5 Appendix

Let $\angle(\boldsymbol{h}, \boldsymbol{h}_0) = \bar{\theta}_0^{(1)}$ and $\angle(\boldsymbol{m}, \boldsymbol{m}_0) = \bar{\theta}_0^{(2)}$ for non-zero $\boldsymbol{h}, \boldsymbol{h}_0 \in \mathbb{R}^n$ and $\boldsymbol{m}, \boldsymbol{m}_0 \in \mathbb{R}^p$. In order to understand how the operators $\boldsymbol{h} \to \boldsymbol{W}_{+,\boldsymbol{h}}^{(1)}\boldsymbol{h}$ and $\boldsymbol{m} \to \boldsymbol{W}_{+,\boldsymbol{m}}^{(2)}\boldsymbol{m}$ distort angles, we define

$$g(\theta) = \cos^{-1}\left(\frac{(\pi-\theta)\cos\theta + \sin\theta}{\pi}\right). \tag{8}$$

Also, for a fixed $\boldsymbol{p},\ \boldsymbol{q} \in \mathbb{R}^n$, define

$$\tilde{t}_{\boldsymbol{p},\boldsymbol{q}}^{(k)} := \frac{1}{2^{a^{(k)}}}\left[\left(\prod_{i=0}^{a^{(k)}-1}\frac{\pi-\bar{\theta}_i^{(k)}}{\pi}\right)\boldsymbol{q} + \sum_{i=0}^{a^{(k)}-1}\frac{\sin\bar{\theta}_i^{(k)}}{\pi}\left(\prod_{j=i+1}^{a^{(k)}-1}\frac{\pi-\bar{\theta}_j^{(k)}}{\pi}\right)\frac{\|\boldsymbol{q}\|_2}{\|\boldsymbol{p}\|_2}\boldsymbol{p}\right], \tag{9}$$

where $\bar{\theta}_i^{(k)} = g(\bar{\theta}_{i-1}^{(k)})$ for g given by (8), $\bar{\theta}_0^{(k)} = \angle(\boldsymbol{p}, \boldsymbol{q})$, $a^{(1)} = d$, and $a^{(2)} = s$.

5.1 Proof of Deterministic Theorem

Theorem 4 (Also Theorem 3)**.** *Fix $\epsilon > 0$, $0 < \alpha_1 \le 1$ and $0 < \alpha_2 \le 1$ such that $K_1(d^7s^2 + d^2s^7)\epsilon^{1/4}/(\alpha_1\alpha_2) < 1$, $d \ge 2$, and $s \ge 2$. Let $\mathcal{K} = \{(\boldsymbol{h}, \boldsymbol{0}) \in \mathbb{R}^{n\times p} \,|\boldsymbol{h} \in \mathbb{R}^n\} \cup \{(\boldsymbol{0}, \boldsymbol{m}) \in \mathbb{R}^{n\times p} \,|\boldsymbol{m} \in \mathbb{R}^p\}$. Suppose that $\boldsymbol{W}_i^{(1)} \in \mathbb{R}^{n_i \times n_{i-1}}$ for $i = 1, \dots, d-1$ and $\boldsymbol{W}_i^{(2)} \in \mathbb{R}^{p_i \times p_{i-1}}$ for $i = 1, \dots, s-1$ satisfy the WDC with constant ϵ and 1. Suppose $\boldsymbol{W}_d^{(1)} \in \mathbb{R}^{\ell \times n_{d-1}}$ satisfy WDC with constants ϵ and α_1, and $\boldsymbol{W}_s^{(2)} \in \mathbb{R}^{\ell \times p_{s-1}}$ satisfy WDC with constants ϵ and α_2. Also, suppose $\left(\boldsymbol{W}_d^{(1)}, \boldsymbol{W}_s^{(2)}\right)$ satisfy joint-WDC with constants ϵ, $\alpha = \alpha_1 \cdot \alpha_2$. Let $\mathcal{K} = \{(\boldsymbol{h}, \boldsymbol{0}) \in \mathbb{R}^{n\times p} \,|\boldsymbol{h} \in \mathbb{R}^n\} \cup \{(\boldsymbol{0}, \boldsymbol{m}) \in \mathbb{R}^{n\times p} \,|\boldsymbol{m} \in \mathbb{R}^p\}$ and $\mathcal{A} = \mathcal{A}_{K_2 d^3 s^3 \epsilon^{\frac{1}{4}}\alpha^{-1}, (\boldsymbol{h}_0, \boldsymbol{m}_0)} \cup \mathcal{A}_{K_2 d^8 s^3 \epsilon^{\frac{1}{4}}\alpha^{-1}, \left(-\rho_d^{(1)}\boldsymbol{h}_0, \boldsymbol{m}_0\right)} \cup \mathcal{A}_{K_2 d^3 s^8 \epsilon^{\frac{1}{4}}\alpha^{-1}, \left(\rho_s^{(2)}\boldsymbol{h}_0, -\boldsymbol{m}_0\right)} \cup \mathcal{A}_{K_2 d^8 s^8 \epsilon^{\frac{1}{4}}\alpha^{-1}, \left(-\rho_d^{(1)}\rho_s^{(2)}\boldsymbol{h}_0, -\boldsymbol{m}_0\right)}$. Then, for $(\boldsymbol{h}_0, \boldsymbol{m}_0) \ne (\boldsymbol{0}, \boldsymbol{0})$, and*

