A. Proofs for Sec 4.1

Proposition 2. *For a fixed unit vector $\mathbf{z}^{(0)}$, fixed input data $\hat{\mathbf{x}}$ and a network of depth L at random initialization, with a Lipschitz nonlinearity σ, and in the limit $n_1, ..., n_{L-1} \to \infty$, $\mathbf{J}(\hat{\mathbf{x}})\mathbf{z}^{(0)}$ has the following recursion with $\mathbf{z}_i^{(\ell)} = \hat{z}^{(\ell)}$:*

$$
\begin{aligned}
\hat{z}^{(1)} &= \sigma'(a)b \quad (a,b) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \frac{\|\hat{\mathbf{x}}\|_2^2}{n_0}, & \frac{\hat{\mathbf{x}}^T\mathbf{z}^{(0)}}{n_0} \\ \frac{\hat{\mathbf{x}}^T\mathbf{z}^{(0)}}{n_0}, & \frac{\|\mathbf{z}^{(0)}\|_2^2}{n_0} \end{bmatrix}\right), \\
\hat{z}^{(\ell+1)} &= \sigma'(a)b \\
&\qquad (a,b) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbb{E}[(\hat{\alpha}^{(\ell)})^2], & \mathbb{E}[\hat{\alpha}^{(\ell)}\hat{z}^{(\ell)}] \\ \mathbb{E}[\hat{\alpha}^{(\ell)}\hat{z}^{(\ell)}], & \mathbb{E}[(\hat{z}^{(\ell)})^2] \end{bmatrix}\right), \\
\tilde{\mathbf{z}}_i^{(L)} &= \hat{z}^{(L)} \sim \mathcal{N}\left(0, \mathbb{E}[(\hat{z}^{(L-1)})^2]\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{\alpha}^{(1)} &= \sigma(a) \quad a \sim \mathcal{N}\left(0, \frac{\|\hat{\mathbf{x}}\|_2^2}{n_0}\right), \\
\hat{\alpha}^{(\ell+1)} &= \sigma(a) \quad a \sim \mathcal{N}\left(0, \mathbb{E}[(\hat{\alpha}^{(\ell)})^2]\right).
\end{aligned}
$$

Proof. We will prove this by induction for $\ell = 1, ..., L-1$.

Basic Step

$$
\mathbf{z}_i^{(1)} = \sigma'\left(\frac{1}{\sqrt{n_0}}(\mathbf{W}_i^{(0)})^T\hat{\mathbf{x}}\right)\frac{1}{\sqrt{n_0}}(\mathbf{W}_i^{(0)})^T\mathbf{z}^{(0)}
$$

Notice that $\mathbf{W}_i^{(0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n_0})$. Thus, we have the following:

$$
\begin{aligned}
a &= \frac{1}{\sqrt{n_0}}(\mathbf{W}_i^{(0)})^T\hat{\mathbf{x}} \sim \mathcal{N}\left(0, \frac{\|\hat{\mathbf{x}}\|_2^2}{n_0}\right) \\
b &= \frac{1}{\sqrt{n_0}}(\mathbf{W}_i^{(0)})^T\mathbf{z}^{(0)} \sim \mathcal{N}\left(0, \frac{\|\mathbf{z}^{(0)}\|_2^2}{n_0}\right)
\end{aligned}
$$

a and b are not independent:

$$
\mathbb{E}[ab] = \mathbb{E}[(\frac{1}{\sqrt{n_0}}(\mathbf{W}_i^{(0)})^T\hat{\mathbf{x}})(\frac{1}{\sqrt{n_0}}(\mathbf{W}_i^{(0)})^T\mathbf{z}^{(0)})] = \frac{1}{n_0}\hat{\mathbf{x}}^T\,\mathbb{E}[(\mathbf{W}_i^{(0)})(\mathbf{W}_i^{(0)})^T]\mathbf{z}^{(0)} = \frac{1}{n_0}\hat{\mathbf{x}}^T\mathbf{I}_{n_0}\mathbf{z}^{(0)} = \frac{\hat{\mathbf{x}}^T\mathbf{z}^{(0)}}{n_0}
$$

Note that the result is independent of the index i, we can define $\hat{z}^{(1)} = \mathbf{z}_i^{(1)}$. Therefore, the base step has been proven.

Inductive Step

$$
\mathbf{z}_i^{(\ell+1)} = \sigma'(\frac{1}{\sqrt{n_\ell}}(\mathbf{W}_i^{(\ell)})^T\tilde{\alpha}^{(\ell)}(\hat{\mathbf{x}}))\frac{1}{\sqrt{n_\ell}}(\mathbf{W}_i^{(\ell)})^T\mathbf{z}^{(\ell)}
$$

Then,

$$
a = \frac{1}{\sqrt{n_\ell}}(\mathbf{W}_i^{(\ell)})^T\tilde{\alpha}^{(\ell)}(\hat{\mathbf{x}}) \sim \mathcal{N}(\mathbf{0}, \frac{1}{n_\ell}\sum_{i=0}^{n_\ell}((\tilde{\alpha}^{(\ell)}(\hat{\mathbf{x}})_i)^2)
$$

With $n_1, ..., n_\ell \to \infty$, Var$(a) = \mathbb{E}[(\hat{\alpha}^{(\ell)})^2]$. Similarly,

$$
\begin{aligned}
b &= \frac{1}{\sqrt{n_\ell}}(\mathbf{W}_i^{(\ell)})^T\mathbf{z}^{(\ell)} \\
b &\sim \mathcal{N}(0, \mathbb{E}[(\hat{z}^{(\ell)})^2]) \quad \text{if } n_1, ..., n_\ell \to \infty
\end{aligned}
$$

On the other hand,

$$\mathbb{E}[ab] = \mathbb{E}[(\frac{1}{\sqrt{n_\ell}}(\mathbf{W}_i^{(\ell)})^T\tilde{\alpha}^{(\ell)}(\mathbf{x}))(\frac{1}{\sqrt{n_\ell}}(\mathbf{W}_i^{(\ell)})^T\mathbf{z}^{(\ell)})] = \frac{1}{n_\ell}(\tilde{\alpha}^{(\ell)}(\mathbf{x}))^T\,\mathbb{E}[(\mathbf{W}_i^{(\ell)})(\mathbf{W}_i^{(\ell)})^T]\mathbf{z}^{(\ell)} = \frac{1}{n_\ell}(\tilde{\alpha}^{(\ell)}(\mathbf{x}))^T\mathbf{z}^{(\ell)}$$
$$= \mathbb{E}[\hat{\alpha}^{(\ell)}\hat{z}^{(\ell)}] \quad \text{if } n_1, ..., n_\ell \to \infty$$

The recursive definition is now proven up to layer $\ell - 1$. Now let's look at the last layer.

$$\tilde{\mathbf{z}}_i^{(L)} = \frac{1}{\sqrt{n_{L-1}}}\mathbf{W}_i^{(L-1)}\mathbf{z}^{(L-1)}$$

By similar arguments as before, it is easy to show that with $n_1, ..., n_{L-1} \to \infty$, $\tilde{\mathbf{z}}_i^{(L)} \sim \mathcal{N}(0, \mathbb{E}[(\hat{z}^{(L-1)})^2])$. This concludes the proof. □

Theorem 1. *For any data point* $\mathbf{x}_i$*,* $i \in [1, .., n]$*, with probability at least* $1 - O(n)e^{-O(n_0)}$*,*

$$\|\mathbf{J}(\mathbf{x}_i)\|_{op} \leq c\sqrt{n_0\tau}$$

where c *is a constant and*

$$\tau = \sup_{\mathbf{x_i}\in\hat{\mathbf{X}},\ \|\mathbf{z}^{(0)}\|_2=1} \mathbb{E}[(\hat{z}^{(L-1)})^2|\mathbf{z}^{(0)}, \mathbf{x}_i]$$

Proof. For a fixed unit vector $\mathbf{z}^{(0)}$ and fixed input $\hat{\mathbf{x}}$, we know that based on Proposition 2, $\tilde{\mathbf{z}}_i^{(L)} \sim \mathcal{N}(0, \mathbb{E}[(\hat{z}^{(L-1)})^2|\mathbf{z}^{(0)}, \hat{\mathbf{x}}])$. Define z as

$$z = \frac{1}{\mathbb{E}[(\hat{z}^{(L-1)})^2|\mathbf{z}^{(0)}, \hat{\mathbf{x}}]}\|\tilde{\mathbf{z}}^{(L)}\|_2^2 = \chi^2_{n_0}$$

