z-x) - 1\right) f(z)p(z)dz \end{aligned}$$

[1] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel [2] Department of Computer Science, Univeristy of Maryland, College Park, MD, USA. Correspondence to: Ronen Basri <ronen.basri@weizmann.ac.il>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

Fourth derivative

$$\begin{aligned}
&4\pi\lambda f^{(4)}(x) + 4\pi\lambda f^{(2)}(x) + 4f^{(2)}(x)p(x) + 8f^{(1)}(x)p^{(1)}(x) + 4f(x)p^{(2)}(x) - 2f(x)p(x) = \\
&3f(x-\pi)p(x-\pi) + 3f(x+\pi)p(x+\pi) - \int_{x}^{x+\pi} 2\sin(z-x)f(z)p(z)dz + \int_{x-\pi}^{x} 2\sin(z-x)f(z)p(z)dz
\end{aligned}$$

Adding this to (3)

$$\begin{aligned}
&4\pi\lambda f^{(4)}(x) + 8\pi\lambda f^{(2)}(x) + 4\pi\lambda f(x) + 2f(x)p(x) + 4p(x)f^{(2)}(x) + 8f^{(1)}(x)p^{(1)}(x) + 4f(x)p^{(2)}(x) = \\
&3f(x-\pi)p(x-\pi) + 3f(x+\pi)p(x+\pi) + \int_{x}^{x+\pi} (\pi - z + x)f(z)p(z)dz + \int_{x-\pi}^{x} (\pi + z - x)f(z)p(z)dz
\end{aligned}$$

Fifth derivative

$$\begin{aligned}
&4\pi\lambda f^{(5)}(x) + 8\pi\lambda f^{(3)}\,(x) + 4\pi\lambda f^{(1)}(x) + 4f^{(3)}(x)p(x) + f^{(2)}(x)p^{(1)}(x) + 12f^{(1)}(x) + p^{(2)}(x) \\
&+2f^{(1)}(x)p(x) + 4f(x)p^{(3)}(x) = -2f(x)p^{(2)}(x) + 3f^{(1)}(x-\pi)p(x-\pi) + 3f(x-\pi)p^{(1)}(x-\pi) \\
&+3f^{(1)}(x+\pi)p(x+\pi) + 3f(x+\pi)p^{(1)}(x+\pi) - \int_{x-\pi}^{x} f(z)p(z)dz + \int_{x}^{x+\pi} f(z)p(z)dz
\end{aligned}$$

Sixth derivative

$$\begin{aligned}
&4\pi\lambda f^{(6)}(x) + 8\pi\lambda f^{(4)}(x) + 4\pi\lambda f^{(2)}(x) = 3f^{(2)}(x+\pi)p\,(x+\pi) + 3p^{(2)}(x+\pi)f(x+\pi) \\
&+6f^{(1)}(x+\pi)p^{(1)}(x+\pi) - 2f(x)p(x) + f(x-\pi)p(x-\pi) - 4f(x)p^{(4)}(x) - 4p(x)f^{(4)}\,(x) \\
&-2f(x)p^{(2)}(x) - 2\,p(x)\,f^{(2)}(x) + f(x+\pi)p(x+\pi) + 6f^{(1)}(x-\pi)p^{(1)}(x-\pi) + 3f^{(2)}(x-\pi)p(x-\pi) \\
&+3p^{(2)}(x-\pi)f(x-\pi) - 16f^{(1)}(x)p^{(3)}(x) - 16f^{(3)}(x)p^{(1)}(x) - 24p^{(2)}(x)f^{(2)}(x) - 4f^{(1)}(x)p^{(1)}(x)
\end{aligned}$$

Next, we simplify and rearrange. We omit dependence on x, note that $f(x-\pi) = f(x+\pi)$ and $p(x-\pi) = p(x+\pi)$ and respectively denote them by $\bar{f}$ and $\bar{p}$.

$$\begin{aligned}
&2\pi\lambda f^{(6)} + 2(p + 2\pi\lambda)f^{(4)} + 8p^{(1)}f^{(3)} + (p + 12p^{(2)} + 2\pi\lambda)f^{(2)} + \\
&2(p^{(1)} + 4p^{(3)})f^{(1)} + (p + p^{(2)} + 2p^{(4)})f = (\bar{p} + 3\bar{p}^{(2)})\bar{f} + 6\bar{p}^{(1)}\bar{f}^{(1)} + 3\bar{p}\bar{f}^{(2)}
\end{aligned}$$

Assume next that $p(x)$ is constant around x and $x - \pi$, so its derivatives at these points vanish. Then,

$$2\pi\lambda f^{(6)} + (2p + 4\pi\lambda)f^{(4)} + (p + 2\pi\lambda)f^{(2)} + pf = \bar{p}\bar{f} + 3\bar{p}\bar{f}^{(2)}$$

We next make the assumption that $p(x)$ has a period of π (so $p = \bar{p}$) in which case $f(x+\pi) = -f(x)$ (i.e., $\bar{f} = -f$). These assumptions will be removed later. With these assumptions we have

$$2\pi\lambda f^{(6)} + (2p + 4\pi\lambda)f^{(4)} + (4p + 2\pi\lambda)f^{(2)} + 2pf = 0$$

It can be readily verified that this equation is solved by (1).

Finally, if $p(x)$ does not have a period of π we can preprocess the data in a straightforward way to make p have a period of π (by mapping the interval $[0, 4\pi)$ to $[0, 2\pi)$) without changing the function that needs to be learned. □

B. The amplitudes of the eigenfunctions in different regions

In this section for the NTK of a 2-layer network for which only the first layer is trained we compute bounds on the amplitudes of its eigenfunctions. We first bound the ratios between the amplitudes in two neighboring regions, and use this in the following section to bound the amplitude in any one region.

B.1. Ratios between the amplitudes of neighboring regions

If $p(x) = p_j$ is constant in each region $R_j \subseteq \mathbb{S}^1$, $1 \leq j \leq l$, then the eigenfunction or order q $f_q(x)$ for $x \in R_j$ can be written as

$$f_q(x) = a_j \cos\left(\frac{q\sqrt{p_j}x}{Z} + b_j\right)$$

where $a_j \geq 0$. In this part we characterize the amplitudes the different regions a_j for $j = 1, ..., l$.

We notice that the eigenfunctions appear to be continuous and differentiable. Without loss of generality, assume that the boundary between region j to region $j+1$ happens at $x = 0$. Then the eigenfunction in the vicinity of 0 is defined as follows:

$$f_q(x) = \begin{cases} a_j \cos(q\frac{\sqrt{p_j}}{Z}x + b_j) & x \leq 0 \\ a_{j+1} \cos(q\frac{\sqrt{p_{j+1}}}{Z}x + b_{j+1}) & x \geq 0 \end{cases}$$

Continuity at $x = 0$ implies that

$$a_j \cos(b_j) = a_{j+1} \cos(b_{j+1}) \;\Rightarrow\; \frac{a_j}{a_{j+1}} = \frac{\cos(b_{j+1})}{\cos(b_j)} \tag{4}$$

Differentiability at $x = 0$ implies

$$a_j \sqrt{p_j} \sin(b_j) = a_{j+1} \sqrt{p_{j+1}} \sin(b_{j+1}) \;\Leftrightarrow\; \frac{a_j}{a_{j+1}} = \frac{\sqrt{p_{j+1}} \sin(b_{j+1})}{\sqrt{p_j} \sin(b_j)}$$

These allow us to bound the ratio a_j/a_{j+1}. We have

$$\frac{a_j}{a_{j+1}} = \frac{\sqrt{p_{j+1}} \sin(b_{j+1})}{\sqrt{p_j} \sin(b_j)} \;\Rightarrow\; \left(\frac{a_j}{a_{j+1}}\right)^2 = \frac{p_{j+1} \sin^2(b_{j+1})}{p_j \sin^2(b_j)} = \frac{p_{j+1}(1 - \cos^2(b_{j+1}))}{p_j(1 - \cos^2(b_j))} \tag{5}$$

On the other hand, from (4) we know that

$$\frac{a_j}{a_{j+1}} = \frac{\cos(b_{j+1})}{\cos(b_{j+1})} \;\Rightarrow\; \left(\frac{a_j}{a_{j+1}}\right)^2 = \frac{\cos^2(b_{j+1})}{\cos^2(b_j)} \;\Rightarrow\; \cos^2(b_{j+1}) = \cos^2(b_j)\left(\frac{a_j}{a_{j+1}}\right)^2 \tag{6}$$

Substitute (6) in (5) we get

$$\left(\frac{a_j}{a_{j+1}}\right)^2 = \frac{p_{j+1}}{p_j} \frac{1 - \cos^2(b_j)(\frac{a_j}{a_{j+1}})^2}{1 - \cos^2(b_j)} \;\Rightarrow\; \left(\frac{a_j}{a_{j+1}}\right)^2 (1 - \cos^2(b_j)) = \frac{p_{j+1}}{p_j}\left(1 - \cos^2(b_j)\left(\frac{a_j}{a_{j+1}}\right)^2\right)$$