$$(\boldsymbol{h}, \boldsymbol{m}) \notin \mathcal{A} \cup \mathcal{K}$$

the one-sided directional derivative of f in the direction of $\boldsymbol{g} = \boldsymbol{g}_{1,(\boldsymbol{h},\boldsymbol{m})}$ or $\boldsymbol{g} = \boldsymbol{g}_{2,(\boldsymbol{h},\boldsymbol{m})}$ satisfy $D_{-\boldsymbol{g}}f(\boldsymbol{h}, \boldsymbol{m}) < 0$. Additionally, for all $(\boldsymbol{h}, \boldsymbol{m}) \in \mathcal{K}$ and for all $(\boldsymbol{x}, \boldsymbol{y})$

$$D_{(\boldsymbol{x},\boldsymbol{y})}f(\boldsymbol{h}, \boldsymbol{m}) \le 0.$$

Here, $\rho_d^{(k)}$ are positive numbers that converge to 1 as $d \to \infty$, and K_1, and K_2 are absolute constants.

Proof. Recall that

$$\boldsymbol{v}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}^{(1)} = \begin{cases} \nabla_{\boldsymbol{h}} f(\boldsymbol{h}, \boldsymbol{m}) & G \text{ is differentiable at } (\boldsymbol{h}, \boldsymbol{m}), \\ \lim_{\delta\to 0^+} \nabla_{\boldsymbol{h}} f((\boldsymbol{h}, \boldsymbol{m}) + \delta\boldsymbol{w}) & \text{otherwise} \end{cases}$$

$$\boldsymbol{v}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}^{(2)} = \begin{cases} \nabla_{\boldsymbol{m}} f(\boldsymbol{h}, \boldsymbol{m}) & G \text{ is differentiable at } (\boldsymbol{h}, \boldsymbol{m}), \\ \lim_{\delta\to 0^+} \nabla_{\boldsymbol{m}} f((\boldsymbol{h}, \boldsymbol{m}) + \delta\boldsymbol{w}) & \text{otherwise} \end{cases}$$

where $G(\boldsymbol{h}, \boldsymbol{m})$ is differentiable at $(\boldsymbol{h}, \boldsymbol{m}) + \delta\boldsymbol{w}$ for sufficiently small δ. Such a δ exists because of piecewise linearity of $G(\boldsymbol{h}, \boldsymbol{m})$ and any such $\boldsymbol{w}$ can be arbitrarily selected. Also, recall that for any differentiable point $(\boldsymbol{h}, \boldsymbol{m})$, we have