First, notice that we can have the following tail bound for chi-square distribution (for instance, (Kolar & Liu, 2012))

$$\Pr[|z/n_0 - 1| \geq \epsilon] \leq \exp(-\frac{3}{16}n_0\epsilon^2)$$

when $\epsilon \in [0, 1/2)$. In this case, let $\epsilon = \frac{1}{3}$. Consider a subset of coordinates M with cardinality $|M| \leq O(n_0)$ (Allen-Zhu et al., 2018). Taking the ϵ ball $\mathcal{B}$ of this subspace with $\epsilon = 1/3$, we know what

$$|\mathcal{B}| \leq 7^{|M|} = e^{|M|ln7} = e^{O(n_0)}$$

Then, taking the union bound for all unit vectors in $\mathcal{B}$, we know that

$$\forall \mathbf{z}_0 \in \mathcal{B} \quad \bigcup_{z_0}\Pr[|z/n_0 - 1| \geq \frac{1}{3}]$$
$$\leq \exp(-\frac{1}{48}n_0)\exp(O(n_0)) \leq \exp(-O(n_0))$$

Therefore, by the ϵ-net argument (Tao, 2012), for any unit vector $\mathbf{u}$ with only non-zero entries in M, we have with probability $1 - \exp(-O(n_0))$,

$$\|\mathbf{J}(\hat{\mathbf{x}})\mathbf{u}\|_2^2 \leq 2n_0\tau\|\mathbf{u}\|_2^2 = C^2\|\mathbf{u}\|_2^2$$

For any arbitrary vector $\mathbf{v}$, we can decompose it in the following way: $\mathbf{v} = \mathbf{u}_1 + \mathbf{u}_2 + ... + \mathbf{u}_K$ with $K = O(1)$ where each $\mathbf{u}_i$ comes from a different non-overlapping coordinate set M.

$$\|\mathbf{J}(\hat{\mathbf{x}})\mathbf{v}\|_2 \leq C\sum_{i=1}^{K}\|\mathbf{u}_i\|_2 \leq C\sqrt{K}(\sum_{i=1}^{K}\|\mathbf{u}_i\|_2^2)^{1/2}$$
$$\leq O(1)C\|\mathbf{v}\|.$$

Thus, with probability at least $1 - O(1)\exp(-O(n_0))$,

$$\|\mathbf{J}(\hat{\mathbf{x}})\|_{op} \leq O(1)C = O(1)\sqrt{2n_0\tau}$$
$$= c\sqrt{n_0\tau},$$

where c is a constant. Taking the union bound over all the data points concludes the proof. □

B. Proofs for Sec 4.2

Lemma 5. *Under the setting in Section 3.1 with* sigmoid *as the activation function,*

$$\Theta_{\infty}^{(L)}(\mathbf{x},\mathbf{x}) \geq \frac{1}{4}$$

Proof. For any ℓ, we have

$$\begin{aligned}
\Theta_{\infty}^{(\ell+1)}(\mathbf{x},\mathbf{x}) &= \Theta_{\infty}^{(\ell)}(\mathbf{x},\mathbf{x})\dot{\Sigma}^{(\ell+1)}(\mathbf{x},\mathbf{x}) + \Sigma^{(\ell+1)}(\mathbf{x},\mathbf{x}) \\
&\geq \Sigma^{(\ell+1)}(\mathbf{x},\mathbf{x}) \\
&= \mathbb{E}_{g\sim\mathcal{N}(0,\Sigma^{(\ell)})}[\sigma(g(\mathbf{x}))^2] \\
&= \mathbb{E}_{g\sim\mathcal{N}(0,\Sigma^{(\ell)})}\left[\left(\sigma(g(\mathbf{x})) - \frac{1}{2}\right)^2\right] + \frac{1}{4} \\
&\geq \frac{1}{4}
\end{aligned}$$

where $\sigma(f(\mathbf{x})) - \frac{1}{2}$ moves sigmoid function to the origin such that it is an odd function. □

C. Proofs for Sec 4.3

C.1. Main Lemmas

Lemma 2. *Suppose there is a 2-layer network. If the activation function is $\sigma(x) = \alpha x$, $n = n_0$ and the data matrix is full rank. Then at NTK limit, $\mathbf{J}_{\infty}(\mathbf{x}) = \mathbf{I}_{n_0}$.*

Proof.

$$\mathbf{J}_{\infty}(\mathbf{x}) = \frac{2\alpha^2}{n_0}(\hat{\mathbf{X}} - f_0(\hat{\mathbf{X}}))(\frac{2\alpha^2}{n_0}\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T + J_0(\mathbf{x})$$

Notice that $\mathbf{J}_0(\mathbf{x}) = \alpha\frac{1}{\sqrt{n_1}}\frac{1}{\sqrt{n_0}}\mathbf{W}^{(1)}\mathbf{W}^{(0)}$ and $f_0(\hat{\mathbf{X}}) = \alpha\frac{1}{\sqrt{n_1}}\frac{1}{\sqrt{n_0}}\mathbf{W}^{(1)}\mathbf{W}^{(0)}\hat{\mathbf{X}} = \mathbf{J}_0(x)\hat{\mathbf{X}}$.

$$\begin{aligned}
\mathbf{J}_{\infty}(\mathbf{x}) &= \mathbf{J}_0(\mathbf{x}) - f_0(\hat{\mathbf{X}})(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T + \hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T \\
&= \alpha\frac{1}{\sqrt{n_1}}\frac{1}{\sqrt{n_0}}\mathbf{W}^{(1)}\mathbf{W}^{(0)} - \alpha\frac{1}{\sqrt{n_1}}\frac{1}{\sqrt{n_0}}\mathbf{W}^{(1)}\mathbf{W}^{(0)}\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T + \hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T \\
&= \alpha\frac{1}{\sqrt{n_1}}\frac{1}{\sqrt{n_0}}\mathbf{W}^{(1)}\mathbf{W}^{(0)} - \alpha\frac{1}{\sqrt{n_1}}\frac{1}{\sqrt{n_0}}\mathbf{W}^{(1)}\mathbf{W}^{(0)}\mathbf{I}_{n_0} + \mathbf{I}_{n_0} \\
&= \mathbf{I}_{n_0}
\end{aligned}$$

□

Lemma 3. *Suppose there is a 2-layer network with activation function $\sigma(x) = \alpha x$ and given initial weights $\mathbf{W}^{(1)} \in \mathbb{R}^{n_0\times n_1}$, $\mathbf{W}^{(0)} \in \mathbb{R}^{n_1\times n_0}$. If the data matrix is full rank with $n \leq n_0$, then, at the NTK limit ($n_1 \to \infty$), $\mathbf{J}_{\infty}(\mathbf{x})$ has eigenvalue 1 with multiplicity at least n. If at the NTK limit, α is chosen such that $\|\mathbf{J}_0(\mathbf{x})\|_{op} < 1$, then the multiplicity is exactly n and 1 is the largest eigenvalue norm.*

Proof. Based on the proof of last section, we know that

$$\begin{aligned}
\mathbf{J}_{\infty}(\mathbf{x}) &= \mathbf{J}_0(\mathbf{x}) - f_0(\hat{\mathbf{X}})(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T + \hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T \\
&= \mathbf{J}_0(\mathbf{x}) - \mathbf{J}_0(\mathbf{x})\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T + \hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T
\end{aligned}$$

In this case,

$$\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T = V\Sigma V^T$$

where V is an orthgonal matrix and

$$\Sigma = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \multicolumn{4}{c}{\dots\dots\dots\dots\dots} \\ 0 & \dots & 1 & 0 \\ \multicolumn{4}{c}{\dots\dots\dots\dots\dots} \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

So

$$\begin{aligned} \mathbf{J}_\infty(\mathbf{x}) &= \mathbf{J}_0(\mathbf{x})(\mathbf{I}_{n_0} - V\Sigma V^T) + V\Sigma V^T \\ &= \alpha \frac{1}{\sqrt{n_1}} \frac{1}{\sqrt{n_0}} \mathbf{W}^{(1)}\mathbf{W}^{(0)}(\mathbf{I}_{n_0} - V\Sigma V^T) + V\Sigma V^T \end{aligned}$$

Interestingly, $(I_d - V\Sigma V^T)$ and $V\Sigma V^T$ contain orthogonal eigenvectors. For convenience, let $\{v_i\}_{i=1}^n$ be the set of eigenvectors of $V\Sigma V^T$ with eigenvalue 1. Furthermore, let $V_\| = \text{span}(\{v_i\}_{i=1}^n)$ and $V_\perp = \text{span}(\{v_i\}_{i=1}^n)^\perp$. Because we are in the linear region, $\mathbf{J}_\infty(\mathbf{x})$ and $\mathbf{J}_0(\mathbf{x})$ do not depend on $\mathbf{x}$. We'll use $\mathbf{J}_\infty$ to refer $\mathbf{J}_\infty(\mathbf{x})$ and $\mathbf{J}_0$ as $\mathbf{J}_0(\mathbf{x})$.