And we have

$$\left(\frac{a_j}{a_{j+1}}\right)^2 (1 - \cos^2(b_j) + \frac{p_{j+1}}{p_j}\cos^2(b_j)) = \frac{p_{j+1}}{p_j}$$

implying that

$$\left(\frac{a_j}{a_{j+1}}\right)^2 = \frac{\frac{p_{j+1}}{p_j}}{1 - \cos^2(b_j)\left(1 - \frac{p_{j+1}}{p_j}\right)} \tag{7}$$

WLOG assume that $p_{j+1}/p_j \geq 1$ then

$$\cos^2(b_j)\left(1 - \frac{p_{j+1}}{p_j}\right) \leq 0 \;\Rightarrow\; \frac{1}{1 - \cos^2(b_j)\left(1 - \frac{p_{j+1}}{p_j}\right)} \leq 1$$

As a result we get

$$\left(\frac{a_j}{a_{j+1}}\right)^2 = \frac{\frac{p_{j+1}}{p_j}}{1 - \cos^2(b_j)(1 - \frac{p_{j+1}}{p_j})} \leq \frac{p_{j+1}}{p_j} \;\Rightarrow\; \frac{a_j}{a_{j+1}} \leq \sqrt{\frac{p_{j+1}}{p_j}}$$

For a lower bound note that the denominator in (7) satisfies

$$1-\cos^2(b_j)(1-\frac{p_{j+1}}{p_j})=\sin^2(b_j)+\frac{p_{j+1}}{p_j}\cos^2(b_j)\leq\frac{p_{j+1}}{p_j}$$

where the inequality is due to the assumption that $p_{j+1}\geq p_j$. Consequently, $(a_{j+1}/a_j)^2\geq 1$. In summary, we have bounded the ratios between the amplitudes of neighboring regions by

$$1\leq\frac{a_j}{a_{j+1}}\leq\sqrt{\frac{p_{j+1}}{p_j}} \tag{8}$$

We next note that these bounds are tight and are obtained in the following setup. Assume we have an even number of regions of constant density l each with equal size. Suppose that in each region the eigenfunction includes an integer number of cycles. For each q we construct an eigenfunction, by choosing a phase $b_j=0$ for $j=1,...,l$, and it holds that the border between region $l/2$ and $l/2+1$ lies at $x=0$. As a result, at this point we have

$$a_{\frac{l}{2}}\cos\left(\frac{q\sqrt{p_{\frac{l}{2}}}\,0}{Z}\right)=a_{\frac{l}{2}+1}\cos\left(\frac{q\sqrt{p_{\frac{l}{2}+1}}\,0}{Z}\right)\;\Rightarrow\;a_{\frac{l}{2}}=a_{\frac{l}{2}+1}$$

But since each region contains an integer number of cycles we get for $j=1,...,l$

$$\cos\left(\frac{q\sqrt{p_{\frac{l}{2}}}\,0}{Z}\right)=\cos\left(\frac{q\sqrt{p_j}}{Z}\left(\frac{2\pi}{l}j-\pi\right)\right)=1 \tag{9}$$

Continuity implies for $j=2,...,l$

$$a_{j-1}\cos\left(\frac{q\sqrt{p_{j-1}}}{Z}\left(\frac{2\pi(j-1)}{l}-\pi\right)\right)=a_j\cos\left(\frac{q\sqrt{p_j}}{Z}\left(\frac{2\pi(j-1)}{l}-\pi\right)\right)\Rightarrow a_{j-1}=a_j$$

As a result, for each q we get one eigenfunction (up to a global scale)

$$f_q^1(x)=\cos\left(\frac{q\sqrt{p_j}x}{Z}\right)\quad\text{, for }\;x\in\left[\frac{2\pi(j-1)}{l}-\pi,\frac{2\pi j}{l}-\pi\right] \tag{10}$$

We next construct a second eigenfunction for each q. Since there is an integer number of cycles in each region, to keep the second eigenfunction of each q orthogonal to the first one, we choose a phase of $-\pi/2$:

$$f_q^1(x)=a_j\sin\left(\frac{q\sqrt{p_j}x}{Z}\right)\quad\text{, for }\;x\in\left[\frac{2\pi(j-1)}{l}-\pi,\frac{2\pi j}{l}-\pi\right]$$

Next, to maintain differentiability, the derivative at the border between regions R_j and R_{j+1} must be equal. So at $x=2\pi j/l-\pi$ we have for $j=1,...,l-1$

$$\frac{d}{dx}\left(a_j\sin\left(\frac{q\sqrt{p_j}x}{Z}\right)\right)=\frac{d}{dx}\left(a_{j+1}\sin\left(\frac{q\sqrt{p_{j+1}}x}{Z}\right)\right)\Rightarrow$$

$$-\frac{a_jq\sqrt{p_j}}{Z}\cos\left(\frac{q\sqrt{p_j}x}{Z}\right)=-\frac{a_{j+1}q\sqrt{p_{j+1}}}{Z}\cos\left(\frac{q\sqrt{p_{j+1}}x}{Z}\right)\Rightarrow$$

$$a_j\sqrt{p_j}\cos\left(\frac{q\sqrt{p_j}x}{Z}\right)=a_{j+1}\sqrt{p_{j+1}}\cos\left(\frac{q\sqrt{p_{j+1}}x}{Z}\right)$$

From (9) we have

$$a_j\sqrt{p_j}=a_{j+1}\sqrt{p_{j+1}}\Rightarrow\frac{a_j}{a_{j+1}}=\frac{\sqrt{p_{j+1}}}{\sqrt{p_j}}$$

And we can choose for the second eigenfunction for each q (up to a global scale)

$$f_q^2(x)=\frac{1}{\sqrt{p_j}}\sin\left(\frac{q\sqrt{p_j}x}{Z}\right)\quad\text{, for }\;x\in\left[\frac{2\pi(j-1)}{l}-\pi,\frac{2\pi j}{l}-\pi\right] \tag{11}$$

In Figure 1 we show an example for this setup.

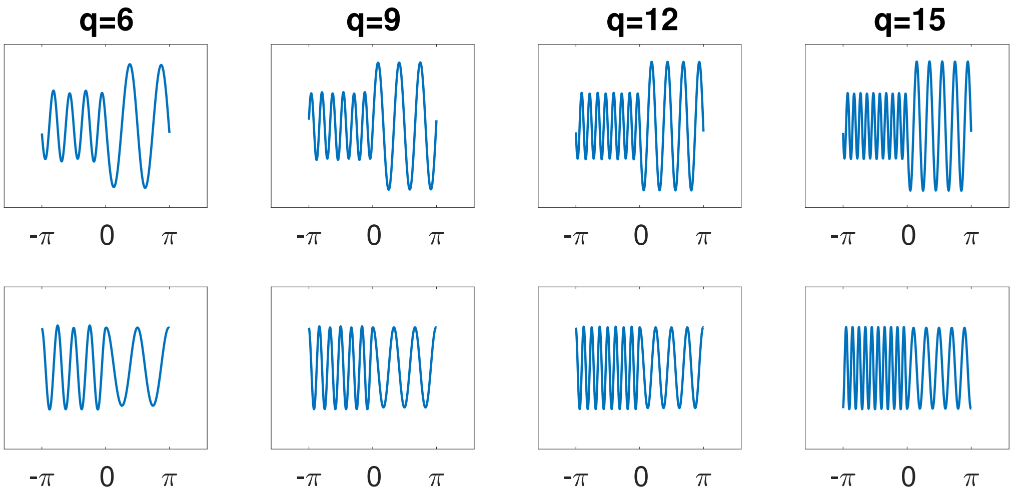

Figure 1. For the NTK of a two-layer network with bias we plot in each of the four columns four of its eigenfunction pairs (each of the same eigenvalue) under a non-uniform data distribution of $p(x) \in 1/\pi\{4/5, 1/5\}$ in $\mathbb{S}^1$. For this distribution whenever $\mathrm{mod}(q, 3) = 0$ there is an integer number of cycles in each region. As a result, for each q we obtain two eigenfunctions of the form of (10) and (11).