$$\begin{aligned}
\nabla_{\boldsymbol{h}} f(\boldsymbol{h}, \boldsymbol{m}) &= \left(\boldsymbol{\Lambda}_{d,+,\boldsymbol{h}}^{(1)}\right)^\top \operatorname{diag}\left(\boldsymbol{\Lambda}_{s,+,\boldsymbol{m}}^{(2)}\boldsymbol{m}\right)^2 \boldsymbol{\Lambda}_{d,+,\boldsymbol{h}}^{(1)}\boldsymbol{h} \\
&\quad - \left(\boldsymbol{\Lambda}_{d,+,\boldsymbol{h}}^{(1)}\right)^\top \operatorname{diag}\left(\boldsymbol{\Lambda}_{s,+,\boldsymbol{m}}^{(2)}\boldsymbol{m} \odot \boldsymbol{\Lambda}_{s,+,\boldsymbol{m}_0}^{(2)}\boldsymbol{m}_0\right) \boldsymbol{\Lambda}_{d,+,\boldsymbol{h}_0}^{(1)}\boldsymbol{h}_0, \\
\nabla_{\boldsymbol{m}} f(\boldsymbol{h}, \boldsymbol{m}) &= \left(\boldsymbol{\Lambda}_{s,+,\boldsymbol{m}}^{(2)}\right)^\top \operatorname{diag}\left(\boldsymbol{\Lambda}_{d,+,\boldsymbol{h}}^{(1)}\boldsymbol{h}\right)^2 \boldsymbol{\Lambda}_{s,+,\boldsymbol{m}}^{(2)}\boldsymbol{m} \\
&\quad - \left(\boldsymbol{\Lambda}_{s,+,\boldsymbol{m}}^{(2)}\right)^\top \operatorname{diag}\left(\boldsymbol{\Lambda}_{d,+,\boldsymbol{h}}^{(1)}\boldsymbol{h} \odot \boldsymbol{\Lambda}_{d,+,\boldsymbol{h}_0}^{(1)}\boldsymbol{h}_0\right) \boldsymbol{\Lambda}_{s,+,\boldsymbol{m}_0}^{(2)}\boldsymbol{m}_0.
\end{aligned}$$

Let

$$\boldsymbol{g}_{1,(\boldsymbol{h},\boldsymbol{m})} = \begin{bmatrix} \boldsymbol{v}^{(1)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)} \\ \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{n\times p},$$

$$\boldsymbol{g}_{2,(\boldsymbol{h},\boldsymbol{m})} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{v}^{(2)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)} \end{bmatrix} \in \mathbb{R}^{n\times p},$$

$$\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)} = \frac{\alpha}{2^{d+s}\ell}\|\boldsymbol{m}\|_2^2\boldsymbol{h} - \frac{\alpha}{\ell}\boldsymbol{m}^\intercal\tilde{\boldsymbol{t}}^{(2)}_{\boldsymbol{m},\boldsymbol{m}_0}\tilde{\boldsymbol{t}}^{(1)}_{\boldsymbol{h},\boldsymbol{h}_0}, \tag{10}$$

$$\boldsymbol{t}^{(2)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)} = \frac{\alpha}{2^{d+s}\ell}\|\boldsymbol{h}\|_2^2\boldsymbol{m} - \frac{\alpha}{\ell}\boldsymbol{h}^\intercal\tilde{\boldsymbol{t}}^{(1)}_{\boldsymbol{h},\boldsymbol{h}_0}\tilde{\boldsymbol{t}}^{(2)}_{\boldsymbol{m},\boldsymbol{m}_0}, \tag{11}$$

$$S^{(1)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)} = \left\{ (\boldsymbol{h},\boldsymbol{m}) \in \mathbb{R}^{n\times p}\backslash\mathcal{K} \,\middle|\, \frac{\|\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}\|_2}{\|\boldsymbol{m}\|_2} \le \frac{\epsilon\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)}{2^{d+s}\ell} \right\},$$

$$S^{(2)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)} = \left\{ (\boldsymbol{h},\boldsymbol{m}) \in \mathbb{R}^{n\times p}\backslash\mathcal{K} \,\middle|\, \frac{\|\boldsymbol{t}^{(2)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}\|_2}{\|\boldsymbol{h}\|_2} \le \frac{\epsilon\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)}{2^{d+s}\ell} \right\},$$

where $\tilde{\boldsymbol{t}}^{(2)}_{\boldsymbol{m},\boldsymbol{m}_0}$ and $\tilde{\boldsymbol{t}}^{(1)}_{\boldsymbol{h},\boldsymbol{h}_0}$ is as defined in (9). For brevity of notation, write $\boldsymbol{v}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})} = \boldsymbol{v}^{(1)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}$, $\boldsymbol{v}^{(2)}_{(\boldsymbol{h},\boldsymbol{m})} = \boldsymbol{v}^{(2)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}$, $\boldsymbol{t}_{(\boldsymbol{h},\boldsymbol{m})} = \boldsymbol{t}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}$, $\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})} = \boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}$ and $\boldsymbol{t}^{(2)}_{(\boldsymbol{h},\boldsymbol{m})} = \boldsymbol{t}^{(2)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}$.