- For any vector $v^\| \in V_\|$,

  $$\mathbf{J}_0(\mathbf{I}_{n_0} - V\Sigma V^T)v^\| = 0$$

  and

  $$\mathbf{J}_\infty v^\| = v^\|$$

  Thus, all vectors in $\{v_i\}_{i=1}^n$ are eigenvetors of $\mathbf{J}_\infty$ with eigenvalue 1 regardless of the choice of α.

- On the other hand, let v be any complex vector such that

  $$v = \text{Re}(v) + i\text{Im}(v)$$

  If v is an eigenvector of $\mathbf{J}_\infty$ with eigenvalue $\lambda = a + ib$, then

  $$\begin{aligned} \mathbf{J}_\infty \text{Re}(v) &= a\text{Re}(v) - b\text{Im}(v) \\ \mathbf{J}_\infty \text{Im}(v) &= b\text{Re}(v) + a\text{Im}(v) \end{aligned}$$

  Let's first decompose Re(v) and Im(v).

  $$\begin{aligned} \text{Re}(v) &= v_r^\perp + v_r^\| \\ \text{Im}(v) &= v_i^\perp + v_i^\| \end{aligned}$$

  where $v_r^\perp, v_i^\perp \in V_\perp$ and $v_r^\|, v_i^\| \in V_\|$.

  $$\begin{aligned} \mathbf{J}_\infty(v_r^\perp + v_r^\|) &= J_0 v_r^\perp + v_r^\| = (av_r^\perp - bv_i^\perp) + (av_r^\| - bv_i^\|) \\ \mathbf{J}_\infty(v_i^\perp + v_i^\|) &= J_0 v_i^\perp + v_i^\| = (bv_r^\perp + av_i^\perp) + (bv_r^\| + av_i^\|) \end{aligned}$$

  By adding and subtracting two equations,

  $$\begin{aligned} \mathbf{J}_0(v_r^\perp + v_i^\perp) + v_r^\| + v_i^\| &= \Big[(a+b)v_r^\perp + (a-b)v_i^\perp\Big] + \Big[(a+b)v_r^\| + (a-b)v_i^\|\Big] \\ \mathbf{J}_0(v_r^\perp - v_i^\perp) + v_r^\| - v_i^\| &= \Big[(a-b)v_r^\perp - (a+b)v_i^\perp\Big] + \Big[(a-b)v_r^\| - (a+b)v_i^\|\Big] \end{aligned}$$

  When α is chosen such that $\|\mathbf{J}_0\| < 1$,

  $$\begin{aligned} \|(a+b)v_r^\perp + (a-b)v_i^\perp\|_2 &< \|v_r^\perp + v_r^\|\|_2 \\ \|(a-b)v_r^\perp - (a+b)v_i^\perp\|_2 &< \|v_r^\perp - v_r^\|\|_2 \end{aligned}$$

Then,

$$(a^2+b^2)\|v_r^\perp\|_2^2 + (a^2+b^2)\|v_i^\perp\|_2^2 < \|v_r^\perp\|_2^2 + \|v_i^\perp\|_2^2$$
$$|\lambda|^2 = a^2+b^2 < 1$$

This suggests that any complex eigenvector with components from $V_\perp$ would have eigenvalue with norm smaller than 1.

□

Lemma 4. *Suppose there is a 2-layer network with activation function $\sigma(x) = \alpha x + \beta$, given initial weights $\mathbf{W}^{(1)} \in \mathbb{R}^{n_0 \times n_1}$, $\mathbf{W}^{(0)} \in \mathbb{R}^{n_1 \times n_0}$ and every data point has the same norm r (i.e. $\forall i \in [n]$ $\|\mathbf{x}\|_2 = r$). If the data matrix is full rank with $n \le n_0$, then, at the NTK limit $n_1 \to \infty$, $\mathbf{J}_\infty(\mathbf{x})$ has eigenvalues 1 with multiplicity at least $n-1$. If at the NTK limit, α and β are chosen such that*

$$\|\mathbf{J}_0(\mathbf{x})\|_{op} = 1 - \Delta, \quad \left\|\frac{1}{\sqrt{n_1}}\mathbf{W}^{(1)}\mathbf{1}_{n_1}\right\|_2 < \frac{\beta n_0 \Delta}{2r\alpha^2},$$

where $0 < \Delta \le 1$, then the multiplicity is exactly $n-1$ and 1 is the largest eigenvalue norm.

Proof. First of all, let $\mathbf{B}$ be an all-one matrix

$$\begin{aligned}
\mathbf{J}_\infty(\mathbf{x}) &= \Big(\hat{\mathbf{X}} - f_0(\hat{\mathbf{X}})\Big)\tilde{\mathbf{K}}^{-1}\frac{\partial k_x}{\partial \mathbf{x}} + \mathbf{J}_0(\mathbf{x}) \\
&= \Big(\hat{\mathbf{X}} - f_0(\hat{\mathbf{X}})\Big)\Big(\frac{2\alpha^2}{n_0}\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \beta^2\mathbf{B}\Big)^{-1}\Big(\frac{2\alpha^2}{n_0}\hat{\mathbf{X}}^T\Big) + \mathbf{J}_0(\mathbf{x}) \\
&= \Big(\hat{\mathbf{X}} - f_0(\hat{\mathbf{X}})\Big)\Big(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \frac{n_0\beta^2}{2\alpha^2}\mathbf{B}\Big)^{-1}\hat{\mathbf{X}}^T + \mathbf{J}_0(\mathbf{x}) \\
&= \mathbf{J}_0(\mathbf{x}) + \hat{\mathbf{X}}\Big(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \frac{n_0\beta^2}{2\alpha^2}\mathbf{B}\Big)^{-1}\hat{\mathbf{X}}^T - \big(\frac{\alpha}{\sqrt{n_1 n_0}}\mathbf{W}^{(1)}\mathbf{W}^{(0)}\hat{\mathbf{X}} + \beta\frac{1}{\sqrt{n_1}}\mathbf{W}^{(1)}\mathbf{1}_{n_1}\mathbf{1}_n^T\big)\Big(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \frac{n_0\beta^2}{2\alpha^2}\mathbf{B}\Big)^{-1}\hat{\mathbf{X}}^T \\
&= \mathbf{J}_0(\mathbf{x}) + \hat{\mathbf{X}}\Big(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \frac{n_0\beta^2}{2\alpha^2}\mathbf{B}\Big)^{-1}\hat{\mathbf{X}}^T - \big(\mathbf{J}_0(\mathbf{x})\hat{\mathbf{X}} + \beta\frac{1}{\sqrt{n_1}}\mathbf{W}^{(1)}\mathbf{1}_{n_1}\mathbf{1}_n^T\big)\Big(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \frac{n_0\beta^2}{2\alpha^2}\mathbf{B}\Big)^{-1}\hat{\mathbf{X}}^T
\end{aligned}$$

Because in the linearized region, $\mathbf{J}_\infty(\mathbf{x})$ and $\mathbf{J}_0(\mathbf{x})$ do not depend on $\mathbf{x}$. We'll use $\mathbf{J}_\infty$ to refer $\mathbf{J}_\infty(\mathbf{x})$ and $\mathbf{J}_0$ as $\mathbf{J}_0(\mathbf{x})$. For simplicity, we'll also use $c = \frac{n_0\beta^2}{2\alpha^2}$.