B.2. Bounding a_j

Assuming $p(x)$ is constant in l regions and that WLOG up to a global scale, the minimal amplitude is $a_{\min} = 1$. Then for two neighboring regions R_j and R_{j+1} if $p_j \geq p_{j+1} \Rightarrow \frac{a_{j+1}}{a_j} \leq \sqrt{\frac{p_j}{p_{j+1}}} \leq \sqrt{\frac{p_{\max}}{p_{\min}}}$ and if $p_{j+1} \geq p_j \Rightarrow \frac{a_j}{a_{j+1}} \geq 1 \Rightarrow \frac{a_{j+1}}{a_j} \leq 1 \leq \sqrt{\frac{p_{\max}}{p_{\min}}}$. As a result in each transition between two regions we have

$$\frac{a_{i+1}}{a_i} \leq \sqrt{\frac{p_{\max}}{p_{\min}}}$$

Starting from a minimal amplitude of magnitude 1. For l regions there are no more than l transitions so each amplitude is (loosely) bounded as follows

$$a_j \leq a_{min} \left(\sqrt{\frac{p_{\max}}{p_{\min}}} \right)^l = \left(\frac{p_{\max}}{p_{\min}} \right)^{\frac{l}{2}}$$

Next we bound the global scale factor. Let $s = \int_{-\pi}^{\pi} (f(x))^2 dx$. Then we have that after normalizing the global scale factor

$$a_j \leq \frac{1}{\sqrt{s}} \left(\frac{p_{\max}}{p_{\min}} \right)^{\frac{l}{2}}$$

To simplify notation we denote the frequency of each region by $q_j = \frac{\sqrt{p_j} q}{Z}$. Then for s we have:

$$s = \int_{-\pi}^{\pi} (f(x))^2 dx = \sum_{j=1}^{l} a_j^2 \int_{R_j} \cos^2(q_j x + b_j) dx \geq \sum_{j=1}^{l} a_{\min}^2 \int_{R_j} \cos^2(q_j x + b_j) dx = \sum_{j=1}^{l} \int_{R_j} \cos^2(q_j x + b_j) dx$$

For each region we have

$$\int_{R_j} \cos^2(q_j x + b_j) dx = \int_{-\pi + \frac{2\pi}{l}(j-1)}^{-\pi + \frac{2\pi}{l} j} \cos^2(q_j x + b_j) dx =$$

$$\frac{1}{2} \int_{-\pi + \frac{2\pi}{l}(j-1)}^{-\pi + \frac{2\pi}{l} j} (1 + \cos(2q_j x + 2b_j)) dx = \frac{1}{2} \left(x + \frac{\sin(2q_j x + 2b_j)}{2q_j} \right) \Bigg|_{-\pi + \frac{2\pi}{l}(j-1)}^{-\pi + \frac{2\pi}{l} j} =$$

$$\frac{1}{2}\left(-\pi+\frac{2\pi}{l}j+\frac{\sin(2q_j(-\pi+\frac{2\pi}{l}j)+2b_j)}{2q_j}-(-\pi+\frac{2\pi}{l}(j-1))-\frac{\sin(2q_j(-\pi+\frac{2\pi}{l}(j-1))+2b_j)}{2q_j}\right)=$$

$$\frac{1}{2}\left(\frac{2\pi}{l}+\frac{\sin(2q_j(-\pi+\frac{2\pi}{l}j)+2b_j)}{2q_j}-\frac{\sin(2q_j(-\pi+\frac{2\pi}{l}(j-1))+2b_j)}{2q_j}\right)\geq\frac{\pi}{l}-\frac{1}{2q_j}$$

So we get $s \geq \sum_{j=1}^{l}\frac{\pi}{l}-\frac{1}{2q_j}=\pi-\frac{1}{2}\sum_{j=1}^{l}\frac{1}{q_j}=\pi-\frac{1}{2}\sum_{j=1}^{l}\frac{Z}{\sqrt{p_j}q}$.

And we get:

$$s\geq\pi-\frac{1}{2}\sum_{j=1}^{l}\frac{Z}{\sqrt{p_j}q}=\pi-\frac{Z}{2q}\sum_{j=1}^{l}\frac{1}{\sqrt{p_j}}$$

As a result all the amplitudes in an eigenfunction of order q are bounded by

$$a_i\leq\frac{1}{\sqrt{\pi-\frac{Z}{2q}\sum_{j=1}^{l}\frac{1}{\sqrt{p_j}}}}\left(\frac{p_{\max}}{p_{\min}}\right)^{\frac{l}{2}}\quad\text{for all } 1\leq i\leq l \tag{12}$$

C. Convergence rate as a function of frequency

To derive the rate of convergence as a function of frequency and density we assume that $p(x)$ forms a piecewise-constant distribution (PCD) with a fixed number of pieces l of equal sizes, $p(x)=p_j$ in R_j, $1\leq j\leq l$. Our proof will rely on a lemma that states informally that not too many eigenfunctions need to be taken into account for convergence – more precisely, only a number linear in k and inversely linear in $\sqrt{p^*}$, where $p^*>0$ denotes the minimal density. Convergence rate is then determined by the eigenfunction with highest eigenvalue included in the approximation for $g(x)$.

Lemma 2. *Let $p(x)$ be PCD. For any $\epsilon>0$, there exist n_k such that $\sum_{i=n_k+1}^{\infty}g_i^2<\epsilon^2$, where $g_i=\int_{-\pi}^{\pi}v_i(x)g(x)p(x)dx$ and n_k is bounded as in (15) below.*

Proof. Given a target function $g(x)=\cos(kx)$ and a basis function $v_i(x)=a(x)\cos(\frac{q_i\sqrt{p(x)}x}{Z}+b(x))$ where $q_i=\lfloor i/2\rfloor$. Their inner product can be written as

$$g_i=\sum_{j=1}^{l}a_jp_j\int_{R_j}\cos(kx)\cos(q_{ij}x+b_j)dx \tag{13}$$

where $q_{ij}=q_i\sqrt{p_j}/Z$ denotes the local frequency of $v_i(x)$ at R_j. Next, to derive a bound we will restrict our treatment to $q_{ij}\geq 2k$ (and by that bound n_k from below). With this assumption we obtain

$$\left|\int_{R_j}\cos(kx)\cos(q_{ij}x+b_j)dx\right|\leq\left|\int_{-\frac{\pi}{l}}^{\frac{\pi}{l}}\cos(kx)\cos(q_{ij}x)dx\right|=$$

$$\left|\frac{\sin\left(\frac{\pi(q_{ij}+k)}{l}\right)}{q_{ij}+k}+\frac{\sin\left(\frac{\pi(q_{ij}-k)}{l}\right)}{q_{ij}-k}\right|\leq\frac{1}{q_{ij}+k}+\frac{1}{q_{ij}-k}=\frac{2q_{ij}}{q_{ij}^2-k^2}\leq\frac{8}{3q_{ij}}$$

Let $p^*=\min_j p_j$ and let $q_i^*=q_i\sqrt{p^*}/Z$, q_i^* denotes the frequency associated with the corresponding region (which is the lowest within v_i). Our requirement that $q_{ij}>2k$ for all $1\leq j\leq l$ implies that $q_i^*>2k$, and therefore

$$q_i>\frac{2Zk}{\sqrt{p^*}} \tag{14}$$

Additionally, using (13)

$$|g_i|\leq\frac{8}{3}\sum_{j=1}^{l}\frac{a_jp_j}{q_{ij}}\leq\frac{8}{3q_i^*}\sum_{j=1}^{l}a_jp_j=\frac{8B}{3q_i^*}=\frac{8BZ}{3q_i\sqrt{p^*}}$$

where we denote by $B = \sum_{j=1}^{l} a_j p_j$ and the equality on the right is obtained by plugging in the definition of q_i^*. Note that $\sum_{j=1}^{l} p_j = l/(2\pi)$ (since $1 = \int_{-\pi}^{\pi} p(x)dx = \sum_{j=1}^{l} 2\pi p_j/l$), implying that $B \leq la^*/(2\pi)$, where $a^* = \max_j a_j$ and a^* is bounded by (12).

Next, for a given $\epsilon > 0$ we wish to bound the sum $\sum_{i=n_k}^{\infty} g_i^2$ by starting from a sufficiently high index n_k, i.e.,

$$\sum_{i=n_k+1}^{\infty} g_i^2 \leq \left(\frac{8BZ}{3\sqrt{p^*}}\right)^2 \sum_{i=n_k+1}^{\infty} \frac{1}{q_i^2} < \frac{1}{q_{n_k}} \left(\frac{8BZ}{3\sqrt{p^*}}\right)^2 < \epsilon^2$$

By the definition of q_i, $n_k \geq 2q_{n_k}$, so

$$n_k > \frac{2}{\epsilon^2} \left(\frac{8BZ}{3\sqrt{p^*}}\right)^2 = \frac{128B^2Z^2}{9\epsilon^2 p^*}$$

So in conclusion,

$$n_k > \max\left\{\frac{4Zk}{\sqrt{p^*}}, \frac{128B^2Z^2}{9\epsilon^2 p^*}\right\} \tag{15}$$

□

Theorem 1. *Let $p(x)$ be a PCD, for any $\delta > 0$ the number of iterations t needed to achieve $\|g(x) - u^{(t)}(x)\| < \delta$ is $\tilde{O}(k^2/p^*)$, where $\tilde{O}$ hides logarithmic terms.*

Proof. Let n_k be chosen as in Lemma 2 with $\epsilon = \delta/2$, i.e.

$$n_k = \left\lceil \max\left\{\frac{4Zk}{\sqrt{p^*}}, \frac{256B^2Z^2}{9\delta^2 p^*}\right\}\right\rceil$$