Since $\boldsymbol{W}^{(1)}_i \in \mathbb{R}^{n_i\times n_{i-1}}$ for $i = 1,\dots,d-1$ and $\boldsymbol{W}^{(2)}_i \in \mathbb{R}^{p_i\times p_{i-1}}$ for $i = 1,\dots,s-1$ satisfy the WDC with constant ϵ and 1, $\boldsymbol{W}^{(1)}_d \in \mathbb{R}^{\ell\times n_{d-1}}$ satisfy WDC with constants ϵ and α_1, $\boldsymbol{W}^{(2)}_s \in \mathbb{R}^{\ell\times p_{s-1}}$ satisfy WDC with constants ϵ and α_2, and $\left(\boldsymbol{W}^{(1)}_d, \boldsymbol{W}^{(2)}_s\right)$ satisfy joint-WDC with constants ϵ, $\alpha = \alpha_1\cdot\alpha_2$, we have lemma 2 implying for all nonzero $\boldsymbol{h},\ \boldsymbol{h}_0 \in \mathbb{R}^n$ and nonzero $\boldsymbol{m},\ \boldsymbol{m}_0 \in \mathbb{R}^p$

$$\|\nabla_{\boldsymbol{h}} f(\boldsymbol{h},\boldsymbol{m}) - \boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}\|_2 \le K\frac{d^3s^3\sqrt{\epsilon}}{2^{d+s}\ell}\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)\|\boldsymbol{m}\|_2, \tag{12}$$

$$\|\nabla_{\boldsymbol{m}} f(\boldsymbol{h},\boldsymbol{m}) - \boldsymbol{t}^{(2)}_{(\boldsymbol{h},\boldsymbol{m})}\|_2 \le K\frac{d^3s^2\sqrt{\epsilon}}{2^{d+s}\ell}\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)\|\boldsymbol{h}\|_2. \tag{13}$$

Thus, we have, for all nonzero $\boldsymbol{h},\ \boldsymbol{h}_0 \in \mathbb{R}^n$ and nonzero $\boldsymbol{m},\ \boldsymbol{m}_0 \in \mathbb{R}^p$,

$$\begin{aligned}
\|\boldsymbol{v}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})} - \boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}\|_2 &= \lim_{\delta\to 0^+}\|\nabla_{\boldsymbol{h}} f\left((\boldsymbol{h},\boldsymbol{x}) + \delta\boldsymbol{w}\right) - \boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})+\delta\boldsymbol{w}}\|_2 \\
&\le K\frac{d^3s^3\sqrt{\epsilon}}{2^{d+s}\ell}\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)\|\boldsymbol{m}\|_2, \text{ and} \\
\|\boldsymbol{v}^{(2)}_{(\boldsymbol{h},\boldsymbol{m})} - \boldsymbol{t}^{(2)}_{(\boldsymbol{h},\boldsymbol{m})}\|_2 &= \lim_{\delta\to 0^+}\|\nabla_{\boldsymbol{m}} f\left((\boldsymbol{h},\boldsymbol{x}) + \delta\boldsymbol{w}\right) - \boldsymbol{t}^{(2)}_{(\boldsymbol{h},\boldsymbol{m})+\delta\boldsymbol{w}}\|_2 \\
&\le K\frac{d^3s^3\sqrt{\epsilon}}{2^{d+s}\ell}\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)\|\boldsymbol{h}\|_2,
\end{aligned}$$

where the inequalities follow from (12) and (13).

Note that the one-sided directional derivative of f in the direction of $(\boldsymbol{x},\boldsymbol{y}) \neq \boldsymbol{0}$ at $(\boldsymbol{h},\boldsymbol{y})$ is $D_{(\boldsymbol{x},\boldsymbol{y})}f(\boldsymbol{h},\boldsymbol{x}) = \lim_{t\to 0^+}\frac{1}{t}\left(f((\boldsymbol{h},\boldsymbol{x}) + t(\boldsymbol{x},\boldsymbol{y})) - f(\boldsymbol{h},\boldsymbol{x})\right)$. Due to the continuity and piecewise linearity of the function

$$\mathcal{G}(\boldsymbol{h},\boldsymbol{m}) = \boldsymbol{\Lambda}^{(1)}_{d,+,\boldsymbol{h}}\boldsymbol{h} \odot \boldsymbol{\Lambda}^{(2)}_{s,+,\boldsymbol{m}}\boldsymbol{m},$$