Based on Lemma 6,

$$\hat{\mathbf{X}}\Big(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + c\mathbf{B}\Big)^{-1}\hat{\mathbf{X}}^T = V\Lambda V^T$$

where $\Lambda = \text{diag}(\underbrace{1,...,1}_{n-1}, \hat{\lambda}, \underbrace{0,...,0}_{n_0-n})$ where $0 < \hat{\lambda} < 1$. Now,

$$\mathbf{J}_\infty = \mathbf{J}_0(I_{n_0} - V\Lambda V^T) + V\Lambda V^T - \beta\frac{1}{\sqrt{n_1}}W^{(1)}\mathbf{1}_{n_1}\mathbf{1}_n^T\Big(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + c\mathbf{B}\Big)^{-1}\hat{\mathbf{X}}^T$$

From Corollary 7, we know that the following two vectors are eigenvectors of $V\Lambda V^T$ with eigenvalue $\hat{\lambda}$,

$$\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + c\mathbf{B})^{-1}\mathbf{1}_n \qquad \hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\mathbf{1}_n$$

Furthermore,

$$\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + c\mathbf{B})^{-1}\mathbf{1}_n = \hat{\lambda}\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\mathbf{1}_n$$

And

$$\hat{\lambda} = \frac{1}{1+cg}$$

where

$$g = \text{trace}(\mathbf{B}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}) = \|\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\mathbf{1}_n\|_2^2$$

Let $\hat{u}$ be a rescaled unit vector of $\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\mathbf{1}_n$, then

$$\mathbf{J}_\infty \hat{u} = \mathbf{J}_0(1-\hat{\lambda})\hat{u} + \hat{\lambda}\hat{u} - \sqrt{g}\hat{\lambda}\beta\frac{1}{\sqrt{n_1}}W^{(1)}\mathbf{1}_{n_1}\hat{u}$$

$$\begin{aligned}
\|\mathbf{J}_\infty \hat{u}\|_2 &= \|\mathbf{J}_0(1-\hat{\lambda})\hat{u} + \hat{\lambda}\hat{u} - \sqrt{g}\hat{\lambda}\beta\frac{1}{\sqrt{n_1}}\mathbf{W}^{(1)}\mathbf{1}_{n_1}\hat{u}\|_2 \\
&\leq \|\mathbf{J}_0\|_{op}\|(1-\hat{\lambda})\hat{u}\|_2 + \|\hat{\lambda}\hat{u}\|_2 + \|\sqrt{g}\hat{\lambda}\beta\frac{1}{\sqrt{n_1}}\mathbf{W}^{(1)}\mathbf{1}_{n_1}\hat{u}\|_2 \\
&= (1-\hat{\lambda})\|\mathbf{J}_0\|_{op} + \hat{\lambda} + \sqrt{g}\hat{\lambda}\|\beta\frac{1}{\sqrt{n_1}}\mathbf{W}^{(1)}\mathbf{1}_{n_1}\|_2 \\
&< (1-\hat{\lambda})(1-\Delta) + \hat{\lambda} + \sqrt{g}\hat{\lambda}\frac{\beta^2 n_0\Delta}{2r\alpha^2} \\
&\leq (1-\hat{\lambda})(1-\Delta) + \hat{\lambda} + g\hat{\lambda}\frac{\beta^2 n_0\Delta}{2\alpha^2} \qquad (Lemma\ 9) \\
&= \frac{(1-\Delta)cg + 1 + \frac{g\beta^2 n_0 \Delta}{2\alpha^2}}{1+cg} = 1
\end{aligned}$$

Therefore, $\mathbf{J}_\infty$ will shrink every vectors orthogonal to the eigenvectors in V with eigenvalue 1. By the same arguments in the proof of Lemma 3, we can conclude the proof. □

C.2. Useful Lemmas

Lemma 6. *Suppose $\mathbf{X} \in \mathbb{R}^{k\times m}$ is a full-rank matrix with $k \geq m$ and $m \geq 2$. Let c be an arbitrary positive constant and $\mathbf{B}$ an all-one matrix. Consider the following real symmetric matrix,*

$$\mathbf{X}(\mathbf{X}^T\mathbf{X} + c\mathbf{B})^{-1}\mathbf{X}^T$$

It can be characterized by having eigenvalue 1 with multiplicity $m-1$, eigenvalue 0 with multiplicity $k-m$ and another eigenvalue λ such that $0 < \lambda < 1$.

Proof. By (Miller, 1981), if P and $P+Q$ are invertible, and Q has rank 1, then let $g' = \text{trace}(QP^{-1})$, we know that $g' \neq 1$, and

$$(P+Q)^{-1} = P^{-1} - \frac{1}{1+g'}P^{-1}QP^{-1}$$

First of all, it is easy to see that $(\mathbf{X}^T\mathbf{X} + c\mathbf{B})^{-1}$ is invertible. This is because $\mathbf{X}^T\mathbf{X}$ is positive definite and cB is positive semi-definite.

Since B is a rank one matrix,

$$(\mathbf{X}^T\mathbf{X} + c\mathbf{B})^{-1} = \underbrace{(\mathbf{X}^T\mathbf{X})^{-1}}_{I_1} - \underbrace{\frac{c}{1+cg}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{B}(\mathbf{X}^T\mathbf{X})^{-1}}_{I_2}$$

where $g = \text{trace}(\mathbf{B}(\mathbf{X}^T\mathbf{X})^{-1})$.

Let's consider the singular value decomposition of $\mathbf{X}^T = U\Sigma V^T$

- $\mathbf{X}^T\mathbf{X} = U\Sigma^2U^T$ and $(\mathbf{X}^T\mathbf{X})^{-1} = U\Sigma^{-2}U^T$. So

$$\mathbf{X}I_1\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = V\Sigma U^T U\Sigma^{-2}U^T U\Sigma V^T = V\Lambda_m V^T$$

where $\Lambda_m = \text{diag}(\underbrace{1,...,1}_{m}, \underbrace{0,...,0}_{k-m})$

- $$\mathbf{X}^T I_2 \mathbf{X} = \frac{c}{1+cg} M = \frac{c}{1+cg} \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{B}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

  The first thing to notice is that $\mathbf{B} = \mathbf{1}\mathbf{1}^T$ where $\mathbf{1}$ is a vector of ones. Therefore,

  $$M = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{B}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{1}\mathbf{1}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{a}\mathbf{a}^T$$

  where $\mathbf{a} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{1}$.

  This implies that M is a rank one matrix with singular value $\|\mathbf{a}\|^2$. But we also know the following:

  $$\begin{aligned}\|\mathbf{a}\|^2 = \mathbf{a}^T\mathbf{a} = \text{trace}(\mathbf{a}\mathbf{a}^T) = \text{trace}(M)&\\ &= \text{trace}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{B}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \text{trace}(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{B}(\mathbf{X}^T\mathbf{X})^{-1})\\ &= \text{trace}(\mathbf{B}(\mathbf{X}^T\mathbf{X})^{-1}) = g > 0\end{aligned}$$

  The last strict inequality comes from the fact that X is full rank so that $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$ has no zero singular value. Furthermore,

  $$\begin{aligned}\mathbf{X}I_1\mathbf{X}^T\mathbf{a} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{a}\\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{1} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{1}\\ &= \mathbf{a}\end{aligned}$$

  Because $\mathbf{a}$ is not a zero vector, it is also one of the eigenvector of $\mathbf{X}I_1X$ with eigenvalue 1.

  And the eigenvalue of $X^T I_2 X$ is the following:

  $$0 < \frac{cg}{1+cg} < 1$$

  The inequalities comes from the fact that c is also non-negative. We'll denote $\sigma = \frac{cg}{1+cg}$. So

  $$\mathbf{X}I_2\mathbf{X}^T = \sigma\hat{\mathbf{a}}\hat{\mathbf{a}}^T$$

  where $\hat{\mathbf{a}}$ is $\mathbf{a}$ rescaled to have unit length.

Now that we have examined two parts separately. Let's put them together. For convenience, we'll also denote $\mathbf{X}(\mathbf{X}^T\mathbf{X} + c\mathbf{B})^{-1}\mathbf{X}^T = \mathbf{X}I_1\mathbf{X}^T - \mathbf{X}I_2\mathbf{X}^T = \mathbf{M}_1 - \mathbf{M}_2$.

Based on the eigen decomposition of $\mathbf{M}_1$,

$$\mathbf{M}_1 = \sum_{k=1}^{m} \mathbf{u}_k\mathbf{u}_k^T$$

with lost of generality, let's also denote $\hat{\mathbf{a}} = \mathbf{u}_1$. Now,

$$\begin{aligned}\mathbf{M}_1 - \mathbf{M}_2 &= \sum_{k=1}^{m} \mathbf{u}_k\mathbf{u}_k^T - \sigma\mathbf{u}_1\mathbf{u}_1^T\\ &= (1-\sigma)\mathbf{u}_1\mathbf{u}_1^T + \sum_{k=2}^{m} \mathbf{u}_k\mathbf{u}_k^T\end{aligned}$$

Because $0 < \sigma < 1$, $\mathbf{X}(\mathbf{X}^T\mathbf{X} + c\mathbf{B})^{-1}\mathbf{X}^T$ has eigenvalue 1 with multiplicity $m-1$, eigenvalue 0 with multiplicity $k-m$ and another eigenvalue λ such that $0 < \lambda < 1$. □

Corollary 7. Following the setup in Lemma 6, we could also know that $\mathbf{X}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{1}$ is an eigenvector with with eigenvalue λ and

$$\Big(\mathbf{X}(\mathbf{X}^T\mathbf{X} + c\mathbf{B})^{-1}\mathbf{X}^T\Big)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{1} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + c\mathbf{B})^{-1}\mathbf{1} = \lambda\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{1}$$

Corollary 8. Suppose $\mathbf{X} \in \mathbb{R}^{k \times m}$ is a full-rank matrix with $k \geq m$ and $m \geq 2$. Let c be an arbitrary non-negative constant and $\mathbf{B}$ an all-one matrix.

$$\|\mathbf{X}(\mathbf{X}^T\mathbf{X} + c\mathbf{B})^{-1}\mathbf{X}^T\|_{op} = 1$$

Remark 2. c can also takes on negative values as long as cg is not close to -1.