Let

$$\hat{g}(x) = \sum_{i=1}^{n_k} g_i v(i)$$

Then,

$$\|g(x) - \hat{g}(x)\|^2 = \sum_{i=n_k+1}^{\infty} g_i^2 < \left(\frac{\delta}{2}\right)^2$$

and due to triangle inequality

$$\|g(x) - u^{(t)}(x)\| \leq \|g(x) - \hat{g}(x)\| + \|\hat{g}(x) - u^{(t)}(x)\|$$

it suffices to find t such that

$$\|\hat{g}(x) - u^{(t)}(x)\| < \frac{\delta}{2} = \tilde{\delta}$$

Using (Arora et al., 2019b)'s Theorem 4.1 adapted to continuous operators

$$\Delta^2 = \|\hat{g} - u^{(t)}\|^2 \approx \sum_{i=1}^{n_k} (1 - \eta\lambda_i)^{2t} g_i^2 \leq \pi \sum_{i=1}^{n_k} (1 - \eta\lambda_i)^{2t} \leq \pi n_k (1 - \eta\lambda_{n_k})^{2t} \tag{16}$$

where the left inequality is due to $|g_i|^2 \leq \|\cos^2(kx)\| = \pi$ and the right inequality is because λ_i are arranged in a descending order. Now for a fixed distribution $p(x)$, and since we are interested in the asymptotic rate of convergence (i.e., as $k \to \infty$), as soon as $k > 64B^2Z/(9\tilde{\delta}^2\sqrt{p^*})$ it suffices to only consider the case $q_{n_k} = 2Zk/\sqrt{p^*}$, as in (14). The eigenvalue λ_{n_k} is determined according to

$$\lambda_{n_k} = \frac{Z^2}{\pi^2 q_{n_k}^2} = \frac{p^*}{4\pi^2 k^2}$$

(Here we used the expression for λ_{n_k} assuming n_k is odd. A similar expression of the same order is obtained for even n_k.) Consequently, to bound $\Delta^2 < \tilde{\delta}$ in (16) and substituting for n_k and λ_{n_k} we have

$$\frac{4Zk}{\sqrt{p^*}} \left(1 - \frac{\eta p^*}{4\pi^2 k^2}\right)^{2t} < \tilde{\delta}$$

Taking log

$$2t\log\left(1-\frac{\eta p^*}{4\pi^2k^2}\right)>\log\left(\frac{\delta\sqrt{p^*}}{4Zk}\right)$$

from which we obtain

$$t>\frac{\log\left(\frac{\delta\sqrt{p^*}}{4Zk}\right)}{2\log\left(1-\frac{\eta p^*}{4\pi^2k^2}\right)}\approx-\frac{2\pi^2k^2}{\eta p^*}\log\left(\frac{\delta\sqrt{p^*}}{4Zk}\right)=\tilde{O}\left(\frac{k^2}{p^*}\right)$$

where $\tilde{O}$ hides logarithmic terms.

□

D. Spectral convergence analysis for deep networks - proof of Theorem 2

D.1. The network model

The parameters of the network are $W=(W_1,...,W_L)$ where $W_l\in\mathbb{R}^{m\times m}$ and also $A\in\mathbb{R}^{m\times d}$ and $B\in\mathbb{R}^{1\times m}$. The network function over input $\mathbf{x}_i\in\mathbb{R}^d$ ($i\in[n]$) is given by

$$u_i=f(\mathbf{x}_i;W)=B\sigma(W_L\sigma(W_{L-1}\sigma(....(W_1\sigma(Ax_i))..))$$

where σ stands for element wise RELU activation function. For a tuple $W=(W_1,...,W_L)$ of matrices, we let $\|W\|_2=\max_{l\in[L]}\|W_l\|_2$ and $\|W\|_F=(\sum_{l=1}^L\|W_l\|_F^2)^{1/2}$.

The parameters are initialized randomly from a normal distribution according to

$$\begin{aligned}[W_l]_{ij}&\sim\mathcal{N}(0,\frac{2}{m}),\ l\in[L]\\ A_{ij}&\sim\mathcal{N}(0,\frac{2}{m})\\ B_{ij}&\sim\mathcal{N}(0,\tau^2)\end{aligned}\tag{17}$$

where similarly to (Allen-Zhu et al., 2019) the layers A and B are initialized and held fixed.

The network functionality is summarized as follows

$$\begin{aligned}\mathbf{h}_{i,0}&=\sigma(A\mathbf{x}_i)\\ \mathbf{h}_{i,l}^{(t)}&=\sigma(W_l^{(t)}\mathbf{h}_{i,l-1}^{(t)})\\ \mathbf{u}_i^{(t)}&=B\mathbf{h}_{i,L}^{(t)}\end{aligned}$$

where $i\in[n]$, $l\in[L]$ and t denotes iteration number. In addition, for each input vector $i\in[n]$ and layer $l\in\{0,1,...,L\}$, we associate a diagonal matrix $D_{i,l}$ such that for $j\in[m]$, $(D_{i,l})_{j,j}=\mathbb{I}_{(W_l\mathbf{h}_{i,l-1})_j\geq0}$, where we use the convention $\mathbf{h}_{i,-1}=\mathbf{x}_i$. The network is trained to minimize the ℓ_2 loss

$$\Phi(W)=\frac{1}{2}\sum_{i=1}^n(y_i-f(\mathbf{x}_i;W))^2$$

We will analyze the properties of the matrices $H,H^\infty\in\mathbb{R}^{n\times n}$, comprised of the following entries

$$H_{ij}(t)=\left\langle\frac{\partial u_i^{(t)}}{\partial W},\frac{\partial u_j^{(t)}}{\partial W}\right\rangle$$

$$H_{ij}^\infty=\mathbb{E}_W\left\langle\frac{\partial u_i^{(0)}}{\partial W},\frac{\partial u_j^{(0)}}{\partial W}\right\rangle.$$

We write the eigen-decomposition of $H^\infty=\sum_{i=1}^n\lambda_i\mathbf{v}_i\mathbf{v}_i^T$, where $\mathbf{v}_1,\dots,\mathbf{v}_n$ are the eigenvectors of H^∞ and $\lambda_1,\dots,\lambda_n$ are their corresponding eigenvalues. The minimal eigenvalue is denoted by $\lambda_0=\min(\lambda(H^\infty))$.

Theorem 2. *For any $\epsilon \in (0,1]$ and $\delta \in (0, O(\frac{1}{L})]$, let $\tau = \Theta(\frac{\epsilon\hat{\delta}}{n})$, $m \geq \Omega\left(\frac{n^{24}L^{12}\log^5 m}{\delta^8\tau^6}\right)$, $\eta = \Theta\left(\frac{\delta}{n^4L^2m\tau^2}\right)$. Then, with probability of at least $1-\hat{\delta}$ over the random initialization after t iterations of GD we have that*

$$\|\mathbf{y}-\mathbf{u}^{(t)}\| = \sqrt{\sum_{i=1}^{n}(1-\eta\lambda_i)^{2t}(\mathbf{v}_i^T\mathbf{y})^2} \pm \epsilon. \tag{18}$$

D.2. Proof strategy

The proof of Thm. 2 relies on a theorem, provided by (Allen-Zhu et al., 2019), stated in Thm. 3, and an observation, based the on the derivation of the proof to that theorem, which we state in Lemma 4.

Thm. 3 assumes that the data is normalized, so that $\|\mathbf{x}_i\| = 1$, and there exists $\delta \in (0, O(\frac{1}{L})]$ such that for every pair $i, j \in [n]$, we have $\|\mathbf{x}_i - \mathbf{x}_j\| \geq \delta$ and also it holds that $|y_i| \leq O(1)$.

In addition, we prove Lemma 3, which is the basis for the proof of our Theorem.

Lemma 3. *Suppose $\delta \in (0, O(\frac{1}{L})]$, $m \geq \Omega\left(\frac{n^{24}L^{12}\log^5 m}{\delta^8\tau^2}\right)$, $\eta = \Theta\left(\frac{\delta}{n^4L^2m\tau^2}\right)$ and also let $\omega = O(\frac{n^3\log m}{\delta\tau\sqrt{m}})$. Then, with probability at least $1-e^{-\Omega(m\omega^{2/3}L)}$ over the randomness of A, B and $W^{(0)}$ we have*

$$\mathbf{u}(t+1)-\mathbf{y} = (I-\eta H(t))(\mathbf{u}(t)-\mathbf{y}) + \epsilon(t) \tag{19}$$

with

$$\|\epsilon(t)\| \leq O\left(\frac{L\log^{4/3} m}{\tau^{1/3}m^{1/6}n^{1.5}}\right)\sqrt{\Phi(W^{(t)})} + O\left(\frac{\delta^2}{\tau n^6 m^{0.5}L^{1.5}}\right)\Phi(W^{(t)})$$

The proof of the Lemma is deferred, and will be given after the proof of the theorem.