Lemma 9. *Suppose $\mathbf{X} \in \mathbb{R}^{k \times m}$ is a full-rank matrix with $k \geq m$ and $\mathbf{B}$ an all-one matrix. If*

$$\|\mathbf{X}_{\cdot,i}\|_2 = r \qquad \forall i \in [m]$$

Then,

$$\text{trace}(\mathbf{B}(\mathbf{X}^T\mathbf{X})^{-1}) \geq \frac{1}{r^2}$$

Proof. First of all,

$$\begin{aligned}\text{trace}(\mathbf{B}(\mathbf{X}^T\mathbf{X})^{-1}) &\geq \text{trace}(\mathbf{1}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{1}) \\ &\geq \|\mathbf{1}\|_2^2 \frac{1}{\|\mathbf{X}^T\mathbf{X}\|_{op}} = \frac{m}{\|\mathbf{X}^T\mathbf{X}\|_{op}}\end{aligned}$$

On the hand,

$$\|\mathbf{X}^T\mathbf{X}\|_{op} = \|\mathbf{X}^T\|_{op}^2 \leq \|\mathbf{X}^T\|_f^2 \leq \text{trace}(\mathbf{X}^T\mathbf{X}) \leq r^2 m$$

Therefore,

$$\text{trace}(\mathbf{B}(\mathbf{X}^T\mathbf{X})^{-1}) \geq \frac{1}{r^2}$$

□

D. Proofs for Sec 4.4

D.1. Derivation for the Approximated NTK

The closed form NTK of erf (Lee et al., 2019; Williams, 1997) can be written with the following two components:

$$\begin{aligned}\mathcal{T}(\Sigma, \text{erf}, \text{erf})(\mathbf{x}, \hat{\mathbf{x}}) &= \frac{2}{\pi}\arcsin\left(\frac{\Sigma(\mathbf{x}, \hat{\mathbf{x}})}{\sqrt{(\Sigma(\mathbf{x}, \mathbf{x}) + 0.5)(\Sigma(\hat{\mathbf{x}}, \hat{\mathbf{x}}) + 0.5)}}\right) \\ \mathcal{T}(\Sigma, \dot{\text{erf}}, \dot{\text{erf}})(\mathbf{x}, \hat{\mathbf{x}}) &= \frac{4}{\pi}\det(I + 2\Sigma)^{-\frac{1}{2}} = \frac{4}{\pi}\frac{1}{\sqrt{(1 + 2\Sigma(\mathbf{x}, \mathbf{x})(1 + 2\Sigma(\hat{\mathbf{x}}, \hat{\mathbf{x}})) - 4\Sigma(\mathbf{x}, \hat{\mathbf{x}})^2)}}\end{aligned}$$

Here, we can approximate sigmoid function σ_s by erf function:

$$\sigma_s(x) \approx \sigma_{\hat{s}}(x) = \frac{1}{2}\text{erf}(\frac{1}{2}x) + \frac{1}{2}$$

Then,

$$\begin{aligned}\mathcal{T}(\Sigma, \sigma_{\hat{s}}, \sigma_{\hat{s}})(\mathbf{x}, \hat{\mathbf{x}}) = \mathbb{E}_{u,v \sim \mathcal{N}(0,\Sigma)}[\sigma_{\hat{s}}(u)\sigma_{\hat{s}}(v)] &= \mathbb{E}[\frac{1}{4}\text{erf}(\frac{1}{2}u)\text{erf}(\frac{1}{2}v)] + \mathbb{E}[\frac{1}{4}\text{erf}(\frac{1}{2}u) + \frac{1}{4}\text{erf}(\frac{1}{2}v)] + \frac{1}{4} \\ &= \frac{1}{4}\,\mathbb{E}[\text{erf}(\frac{1}{2}u)\text{erf}(\frac{1}{2}v)] + \frac{1}{4} \\ &= \frac{1}{4}\mathcal{T}(\frac{1}{4}\Sigma, \text{erf}, \text{erf})(\mathbf{x}, \hat{\mathbf{x}}) + \frac{1}{4}\end{aligned}$$

$$\boxed{\mathcal{T}(\Sigma, \sigma_{\hat{s}}, \sigma_{\hat{s}})(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{4} + \frac{1}{2\pi}\arcsin\left(\frac{\Sigma(\mathbf{x}, \hat{\mathbf{x}})}{\sqrt{(\Sigma(\mathbf{x}, \mathbf{x}) + 2)(\Sigma(\hat{\mathbf{x}}, \hat{\mathbf{x}}) + 2)}}\right)}$$

and

$$\begin{aligned}\mathcal{T}(\Sigma, \dot{\sigma}_{\hat{s}}, \dot{\sigma}_{\hat{s}})(\mathbf{x}, \hat{\mathbf{x}}) &= \mathbb{E}_{u,v\sim\mathcal{N}(0,\Sigma)}[\dot{\sigma}_{\hat{s}}(u)\dot{\sigma}_{\hat{s}}(v)] = \frac{1}{16}\,\mathbb{E}[\dot{\mathrm{erf}}(\frac{1}{2}u)\dot{\mathrm{erf}}(\frac{1}{2}v)] \\ &= \frac{1}{16}\mathcal{T}(\frac{1}{4}\Sigma, \dot{\mathrm{erf}}, \dot{\mathrm{erf}})(\mathbf{x}, \hat{\mathbf{x}})\end{aligned}$$

$$\boxed{\mathcal{T}(\Sigma, \dot{\sigma}_{\hat{s}}, \dot{\sigma}_{\hat{s}})(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{2\pi}\frac{1}{\sqrt{(2+\Sigma(\mathbf{x},\mathbf{x})(2+\Sigma(\hat{\mathbf{x}},\hat{\mathbf{x}})) - \Sigma(\mathbf{x},\hat{\mathbf{x}})^2)}}}$$

Based on the definition of NTK, we can derive the following for $\sigma_{\hat{s}}$

$$\begin{aligned}\Theta^1_\infty(\hat{\mathbf{x}}, \mathbf{x}) &= \Sigma^1(\hat{\mathbf{x}}, \mathbf{x}) = \frac{1}{n_0}\hat{\mathbf{x}}^T\mathbf{x} \\ \Theta^2_\infty(\hat{\mathbf{x}}, \mathbf{x}) &= \Theta^1_\infty(\hat{\mathbf{x}}, \mathbf{x})\mathcal{T}(\Theta^1_\infty, \dot{\sigma}_{\hat{s}}, \dot{\sigma}_{\hat{s}})(\mathbf{x}, \hat{\mathbf{x}}) + \mathcal{T}(\Theta^1_\infty, \sigma_{\hat{s}}, \sigma_{\hat{s}})(\mathbf{x}, \hat{\mathbf{x}})\end{aligned}$$

Let's look at the first part

$$\begin{aligned}\Theta^1_\infty(\hat{\mathbf{x}}, \mathbf{x})\mathcal{T}(\Theta^1_\infty, \dot{\sigma}_{\hat{s}}, \dot{\sigma}_{\hat{s}})(\mathbf{x}, \hat{\mathbf{x}}) &= \frac{1}{2\pi}\frac{1}{\sqrt{(2+\frac{1}{n_0}\mathbf{x}^T\mathbf{x})(2+\frac{1}{n_0}\hat{\mathbf{x}}^T\hat{\mathbf{x}}) - (\frac{1}{n_0}\hat{\mathbf{x}}^T\mathbf{x})^2)}}[\frac{1}{n_0}\hat{\mathbf{x}}^T\mathbf{x}] \\ &= \frac{1}{2\pi}\frac{\hat{\mathbf{x}}^T\mathbf{x}}{\sqrt{(2n_0+\mathbf{x}^T\mathbf{x})(2n_0+\hat{\mathbf{x}}^T\hat{\mathbf{x}}) - (\hat{\mathbf{x}}^T\mathbf{x})^2)}}\end{aligned}$$

and the second part

$$\begin{aligned}\mathcal{T}(\Theta^1_\infty, \sigma_{\hat{s}}, \sigma_{\hat{s}})(\mathbf{x}, \hat{\mathbf{x}}) &= \frac{1}{4} + \frac{1}{2\pi}\arcsin\left(\frac{\frac{1}{n_0}\hat{\mathbf{x}}^T\mathbf{x}}{\sqrt{(\frac{1}{n_0}\mathbf{x}^T\mathbf{x}+2)(\frac{1}{n_0}\hat{\mathbf{x}}^T\hat{\mathbf{x}}+2)}}\right) \\ &= \frac{1}{4} + \frac{1}{2\pi}\arcsin\left(\frac{\hat{\mathbf{x}}^T\mathbf{x}}{\sqrt{(\mathbf{x}^T\mathbf{x}+2n_0)(\hat{\mathbf{x}}^T\hat{\mathbf{x}}+2n_0)}}\right)\end{aligned}$$

D.2. Detailed Discussion of $\frac{\partial \mathbf{k}_x}{\partial \mathbf{x}}$

Without loss of generality, we will focus on $\frac{\partial \mathbf{k}_x}{\partial \mathbf{x}}|_{\mathbf{x}_1}$,

$$\frac{\partial \mathbf{k}_x}{\partial \mathbf{x}} = \begin{bmatrix}\frac{\Theta^L_\infty(\mathbf{x}_1,\mathbf{x})}{\partial \mathbf{x}} \\ \cdots \\ \frac{\Theta^L_\infty(\mathbf{x}_n,\mathbf{x})}{\partial \mathbf{x}}\end{bmatrix}$$

where

$$\frac{\partial\Theta^L_\infty(\hat{\mathbf{x}}, \mathbf{x})}{\partial \mathbf{x}} = \underbrace{\frac{\partial\mathcal{T}(\Theta^1_\infty, \sigma_{\hat{s}}, \sigma_{\hat{s}})(\hat{\mathbf{x}}, \mathbf{x}))}{\partial \mathbf{x}}}_{I^g_1(\hat{\mathbf{x}},\mathbf{x})} + \underbrace{\frac{\partial\Theta^1_\infty(\hat{\mathbf{x}}, \mathbf{x}))\mathcal{T}(\Theta^1_\infty, \dot{\sigma}_{\hat{s}}, \dot{\sigma}_{\hat{s}})(\hat{\mathbf{x}}, \mathbf{x}))}{\partial \mathbf{x}}}_{I^g_2(\hat{\mathbf{x}},\mathbf{x})}$$

Let's look at each row separately, and break this down into two parts.

- $I^g_1(\hat{\mathbf{x}}, \mathbf{x})$

  After deriving the derivative, we get this:

$$\begin{aligned}I^g_1(\hat{\mathbf{x}}, \mathbf{x}) &= \frac{1}{2\pi}\frac{1}{\sqrt{1-A^2}}\frac{\hat{\mathbf{x}}\Big[(\hat{\mathbf{x}}^T\hat{\mathbf{x}}+2n_0)(\mathbf{x}^T\mathbf{x}+2n_0)\Big] - \mathbf{x}\Big[(\hat{\mathbf{x}}^T\hat{\mathbf{x}}+2n_0)\mathbf{x}^T\hat{\mathbf{x}}\Big]}{\Big[(\hat{\mathbf{x}}^T\hat{\mathbf{x}}+2n_0)(\mathbf{x}^T\mathbf{x}+2n_0)\Big]^{\frac{3}{2}}} \\ A &= \frac{\hat{\mathbf{x}}^T\mathbf{x}}{\sqrt{(\mathbf{x}^T\mathbf{x}+2n_0)(\hat{\mathbf{x}}^T\hat{\mathbf{x}}+2n_0)}}\end{aligned}$$

Since we are only interested in $\mathbf{J}_\infty(\mathbf{x}_1)$ and each row of $\frac{\partial \mathbf{k}_x}{\partial \mathbf{x}}$, we'll examine $I_1^g(\mathbf{x}_i, \mathbf{x}_1)$.

$$\begin{aligned} I_1^g(\mathbf{x}_i, \mathbf{x}_1) &= \frac{1}{2\pi} \frac{r^2 + 2n_0}{\sqrt{(r^2+2n_0)^2 - (r^2\rho_{i1})^2}} \frac{\mathbf{x}_i (r^2+2n_0)^2 - \mathbf{x}_1 \Big[(r^2+2n_0) r^2 \rho_{i1} \Big]}{(r^2+2n_0)^3} \\ &= \frac{1}{2\pi} \frac{1}{\sqrt{(r^2+2n_0)^2 - (r^2\rho_{i1})^2}} \frac{\mathbf{x}_i (r^2+2n_0) - \mathbf{x}_1 r^2 \rho_{i1}}{r^2 + 2n_0} \end{aligned}$$

It is easy to see that $I_1^g(\mathbf{x}_i, \mathbf{x}_1) \to 0$ as r grows regardless of ρ_{i1}.

- $I_2^g(\hat{\mathbf{x}}, \mathbf{x})$

  We know that

$$I_2^g(\hat{\mathbf{x}}, \mathbf{x}) = \frac{1}{2\pi} \frac{\hat{\mathbf{x}} \Big[(\mathbf{x}^T\mathbf{x} + 2n_0)(\hat{\mathbf{x}}^T\hat{\mathbf{x}} + 2n_0) - (\hat{\mathbf{x}}^T\mathbf{x})^2 \Big] - \hat{\mathbf{x}}^T\mathbf{x} \Big[(2n_0 + \hat{\mathbf{x}}^T\hat{\mathbf{x}})\mathbf{x} - (\hat{\mathbf{x}}^T\mathbf{x})\hat{\mathbf{x}} \Big]}{\Big[(\mathbf{x}^T\mathbf{x} + 2n_0)(\hat{\mathbf{x}}^T\hat{\mathbf{x}} + 2n_0) - (\hat{\mathbf{x}}^T\mathbf{x})^2 \Big]^{\frac{3}{2}}}$$

  Again, let's examine $I_2^g(\mathbf{x}_i, \mathbf{x}_1)$.

$$I_2^g(\mathbf{x}_i, \mathbf{x}_1) = \frac{1}{2\pi} \frac{(r^2+2n_0)^2 \mathbf{x}_i - r^2 \rho_{i1} (2n_0 + r^2) \mathbf{x}_1}{\Big[(r^2+2n_0)^2 - r^4 \rho_{i1}^2 \Big]^{\frac{3}{2}}}$$

$$\begin{aligned} \|I_2^g(\mathbf{x}_i, \mathbf{x}_1)\|_2^2 &= \frac{1}{4\pi^2} \frac{r^2 \Big[(r^2+2n_0)^4 + r^4 \rho_{i1}^2 (2n_0 + r^2)^2 - 2r^2 \rho_{i1}^2 (2n_0 + r^2)^3 \Big]}{\Big[(r^2+2n_0)^2 - r^4 \rho_{i1}^2 \Big]^3} \\ &= \frac{1}{4\pi^2} \frac{16n_0^4 + r^2 \Big[n_0^3 (32 - 16\rho_{i1}^2) + r^2 \Big[n_0^2 (24 - 20\rho_{i1}^2) + r^2 \big[n_0 (8 - 8\rho_{i1}^2) + r^2 (1 - \rho_{i1}^2) \big] \Big] \Big]}{\Big[r^4 (1 - \rho_{i1}^2) + 4n_0 r^2 + 4n_0^2 \Big]^3} \end{aligned}$$

  Based on the equation for $\|I_2^g(\mathbf{x}_i, \mathbf{x}_1)\|_2^2$, we know that if $\rho_{i1}^2 \neq 1$, $\|I_2^g(\mathbf{x}_i, \mathbf{x}_1)\|_2^2$ eventually decays to zero with larger r. But $\|I_2^g(\mathbf{x}_i, \mathbf{x}_1)\|_2^2$ converges to a constant if $\rho_{i1}^2 = 1$. For simplicity, in this section, we do not assume there is any parallel input. Therefore, we can see that all the other terms will go to zero except $I_2^g(\mathbf{x}_1, \mathbf{x}_1)$. It is worth noting that if ρ_{i1} is close to one, the norm will see a spike before going down to zero. But in practice, the data is more than likely to be well separated with small $|\rho_{ij}|$. The discussion here is illustrated in Figure. D.1.