D.3. Proof of Thm 2

Proof. By Lemma 3 we have the following relation

$$\mathbf{u}(t)-\mathbf{y} = (I-\eta H(t-1))(\mathbf{u}(t-1)-\mathbf{y}) + \epsilon(t-1)$$

Adding and subtracting $\eta H^\infty(\mathbf{u}(t-1)-\mathbf{y})$ we have

$$\mathbf{u}(t)-\mathbf{y} = (I-\eta H^\infty)(\mathbf{u}(t-1)-\mathbf{y}) + \eta(H^\infty - H(t-1))(\mathbf{u}(t-1)-\mathbf{y}) + \epsilon(t-1)$$

and this is equivalent to

$$\mathbf{u}(t)-\mathbf{y} = (I-\eta H^\infty)(\mathbf{u}(t-1)-\mathbf{y}) + \xi(t-1). \tag{20}$$

where we denote $\xi(t) = \eta(H^\infty - H(t))(\mathbf{u}(t)-\mathbf{y}) + \epsilon(t)$. Then, by applying (20) recursively, we obtain

$$\mathbf{u}(t)-\mathbf{y} = (I-\eta H^\infty)^t(\mathbf{u}(0)-\mathbf{y}) + \sum_{i=0}^{t-1}(I-\eta H^\infty)^i\xi(t-1-i) \tag{21}$$

We first bound the quantity $\left\|\xi(t-1-i)\right\|_2$

$$
\begin{aligned}
&\left\|\xi(t-1-i)\right\|_2 = \left\|\eta(H(t-1-i)-H^{\infty})(y-u(t-1-i))+\epsilon(t-1-i)\right\|_2 \\
&\leq \left\|\eta(H(t-1-i)-H^{\infty})\right\|_2 \left\|(y-u(t-1-i))\right\|_2 + \left\|\epsilon(t-1-i)\right\|_2 \\
&\eta \leq^{1,2} O\left(\frac{\delta^2 m\tau^3}{n^6}\right)\sqrt{\Phi(W^{(t-1-i)}))} + O\left(\frac{\delta^2}{\tau n^6 m^{0.5} L^{1.5}}\right)\Phi(W^{(t-1-i)}) + O\left(\frac{L\log^{4/3} m}{\tau^{1/3} m^{1/6} n^{1.5}}\right)\sqrt{\Phi(W^{(t-1-i)})} \\
&\leq^{3} \left(1-\Omega\left(\frac{\tau^2\eta\delta m}{n^2}\right)\right)^{\frac{t-1-i}{2}}\left(\eta O\left(\frac{\delta^2 m\tau^3}{n^6}\right)\sqrt{\Phi(W^{(0)})} + O\left(\frac{\delta^2}{\tau n^6 m^{0.5} L^{1.5}}\right)\Phi(W^{(0)}) + O\left(\frac{L\log^{4/3} m}{\tau^{1/3} m^{1/6} n^{1.5}}\right)\sqrt{\Phi(W^{(0)})}\right) \\
&\leq^{4} \left(1-\Omega\left(\frac{\tau^2\eta\delta m}{n^2}\right)\right)^{\frac{t-1-i}{2}}\left(\eta O\left(\sqrt{n}\right) O\left(\frac{\delta^2 m\tau^3}{n^6}\right) + O\left(\frac{\delta^2}{\tau n^6 m^{0.5} L^{1.5}}\right) O\left(n\right) + O\left(\frac{L\log^{4/3} m}{\tau^{1/3} m^{1/6} n^{1.5}}\right) O\left(\sqrt{n}\right)\right) \\
&= \left(1-\Omega\left(\frac{\tau^2\eta\delta m}{n^2}\right)\right)^{\frac{(t-1-i)}{2}}\left(\eta O\left(\frac{\delta^2 m\tau^3}{n^{5.5}}\right) + O\left(\frac{\delta^2}{\tau n^5 m^{0.5} L^{1.5}}\right) + O\left(\frac{L\log^{4/3} m}{\tau^{1/3} m^{1/6} n}\right)\right)
\end{aligned}
$$

where we make the following derivations

1. Using Lemma 14 which states that $\left\|H(t)-H^{\infty}\right\|_2 \leq O(\frac{\delta^2 m\tau^3}{n^6})$.
2. Using the bound in Lemma 3, for $\epsilon(t-1-i)$
3. Using bound over the loss by, Lemma 4 (b).
4. By Lemma 11 the loss at initialization is bounded by $O(n)$.

Using the bound, derived above, (21) yields

$$
\begin{aligned}
&\left\|\mathbf{u}(t)-\mathbf{y}\right\| = \left\|(I-\eta H^{\infty})^t(\mathbf{u}(0)-\mathbf{y}) + \sum_{i=0}^{t-1}((I-\eta H^{\infty})^i\xi(t-1-i))\right\| \\
&\leq^{1} \left\|(I-\eta H^{\infty})^t(\mathbf{u}(0)-\mathbf{y})\right\| \\
&+ \sum_{i=0}^{t-1}(1-\eta\lambda_0)^i\left(1-\Omega\left(\frac{\tau^2\eta\delta m}{n^2}\right)\right)^{\frac{(t-1-i)}{2}}\left(\eta O\left(\frac{\delta^2 m\tau^3}{n^{5.5}}\right) + O\left(\frac{\delta^2}{\tau n^5 m^{0.5} L^{1.5}}\right) + O\left(\frac{L\log^{4/3} m}{\tau^{1/3} m^{1/6} n}\right)\right) \\
&\leq^{2} \left\|(I-\eta H^{\infty})^t(\mathbf{u}(0)-\mathbf{y})\right\| + t\left(\eta O\left(\frac{\delta^2 m\tau^3}{n^{5.5}}\right) + O\left(\frac{\delta^2}{\tau n^5 m^{0.5} L^{1.5}}\right) + O\left(\frac{L\log^{4/3} m}{\tau^{1/3} m^{1/6} n}\right)\right) \\
&\leq^{3} \left\|(I-\eta H^{\infty})^t(\mathbf{u}(0)-\mathbf{y})\right\| + O\left(\frac{n^6 L^2}{\delta^2}\right)\left(\eta O\left(\frac{\delta^2 m\tau^3}{n^{5.5}}\right) + O\left(\frac{\delta^2}{\tau n^5 m^{0.5} L^{1.5}}\right) + O\left(\frac{L\log^{4/3} m}{\tau^{1/3} m^{1/6} n}\right)\right) \\
&\leq \left\|(I-\eta H^{\infty})^t\right\|\left\|\mathbf{u}(0)\right\| + \left\|(I-\eta H^{\infty})^t\mathbf{y}\right\| + O\left(\frac{n^6 L^2}{\delta^2}\right)\left(\eta O\left(\frac{\delta^2 m\tau^3}{n^{5.5}}\right) + O\left(\frac{\delta^2}{\tau n^5 m^{0.5} L^{1.5}}\right) + O\left(\frac{L\log^{4/3} m}{\tau^{1/3} m^{1/6} n}\right)\right)
\end{aligned}
$$

where we make the following derivations

1. $\left\|I-\eta H^{\infty}\right\|_2$ is bounded by the maximal eigenvalue of the positive definite matrix $(I-\eta H^{\infty})$, i.e, $(1-\eta\lambda_0)$.
2. $(1-\eta\lambda_0)^i\left(1-\Omega\left(\frac{\tau^2\eta\delta m}{n^2}\right)\right)^{\frac{(t-1-i)}{2}} \leq 1$
3. By Theorem 3, $t \leq O(\frac{n^6 L^2}{\delta^2})$

Next, it is straightforward to show that

$$\left\|(I-\eta H^{\infty})^{t}\mathbf{y}\right\| = \sqrt{\sum_{i=1}^{n}(1-\eta\lambda_i)^{2t}(\mathbf{v}_i^T\mathbf{y})^2} \tag{22}$$

where $\lambda_i, \mathbf{v}_i$ are the eigenvalues and eigenvectors of H^{∞}, respectively.

For the first term we use lemma 11 which states that $\|\mathbf{u}(0)\| \leq \frac{\sqrt{n}\tau}{\hat{\delta}}$, and by our choice of τ we obtain

$$\left\|(I-\eta H^{\infty})^{t}\right\| \|\mathbf{u}(0)\| \leq (1-\eta\lambda_0)^t O\left(\frac{\sqrt{n}\tau}{\hat{\delta}}\right) \leq \epsilon \tag{23}$$

Finally, by our choice of η, m, τ it holds that

$$O\left(\frac{n^6L^2}{\delta^2}\right)\left(O\left(\frac{\delta^2 m\tau^3}{n^{5.5}}\right)\eta + O\left(\frac{\delta^2}{\tau n^5 m^{0.5} L^{1.5}}\right) + O\left(\frac{L\log^{4/3} m}{\tau^{1/3} m^{1/6} n}\right)\right) \leq \epsilon \tag{24}$$

Combining (22), (23) and (24) yields

$$\|\mathbf{y}-\mathbf{u}(t)\| = \sqrt{\sum_{i=1}^{n}(1-\eta\lambda_i)^{2k}(\mathbf{v}_i^T\mathbf{y})^2} \pm \epsilon \tag{25}$$

□

D.4. Supporting Lemmas

Proof. Proof of Lemma 3.