Combining the above analysis on the two components of gradient, it is easy to see that with large r,

$$\frac{\partial \mathbf{k}_x}{\partial \mathbf{x}}|_{\mathbf{x}_1} \approx \begin{bmatrix} \frac{\Theta_\infty^L(\mathbf{x}_1, \mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}_1} \\ \mathbf{0} \\ \dots \\ \mathbf{0} \end{bmatrix}$$

$$\|\frac{\partial \mathbf{k}_x}{\partial \mathbf{x}}|_{\mathbf{x}_1}\|_{op} \approx \|\frac{\Theta_\infty^L(\mathbf{x}_1, \mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}_1}\|_2 \approx \|\frac{1}{2\pi} \frac{2n_0(r^2+2n_0)}{(4n_0r^2 + 4n_0^2)^{\frac{3}{2}}} \mathbf{x}_1\|_2 = \frac{1}{2\pi} \frac{2n_0 r (r^2+2n_0)}{(4n_0r^2 + 4n_0^2)^{\frac{3}{2}}} \approx \frac{1}{8\pi\sqrt{n_0}}$$

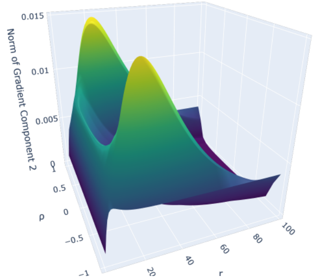


Figure D.1: ρ, r vs Norm of Gradient Component 2

D.3. Parallel Inputs Analysis

In the previous section, we assume that there are no parallel inputs. But this assumption is not necessary. In fact, given training data $\{\mathbf{x}_i\}_1^n$, w.l.o.g, let's impose $\mathbf{x}_1 = -\mathbf{x}_2$. Based on the results we have in Section 4.4, we can still derive a similar approximation for the NTK regression solution.

- $\tilde{\mathbf{K}}$

  Fisrt of all,

$$\mathbf{K}^1_{ij} = \mathcal{T}(\Theta^1_\infty, \sigma_{\hat{s}}, \sigma_{\hat{s}})(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{4} + \frac{1}{2\pi} \arcsin\left(\frac{r^2 \rho_{i,j}}{(r^2 + 2n_0)}\right)$$

  If $\rho_{i,j} = 1$, then $\mathbf{K}^1_{ij}$ is going to converge to $\frac{1}{2}$ as r grows bigger. But if $\rho_{i,j} = -1$, this term is going to zero.

  Therefore, $\tilde{\mathbf{K}}$ can be approximated by this block diagonal matrix.

$$\tilde{\mathbf{K}} \approx \begin{bmatrix} B_1 & \dots & 0 \\ 0 & B_2 & 0 \\ \dots & \dots & \dots \\ 0 & \dots & B_2 \end{bmatrix}$$

  where

$$B_1 = \begin{bmatrix} I_k + \frac{1}{2} & -I_k \\ -I_k & I_k + \frac{1}{2} \end{bmatrix} \qquad B_2 = I_k + \frac{1}{2} \qquad I_k = \frac{1}{2\pi} \frac{r^2}{\sqrt{4n_0^2 + 4n_0 r^2}} \approx \frac{r}{4\pi\sqrt{n_0}}$$

  The inverse of $\tilde{\mathbf{K}}$, is the following, as r grows large:

$$\tilde{\mathbf{K}}^{-1} \approx \begin{bmatrix} B_1^{-1} & \dots & 0 \\ 0 & \frac{1}{I_k + \frac{1}{2}} & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \frac{1}{I_k + \frac{1}{2}} \end{bmatrix} \approx \begin{bmatrix} B_1^{-1} & \dots & 0 \\ 0 & \frac{4\pi\sqrt{n_0}}{r} & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \frac{4\pi\sqrt{n_0}}{r} \end{bmatrix}$$

  where

$$B_1^{-1} = \frac{1}{I_k + \frac{1}{4}} \begin{bmatrix} I_k + \frac{1}{2} & I_k \\ I_k & I_k + \frac{1}{2} \end{bmatrix}$$

- $\frac{\partial \mathbf{k}_x}{\partial \mathbf{x}}$

  Based on the discussion from Section 4.4,

$$\frac{\partial \mathbf{k}_x}{\partial \mathbf{x}}|_{\mathbf{x}_1} \approx \begin{bmatrix} \frac{\Theta^L_\infty(\mathbf{x}_1, \mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}_1} \\ -\frac{\Theta^L_\infty(\mathbf{x}_1, \mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}_1} \\ \dots \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} J_k \mathbf{x}_1 \\ -J_k \mathbf{x}_1 \\ \dots \\ \mathbf{0} \end{bmatrix}$$

where

$$J_k = \frac{1}{2\pi} \frac{2n_0(r^2 + 2n_0)}{(4n_0 r^2 + 4n_0^2)^{\frac{3}{2}}} \approx \frac{1}{8\pi\sqrt{n_0}} \frac{1}{r}$$

Finally,

$$\left(\hat{\mathbf{X}} - f_0(\hat{\mathbf{X}})\right)\tilde{\mathbf{K}}^{-1}\frac{\partial k_x}{\partial \mathbf{x}} \approx \hat{\mathbf{X}}\tilde{\mathbf{K}}^{-1}\frac{\partial k_x}{\partial \mathbf{x}} \approx \hat{\mathbf{X}} \begin{bmatrix} B_1^{-1} & \dots & 0 \\ 0 & \frac{1}{I_k + \frac{1}{2}} & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \frac{1}{I_k + \frac{1}{2}} \end{bmatrix} \begin{bmatrix} J_k \mathbf{x}_1 \\ -J_k \mathbf{x}_1 \\ \mathbf{0} \\ \dots \\ \mathbf{0} \end{bmatrix}$$

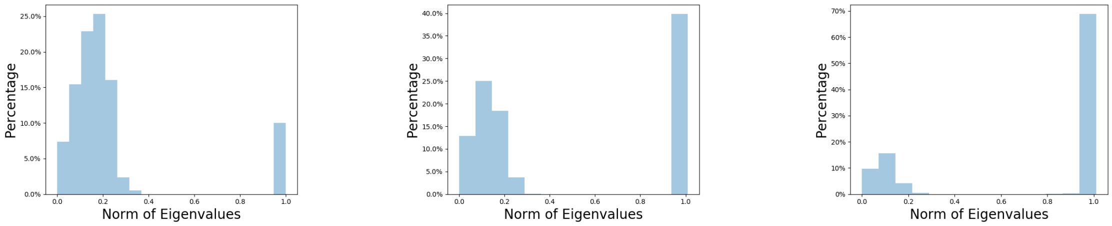
$$= \hat{\mathbf{X}} \begin{bmatrix} \frac{\frac{1}{2}J_k}{I_k + \frac{1}{4}}\mathbf{x}_1 \\ -\frac{\frac{1}{2}J_k}{I_k + \frac{1}{4}}\mathbf{x}_1 \\ \mathbf{0} \\ \dots \\ \mathbf{0} \end{bmatrix} = 2\frac{\frac{1}{2}J_k}{I_k + \frac{1}{4}}\mathbf{x}_1\mathbf{x}_1^T$$

Thus,

$$\|\left(\hat{\mathbf{X}} - f_0(\hat{\mathbf{X}})\right)\tilde{\mathbf{K}}^{-1}\frac{\partial \mathbf{k}_x}{\partial \mathbf{x}}\|_{op} \approx \frac{r^2 J_k}{I_k + \frac{1}{4}} = \frac{r^2 \frac{1}{8\pi\sqrt{n_0}} \frac{1}{r}}{\frac{r}{4\pi\sqrt{n_0}}} = \frac{1}{2}$$

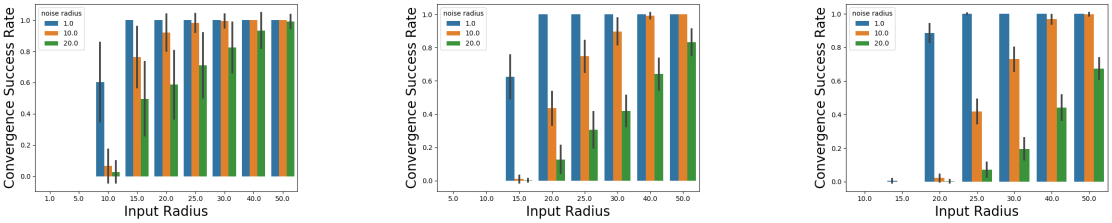
By similar argument, as $r \to \infty$, we have

$$\|\mathbf{J}_\infty(\mathbf{x})\|_{op} \leq \frac{1}{2}$$



(a) 2 training points (b) 5 training points (c) 8 training points

Figure E.1: Eigenvalue distribution of 2-layer sigmoid network trained with input dimension 10



(a) 5 training points (b) 20 training points (c) 40 training points

Figure E.2: Convergence success rate vs input norm: random data with input dimension 32

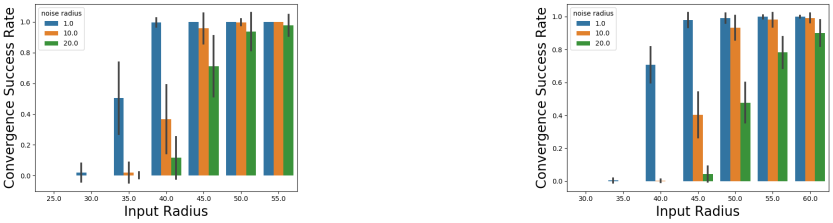
E. Additional Simulations

E.1. Multiple Points: Linear Region

In this section, we first illustrate the eigenvalue distribution in the linear region. Here, we trained 2 layer sigmoid networks with input dimension 10 and hidden size 1000 for 2, 5 and 8 training points. As suggested by Lemma 4, there should be $n - 1$ eigenvalues with norm around 1. This is supported by Figure E.1, as there are 10%, 40% and 70% eigenvalues around that region.