By construction

$$\begin{aligned}
\epsilon_i(t) &= u_i(t+1) - u_i(t) + [\eta H(t)(\mathbf{u}(t)-\mathbf{y})]_i \\
&= u_i(t+1) - u_i(t) + \eta\sum_{j=1}^{n}(u_j(t)-y_j)H_{ij}(t) \\
&= u_i(t+1) - u_i(t) + \eta\sum_{j=1}^{n}(u_j(t)-y_j)\left\langle\frac{\partial u_i(t)}{\partial W}, \frac{\partial u_j(t)}{\partial W}\right\rangle \\
&= u_i(t+1) - u_i(t) + \eta\left\langle\frac{\partial u_i(t)}{\partial W}, \sum_{j=1}^{n}(u_j(t)-y_j)\frac{\partial u_j(t)}{\partial W}\right\rangle \\
&= u_i(t+1) - u_i(t) + \eta\left\langle\frac{\partial u_i}{\partial W}, \nabla\Phi(W^{(t)})\right\rangle.
\end{aligned}$$

We denote $-\eta\nabla\Phi(W^{(t)})$ by $W' = (W_1^{'}, ..., W_L^{'})$, yielding

$$\begin{aligned}
\epsilon_i(t) &= u_i(t+1) - u_i(t) - \left\langle\frac{\partial u_i(t)}{\partial W}, W'\right\rangle \\
&= B(h_{i,L}^{(t+1)} - h_{i,L}^{(t)}) - \left\langle\frac{\partial u_i(t)}{\partial W}, W'\right\rangle \\
&= B(h_{i,L}^{(t+1)} - h_{i,L}^{(t)} - \sum_{l=1}^{L} D_{i,L}^{(t)}W_L^{(t)}D_{i,L-1}^{(t)}W_{L-1}^{(t)}\cdots D_{i,L+1}^{(t)}W_{l+1}^{(t)}D_{i,l}^{(t)}W_l'h_{i,l-1}^{(t)}) \\
&= B\left(\sum_{l=1}^{L}(D_{i,L}^{(t)} + D_{i,L}'')W_L^{(t)}\cdots W_{l+1}^{(t)}(D_{i,l}^{(t)} + D_{i,l}'')W_l'h_{i,l-1}^{(t+1)} - \sum_{l=1}^{L} D_{i,L}^{(t)}W_L^{(t)}\cdots W_{l+1}^{(t)}D_{i,l}^{(t)}W_l'h_{i,l-1}^{(t)}\right)
\end{aligned}$$

where the last equality is obtained by replacing $h_{i,L}^{(t+1)} - h_{i,L}^{(t)}$ by the term provided in Lemma 5, where $D''_{i,l} \in \mathbb{R}^{m\times m}$ are diagonal matrices with entries in $[-1, 1]$.

Now, we derive a bound for $|\epsilon_i(t)|$. We start by subtracting and adding the same term, yielding

$$\begin{aligned}
|\epsilon_i(t)| = |B(&\sum_{l=1}^{L}(D_{i,L}^{(t)} + D''_{i,L})W_L^{(t)}\cdots W_{l+1}^{(t)}(D_{i,l}^{(t)} + D''_{i,l})W'_l h_{i,l-1}^{(t+1)} - D_{i,L}^{(t)}W_L^{(t)}\cdots W_{l+1}^{(t)}D_{i,l}^{(t)}W'_l h_{i,l-1}^{(t+1)} \\
&+ \sum_{l=1}^{L} D_{i,L}^{(t)}W_L^{(t)}\cdots W_{l+1}^{(t)}D_{i,l}^{(t)}W'_l h_{i,l-1}^{(t+1)} - D_{i,L}^{(t)}W_L^{(t)}\cdots W_{l+1}^{(t)}D_{i,l}^{(t)}W'_l h_{i,l-1}^{(t)})| \\
&\leq \sum_{l=1}^{L}\left|B\left((D_{i,L}^{(t)} + D''_{i,L})W_L^{(t)}...W_{l+1}^{(t)}(D_{i,l}^{(t)} + D''_{i,l})W'_l h_{i,l-1}^{(t+1)} - D_{i,L}^{(t)}W_L^{(t)}\cdots W_{l+1}^{(t)}D_{i,l}^{(t)}W'_l h_{i,l-1}^{(t+1)}\right)\right| \\
&+ \sum_{l=1}^{L}\left|B\left(D_{i,L}^{(t)}W_L^{(t)}\cdots W_{l+1}^{(t)}D_{i,l}^{(t)}W'_l h_{i,l-1}^{(t+1)} - D_{i,L}^{(t)}W_L^{(t)}\cdots W_{l+1}^{(t)}D_{i,l}^{(t)}W'_l h_{i,l-1}^{(t)}\right)\right|.
\end{aligned}$$

To construct the bound for $|\epsilon_i(t)|$, we separately bound each of the above two terms. For the first term

$$\begin{aligned}
&\left|B\left((D_{i,L}^{(t)} + D''_{i,L})W_L^{(t)}...W_{l+1}^{(t)}(D_{i,l}^{(t)} + D''_{i,l})W'_l h_{i,l-1}^{(t+1)} - D_{i,L}^{(t)}W_L^{(t)}...W_{l+1}^{(t)}D_{i,l}^{(t)}W'_l h_{i,l-1}^{(t+1)}\right)\right| \\
&\leq \left\|B\left((D_{i,L}^{(t)} + D''_{i,L})W_L^{(t)}...W_{l+1}^{(t)}(D_{i,l}^{(t)} + D''_{i,l}) - D_{i,L}^{(t)}W_L^{(t)}...W_{l+1}^{(t)}D_{i,l}^{(t)}\right)\right\|_2 \left\|W'_l h_{i,l-1}^{(t+1)}\right\|_2 \\
&\overset{1}{\leq} \left\|B\left((D_{i,L}^{(t)} + D''_{i,L})W_L^{(t)}...W_{l+1}^{(t)}(D_{i,l}^{(t)} + D''_{i,l}) - D_{i,L}^{(0)}W_L^{(0)}...W_{l+1}^{(0)}D_{i,l}^{(0)}\right)\right\|_2 O(\|W'_l\|_2) \\
&+ \left\|B\left(D_{i,L}^{(0)}W_L^{(0)}...W_{l+1}^{(0)}D_{i,l}^{(0)} - D_{i,L}^{(t)}W_L^{(t)}...W_{l+1}^{(t)}D_{i,l}^{(t)}\right)\right\|_2 O(\|W'_l\|_2)) \\
&\overset{2}{=} \left\|B\left(D_{i,L}^{(0)} - D_{i,L}^{(0)} + D_{i,L}^{(t)} + D''_{i,L})W_L^{(t)}...W_{l+1}^{(t)}(D_{i,l}^{(0)} - D_{i,l}^{(0)} + D_{i,l}^{(t)} + D''_{i,l}) - D_{i,L}^{(0)}W_L^{(0)}...W_{l+1}^{(0)}D_{i,l}^{(0)}\right)\right\|_2 O(\|W'_l\|_2) \\
&+ \left\|B\left(D_{i,L}^{(0)}W_L^{(0)}...W_{l+1}^{(0)}D_{i,l}^{(0)} - (D_{i,L}^{(0)} - D_{i,L}^{(0)} + D_{i,L}^{(t)})W_L^{(t)}...W_{l+1}^{(t)}(D_{i,l}^{(0)} - D_{i,l}^{(0)} + D_{i,l}^{(t)})\right)\right\|_2 O(\|W'_l\|_2) \\
&\overset{3}{\leq} O(\tau\omega^{1/3}L^2\sqrt{m\log m})O(\|W'_l\|_2)
\end{aligned}$$

where we apply the following derivations

1. We subtract and add the same term, use triangle inequality and the result provided in Lemma 10, $\left\|h_{i,l-1}^{(t+1)}\right\| = O(1)$.
2. Subtract and add $D_{i,l}^{(0)}$ from each coefficient that multiply $W_l^{(t)}$.
3. Due to Lemma 4, it holds that $||W^{(t)} - W^{(0)}|| \leq \omega$. This enables us to use Lemma 6, implying that $\|D_{i,l}^{(t)} - D_{i,l}^{(0)}\|_0 \leq s = O(m\omega^{2/3}L)$. Moreover, in conjunction with Lemma 5, this yields $\left\|D_{i,l}^{(t)} + D''_{i,l} - D_{i,l}^{(0)}\right\|_0 \leq s$. Having that, we can apply Lemma 7, to obtain a bound for the first term.

For the second term we have that:

$$\begin{aligned}
&\left|B(D_{i,L}^{(t)}W_L^{(t)}...W_{l+1}^{(t)}D_{i,l}^{(t)}W'_l h_{i,l-1}^{(t+1)} - D_{i,L}^{(t)}W_L^{(t)}...W_{l+1}^{(t)}D_{i,l}^{(t)}W'_l h_{i,l-1}^{(t)})\right| \\
&= \left|B(D_{i,L}^{(t)}W_L^{(t)}...W_{l+1}^{(t)}D_{i,l}^{(t)}W'_l(h_{i,l-1}^{(t+1)} - h_{i,l-1}^{(t)}))\right| \\
&\leq \left(\left\|B(D_{i,L}^{(t)}W_L^{(t)}...W_{l+1}^{(t)}D_{i,l}^{(t)} - D_{i,L}^{(0)}W_L^{(0)}...W_{l+1}^{(0)}D_{i,l}^{(0)})\right\| + \left\|BD_{i,L}^{(0)}W_L^{(0)}...W_{l+1}^{(0)}D_{i,l}^{(0)}\right\|\right)\|W'_l\|\left\|h_{i,l-1}^{(t+1)} - h_{i,l-1}^{(t)}\right\| \\
&\overset{1}{\leq} \left(O(\tau\omega^{1/3}L^2\sqrt{m\log m}) + \left\|BD_{i,L}^{(0)}W_L^{(0)}...W_{l+1}^{(0)}D_{i,l}^{(0)}\right\|\right)\|W'_l\|\left\|h_{i,l-1}^{(t+1)} - h_{i,l-1}^{(t)}\right\| \\
&\overset{2}{\leq} \tau O(\sqrt{m} + \omega^{1/3}L^2\sqrt{m\log m})\|W'_l\|\left\|h_{i,l-1}^{(t+1)} - h_{i,l-1}^{(t)}\right\| \overset{3}{\leq} \tau O(\sqrt{m} + \omega^{1/3}L^2\sqrt{m\log m})L^{1.5}\|W'\|^2 \\
&\overset{4}{\leq} O(\tau\sqrt{m})L^{1.5}\|W'\|^2
\end{aligned}$$

where we apply the following derivations

1. As in the previous derivation, using Lemma 7.

2. Applying Lemma 8.

3. Using Lemma 5.

4. Plug in $\omega = \frac{n^3 \log m}{\delta\tau\sqrt{m}}$.

Since $W' = -\eta\nabla\Phi(W^{(t)})$, we can get a bound for $\|W'\|_2$ using Lemma 9, yielding $\|W'\|_2 \leq \eta O(\tau\sqrt{nm}\sqrt{\Phi(W^{(t)})})$. Taking into account the two bounds, and summing over the all layers and data points we obtain that

$$\|\epsilon(t)\| \leq nLO(\tau w^{1/3}L^2\sqrt{m\log m})O(\eta\tau\sqrt{nm}\sqrt{\Phi(W^{(t)})}) + nLO(\tau\sqrt{m})L^{1.5}O(\eta^2\tau^2 nm\Phi(W^{(t)}))$$

Using our choice of η and the value of ω, we finally get

$$\|\epsilon(t)\| \leq O\left(\frac{L\log^{4/3} m}{\tau^{1/3}m^{1/6}n^{1.5}}\right)\sqrt{\Phi(W^{(t)})} + O\left(\frac{\delta^2}{\tau n^6 m^{0.5}L^{1.5}}\right)\Phi(W^{(t)})$$

□

Theorem 3. [1] *For any $\epsilon \in (0,1]$ and $\delta \in (0, O(\frac{1}{L})]$, let $m \geq \Omega\left(\frac{n^{24}L^{12}\log^5 m}{\delta^8\tau^2}\right)$, $\eta = \Theta\left(\frac{\delta}{n^4L^2m\tau^2}\right)$ and $W^{(0)}, A, B$ are at random initialization (17). Then, starting from Gaussian initialization, with probability at least $1 - e^{-\Omega(log^2 m)}$, gradient descent with learning rate η achieves*

$$\Phi(W) \leq \epsilon \;\; in \;\; T = \Theta\left(\frac{n^6L^2}{\delta^2}\log\frac{1}{\epsilon}\right)$$

Lemma 4. *Under the assumptions of Thm. 3, it holds that for every $t = 0, 1, .., T-1$*

$$(a) \quad \left\|W^{(t)} - W^{(0)}\right\|_F \leq \omega := O\left(\frac{n^3}{\delta\tau\sqrt{m}}\log m\right)$$

$$(b) \quad \Phi(W^{(t)}) \leq \left(1 - \Omega\left(\frac{\tau^2\eta\delta m}{n^2}\right)\right)^t \Phi(W^{(0)})$$

Lemma 5. *(This Lemma follows Claim 11.2 from (Allen-Zhu et al., 2019)) Let $\omega \in [\Omega(\frac{1}{\tau^3 m^{3/2}L^{3/2}\log^{3/2} m}), O(\frac{1}{L^{4.5}\log^3 m})]$, then under the following assumptions $\left\|W^{(t)} - W^{(0)}\right\|_2 \leq \omega$ and $\|W'\|_2 \leq w$ it holds that there exist diagonal matrices $D''_{i,l} \in \mathbb{R}^{m\times m}$ with entries in [-1,1] such that*

$$\forall i \in [n], \forall l \in [L] : h_{i,l}^{(t+1)} - h_{i,l}^{(t)} = \sum_{a=1}^{l}(D_{i,l}^{(t)} + D''_{i,l})W_l^{(t)}...W_{a+1}^{(t)}(D_{i,a}^{(t)} + D''_{i,a})W'_a h_{i,a-1}^{(t+1)}$$

Furthermore we have $\left\|h_{i,l}^{(t+1)} - h_{i,l}^{(t)}\right\| \leq O(L^{1.5})\|W'\|_2$ and $\left\|Bh_{i,l}^{(t+1)} - Bh_{i,l}^{(t)}\right\| \leq O(L\tau\sqrt{m})\|W'\|_2$ and $\left\|D''_{i,l}\right\|_0 \leq O(m\omega^{2/3}L)$

Lemma 6. *(This Lemma follows Lemma 8.2 from (Allen-Zhu et al., 2019)) Suppose $\omega \leq \frac{1}{CL^{9/2}log^3 m}$ for some sufficiently large constant $C > 1$. With probability at least $1 - e^{-\Omega(m\omega^{2/3}L)}$ for every $(W^{(t)} - W^{(0)})$ satisfying $\left\|W^{(t)} - W^{(0)}\right\|_2 \leq \omega$,*

$$\left\|D_{i,l}^{(t)} - D_{i,l}^{(0)}\right\|_0 \leq O(m\omega^{2/3}L)$$

[1] This theorem was proved in (Allen-Zhu et al., 2019), for $\tau = 1$. However, it is straightforward to generalize it for $\tau \in (0,1]$ at the price of modifying m and η by a factor of $\frac{1}{\tau^2}$

Lemma 7. *(This Lemma follows Lemma 8.7 from (Allen-Zhu et al., 2019)) For $s = O(mw^{2/3}L)$, with probability at least $1 - e^{-\Omega(s\log m)}$ over the randomness of $W^{(0)}, A, B$*

- *for all $i \in [n], a \in [L+1]$*
- *for every diagonal matrices $D'''_{i,0}, \cdots, D'''_{i,L} \in [-3,3]^{m\times m}$ with at most s non-zero entries*
- *for every perturbation with respect to the initialization $W''_1 \cdots W''_L \in \mathbb{R}^{m\times m}$ with $\|W''\|_2 \le \omega = O(1/L^{1.5})$*

it holds $\left\|B(D^{(0)}_{i,L} + D'''_{i,L})(W^{(0)}_L + W''_L)\cdots(W^{(0)}_{a+1} + W''_{a+1})(D^{(0)}_{i,a} + D'''_{i,a}) - BD^{(0)}_{i,L}W^{(0)}_L \cdots W^{(0)}_{a+1}D^{(0)}_{i,a}\right\|_2 \le O(\tau\omega^{1/3}L^2\sqrt{m\log m})$

Lemma 8. *(This Lemma follows Lemma 7.4b from (Allen-Zhu et al., 2019)) Suppose $m \ge \Omega(nL\log(nL))$. If $s = O(m\omega^{2/3}L)$ then with probability at least $1 - e^{-\Omega(s\log m)}$ for all $i \in [n], a \in [L+1]$ it holds that $\left\|v^T BD^{(0)}_{i,L}W^{(0)}_L \cdots D^{(0)}_{i,a}W^{(0)}_a\right\| \le O(\tau\sqrt{m})\,\|v\|$.*