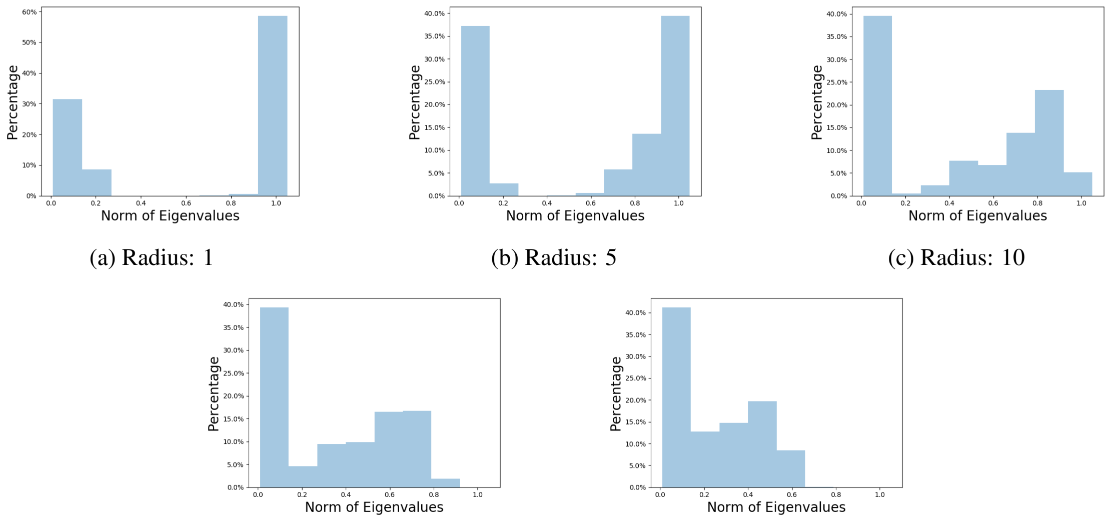
E.2. Basin of Attraction

We test basin of attraction by adding Gaussian noises to training examples and check if the modified examples can converge to the original ones via iterative maps under 50 iterations. The standard deviation of the Gaussian noise is called the noise radius. The network has 2 layers with hidden size 10000 and input dimension 32. Figure E.3 details experiments for 5, 20 and 40 examples. Not surprisingly, the basin of attraction is larger when there are fewer training examples and larger input norms since a level of separation between data is required.



(a) 5 training points (b) 20 training points

Figure E.3: Convergence success rate vs input norm: MNIST dataset



(a) Radius: 1

(b) Radius: 5

(c) Radius: 10

(d) Radius: 15

(e) Radius: 20

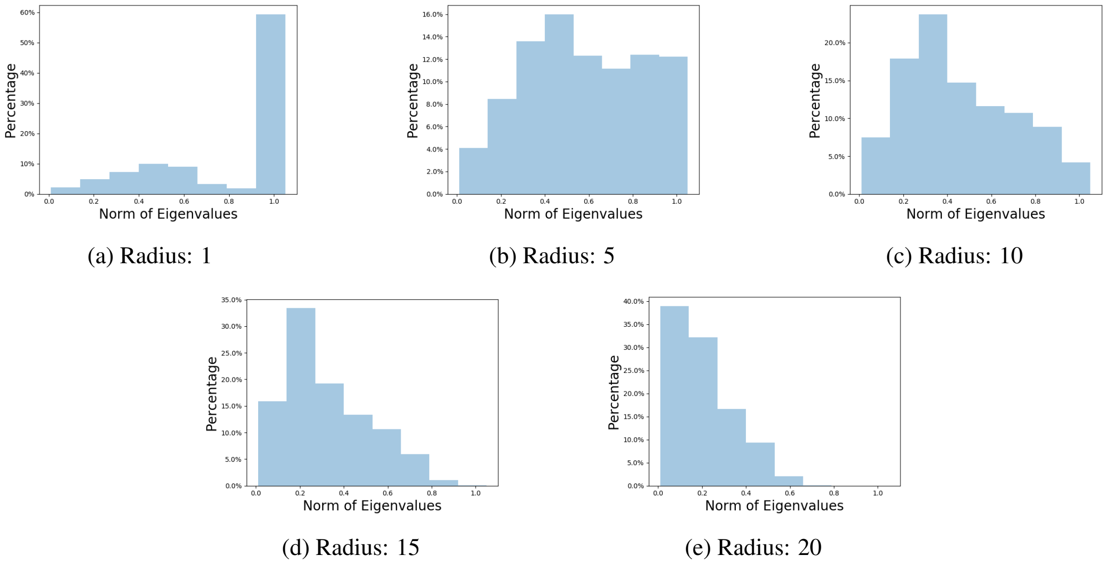
Figure E.4: Specturm Change for Sigmoid

E.3. Basin of Attraction on Mnist

We also test basin of attraction experiments on MNIST dataset to check if we can recover real training examples. The images are prepossessed by subtracting means and rescaled to have different input norms for testing. Similar to the setting before, Figure 4b also shows that larger input norm gives greater basin of attraction for 5 and 20 examples. Notice that because MNIST images have large input dimension, they need larger radius to move out of the linear region.

E.4. Sigmoidal Activations

Finally, we show that our results can be extended to different sigmoidal activation functions as well. We chose 2 layer network with hidden size 10000, input dimension 32 and 20 training examples. As before, only settings that can let network converges to training loss below 10^{-7} are included. Figure 5 clearly suggests all the activation functions share similar curves. Notice that both $\tanh$ and erf have large eigenvalue when r is small. This is not a contradiction to our Lemma 3 as their $\alpha = \dot{\sigma}(0)$ is too large to satisfy the conditions in Lemma 3. The histogram of eigenvalue norm changes for those activation is shown in Figure E.4, Figure E.5, Figure E.6. It is clear that they all follow the same pattern.



(a) Radius: 1

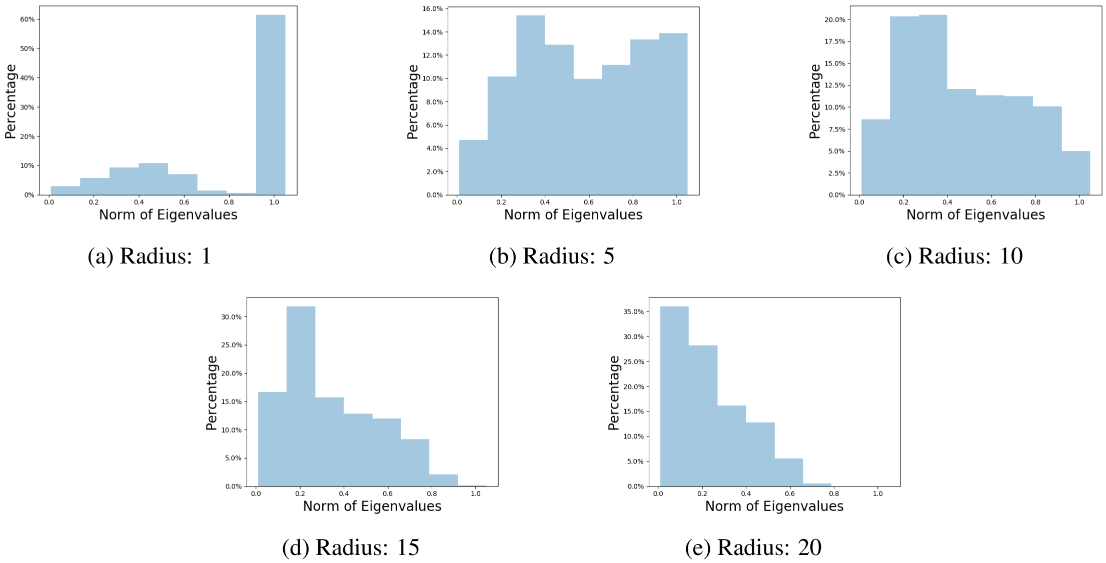
(b) Radius: 5

(c) Radius: 10

(d) Radius: 15

(e) Radius: 20

Figure E.5: Specturm Change for Erf



(a) Radius: 1

(b) Radius: 5

(c) Radius: 10

(d) Radius: 15

(e) Radius: 20

Figure E.6: Specturm Change for Sigmoid