Lemma 9. *(This Lemma follows Theorem 3 from (Allen-Zhu et al., 2019)) Let $\omega = O(\frac{\delta^{3/2}}{n^{9/2}L^6\log^3 m})$. With probability at least $1 - e^{-\Omega(m\omega^{2/3}L)}$ over the randomness of W^0, A, B, it satisfies for every $l \in [L]$ and W with $\left\|W - W^{(0)}\right\|_2 \le \omega$ that*

$$\|\nabla_{W_l}\Phi(W)\|_F^2 \le O(\tau^2\Phi(W)\cdot n\cdot m)$$

Lemma 10. *(This Lemma is based on Lemma 7.1 and Lemma 8.2c from (Allen-Zhu et al., 2019)) With high probability over the randomness of A, W we have*

$$\forall i \in [n], l \in \{0, 1, .., L\} : \|h_{i,l}\| = O(1)$$

Lemma 11. *Let $\delta > 0$ and $m \ge \Omega(L\log(nL/\delta)$ then with probability at least $1 - \delta$ it holds that $||u(0)|| \le \sqrt{n}\tau/\delta$ and as a consequence by using the triangle inequality $\Phi(W(0)) = \frac{1}{2}\|\mathbf{y} - \mathbf{u}(0)\|^2 \le O(n)$*

Proof. Conditioned on W, A it holds that $u_i(0) \backsim N(0, \tau^2\|h_{i,L}\|^2)$ and since by Lemma 10 we have that $\|h_{i,L}\| = O(1)$, this yields $E(\|\mathbf{u}(0)\|^2) = O\left(n\tau^2\right)$. Then by Markov's inequality, $\|\mathbf{u}(0)\|^2 \le n\tau^2/\delta^2$ with probability $1 - \delta$. □

Lemma 12. *(Based on Theorem 3.1 (Arora et al., 2019))*[2] *Fix $\epsilon > 0$ and $\delta \in (0,1)$ and assume $m \ge \Omega(\frac{L^6}{\epsilon^4}log(\frac{L}{\delta}))$. Then for any pair of inputs $\mathbf{x}_i, \mathbf{x}_j$ such that $\|\mathbf{x}_i\| \le 1, \|\mathbf{x}_j\| \le 1$ with probability $1 - \delta$ we have*

$$\left|\frac{1}{m}H_{ij}(0) - \frac{1}{m}H^{\infty}_{ij}\right| \le (L+1)\epsilon$$

Lemma 13. *(Based on Theorem 5c (Allen-Zhu et al., 2019)) Let $W^{(0)}, A, B$ be at random initialization. For any pair of inputs $\mathbf{x}_i, \mathbf{x}_j$ and parameter $\omega \le O(\frac{1}{L^9 log^{3/2}m})$ with probability at least $1 - e^{-\Omega(m\omega^{2/3}L)}$ over $W^{(0)}, A, B$ with $\left\|W^{(0)} - W^{(t)}\right\|_2 \le \omega$ we have*

$$|H_{ij}(t) - H_{ij}(0)| \le O(\sqrt{\log m}\cdot\omega^{1/3}L^3)\sqrt{H_{i,i}(0)H_{j,j}(0)} \tag{26}$$

Lemma 14. *Let $\hat{\delta} \in (0,1]$ and $W^{(0)}, A, B$ be at random initialization. Then, for $m \ge \Omega\left(\frac{n^{24}L^{12}\log^5 m}{\delta^8\tau^6}\right)$ and parameter $\omega = O\left(\frac{n^3}{\delta\tau\sqrt{m}}\log m\right)$ with probability of at least $1 - \hat{\delta}$ over $W^{(0)}, A, B$ with $\left\|W^{(0)} - W^{(t)}\right\|_2 \le \omega$ it holds that*

1. $\|H(t) - H(0)\|_2 \le O(\frac{n^3 log^{5/6}m}{\delta\tau})m^{5/6}$
2. $\|H(0) - H^\infty\|_2 \le O(\frac{\delta^2 m\tau^3}{n^6})$
3. $\|H^\infty - H(t)\|_2 \le O(\frac{n^3 log^{5/6}m}{\delta\tau})m^{5/6} + O(\frac{\delta^2 m\tau^3}{n^6}) \le O(\frac{\delta^2 m\tau^3}{n^6})$

[2] The formulation given in (Arora et al., 2019) considers training w.r.t all layers. The proof can be extended trivially to the case where the first and last layers are held fixed.

Proof. We prove the first claim. Then, the second claim is obtained by plugging m into Lemma 12. The third claim is a direct consequence of the two claims using triangle inequality.

By the definition of $H_{ij}(0)$ we have that

$$\begin{aligned}\sqrt{H_{ii}(0)} &= \sqrt{\left\langle \frac{\partial u_i(0)}{\partial W}, \frac{\partial u_i(0)}{\partial W} \right\rangle} \\ &\leq \sum_{l=1}^{L} \left\| \frac{\partial u_i(0)}{\partial W_l} \right\| = \sum_{l=1}^{L} \left\| h_{i,l-1} B D_{i,L}^{(0)} W_L^{(0)} D_{i,L-1}^{(0)} W_{L-1}^{(0)} \cdots D_{i,L+1}^{(0)} W_{l+1}^{(0)} D_{i,l}^{(0)} \right\| \\ &\leq \sum_{l=1}^{L} \| h_{i,l-1} \| \left\| B D_{i,L}^{(0)} W_L^{(0)} D_{i,L-1}^{(0)} W_{L-1}^{(0)} \cdots D_{i,L+1}^{(0)} W_{l+1}^{(0)} D_{i,l}^{(0)} \right\| \leq O(L\sqrt{m}\tau)\end{aligned}$$

where the last inequality is obtained by applying Lemma 8 and Lemma 10. Applying the obtained bound for $H_{ii}(0)$ and $H_{jj}(0)$ yields a bound for $|H_{ij}(t) - H_{ij}(0)|$, using (26). Finally, $\|H(t) - H(0)\| \leq O(\frac{n^3 log^{5/6} m}{\delta\tau}) m^{5/6}$. □

E. Experiment setup

Below we provide our experimental setup for all the figures in the paper.

Figure 1. Experiments are run with input data in $\mathbb{S}^1$ drawn from a uniform (top plots) and non-uniform (bottom plots) distributions, where the latter densities are of ratio $1:40$. The target function is $y(x) = 0.4\cos(16x) + \cos(x)$. The number of training points is $n = 10000$ and batch size is 100. The network includes $L = 10$ fully connected layers, each with $m = 256$ hidden units. The weights are initialized with normal distribution with standard deviation $\tau = 0.1$, and the learning rate is $\eta = 0.001$.

Figure 2. Eigenfunctions are computed with $n = 2,933$ data points in $\mathbb{S}^1$.

Figure 3. Local frequencies are computed with $n = 1,467$ data points in $\mathbb{S}^1$.

Figure 4. Eigenvalues are computed with $n = 50,000$ data points in $\mathbb{S}^1$.

Figure 5. Eigenvalues are computed with $n = 12,567$ data points in $\mathbb{S}^1$.

Figure 6. Eigenvectors are computed numerically using $n = 10,000$ data points in $\mathbb{S}^1$ drawn from a piecewise constant distribution with densities proportional to $(11, 1, 3)$.

Figure 7. Convergence times are measured by training a two-layer network with bias. The weights of the second layer are set randomly to -1 or 1 (with probability 0.5) and remain fixed throughout training. The bias is initialized to zero. The network parameters are set to $m = 4000$, $\eta = 0.004$, $n = 734$, and $\tau = 0.2$. Convergence for region R_j is declared when $\frac{1}{2|R_j|} \sum_{i \in R_j}^{n} \left(f(x_i; w) - u_i\right)^2 < \frac{\delta}{n}$ with $\delta = 0.05$.

Figure 8. Eigenvectors are computed with $n = 9,926$ data points in $\mathbb{S}^2$.

Figure 9. We used the same setup as in Figure 7 with the parameters: $m = 8000$, $tau = 0.2$, and $\eta = 0.004$. Here n varies between the three plots. We sampled 300 points from a uniform distribution on one hemisphere, and $300p_2/p_1$ points on the other hemisphere, where $p_2/p_1 \in \{2, 3, 4\}$.

Figure 10. Eigenvectors are computed with $n = 1257$ data points in $\mathbb{S}^1$.

Figure 11. Here we compare the number of iterations needed for a deep FC network to converge the number of iterations predicted by the eigenvalue of the corresponding NTK. We used $m = 256$, $\eta = 0.05$ and $\delta = 0.05$. The corresponding NTK was calculated in the $\mathbb{S}^1$ with $n = 630$ points and in $\mathbb{S}^2$ with $n = 1,000$ points, both drawn from a uniform distribution. Note that the plot for $\mathbb{S}^2$ appears on the left and the one for $\mathbb{S}^1$ on the right.

Figure 12. Here, we calculate the eigenvalues of NTK for FC networks with $3 \leq L \leq 50$ layers for data distributed uniformly in $\mathbb{S}^1$ (left) and $\mathbb{S}^2$ (right). The NTK was calculated with $n = 16,383$ and $n = 20,000$ data points in $\mathbb{S}^1$ and $\mathbb{S}^2$, respectively.

References

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252, 2019.

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