we have that for any $(\boldsymbol{h},\boldsymbol{m}) \neq (\boldsymbol{0},\boldsymbol{0})$ and $(\boldsymbol{x},\boldsymbol{y}) \neq \boldsymbol{0}$ that there exists a sequence $\{(\boldsymbol{h}_n,\boldsymbol{m}_n)\} \to (\boldsymbol{h},\boldsymbol{m})$ such that f is differentiable at each $(\boldsymbol{h}_n,\boldsymbol{m}_n)$ and $D_{(\boldsymbol{x},\boldsymbol{y})}f(\boldsymbol{h},\boldsymbol{m}) = \lim_{n\to\infty}\nabla f(\boldsymbol{h}_n,\boldsymbol{m}_n)\cdot(\boldsymbol{x},\boldsymbol{y})$. Thus, as $\nabla f(\boldsymbol{h}_n,\boldsymbol{m}_n) = \begin{bmatrix} \boldsymbol{v}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)} \\ \boldsymbol{v}^{(2)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)} \end{bmatrix}$,

$$D_{-\boldsymbol{g}_{1,(\boldsymbol{h},\boldsymbol{m})}}f(\boldsymbol{h},\boldsymbol{m}) = \lim_{n\to\infty}\nabla f(\boldsymbol{h}_n,\boldsymbol{m}_n)\cdot\frac{-\boldsymbol{g}_{1,(\boldsymbol{h},\boldsymbol{m})}}{\|\boldsymbol{g}_{1,(\boldsymbol{h},\boldsymbol{m})}\|_2} = \frac{-1}{\|\boldsymbol{g}_{1,(\boldsymbol{h},\boldsymbol{m})}\|_2}\lim_{n\to\infty}\boldsymbol{v}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)}\cdot\boldsymbol{v}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})},$$

$$D_{-\boldsymbol{g}_{2,(\boldsymbol{h},\boldsymbol{m})}} f(\boldsymbol{h},\boldsymbol{m}) = \lim_{n\to\infty} \nabla f(\boldsymbol{h}_n,\boldsymbol{m}_n)\cdot \frac{-\boldsymbol{g}_{2,(\boldsymbol{h},\boldsymbol{m})}}{\|\boldsymbol{g}_{2,(\boldsymbol{h},\boldsymbol{m})}\|_2} = \frac{-1}{\|\boldsymbol{g}_{2,(\boldsymbol{h},\boldsymbol{m})}\|_2} \lim_{n\to\infty} \boldsymbol{v}^{(2)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)}\cdot \boldsymbol{v}^{(2)}_{(\boldsymbol{h},\boldsymbol{m})}.$$

Now, we write

$$\begin{aligned}
&\boldsymbol{v}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)}\cdot \boldsymbol{v}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}\\
=&\boldsymbol{t}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)}\cdot \boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})} + (\boldsymbol{v}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)} - \boldsymbol{t}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)})\cdot \boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})} + \boldsymbol{t}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)}\cdot (\boldsymbol{v}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})} - \boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})})\\
&+ (\boldsymbol{v}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)} - \boldsymbol{t}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)})\cdot (\boldsymbol{v}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})} - \boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})})\\
\geq&\boldsymbol{t}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)}\cdot \boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})} - \|\boldsymbol{v}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)} - \boldsymbol{t}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)}\|_2\|\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}\|_2 - \|\boldsymbol{t}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)}\|_2\|\boldsymbol{v}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})} - \boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}\|_2\\
&\|\boldsymbol{v}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)} - \boldsymbol{t}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)}\|_2\|\boldsymbol{v}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})} - \boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}\|_2\\
\geq&\boldsymbol{t}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)}\cdot \boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})} - K\frac{d^3s^3\sqrt{\epsilon}}{2^{d+s}\ell}\max(\|\boldsymbol{h}_n\|_2\|\boldsymbol{m}_n\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)\|\boldsymbol{m}_n\|_2\|\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}\|_2\\
&- K\frac{d^3s^3\sqrt{\epsilon}}{2^{d+s}\ell}\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)\|\boldsymbol{m}\|_2\|\boldsymbol{t}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)}\|_2\\
&- \left(K\frac{d^3s^3\sqrt{\epsilon}}{2^{d+s}\ell}\right)^2 \max(\|\boldsymbol{h}_n\|_2\|\boldsymbol{m}_n\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)\\
&\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)\|\boldsymbol{m}_n\|_2\|\boldsymbol{m}\|_2.
\end{aligned}$$

As $\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}$ is continuous in $(\boldsymbol{h},\boldsymbol{m})$ for all $(\boldsymbol{h},\boldsymbol{m})\notin\mathcal{K}$, we have for all $(\boldsymbol{h},\boldsymbol{m})\notin \mathcal{S}^{(1)}_{4Kd^3s^3\sqrt{\epsilon},(\boldsymbol{h}_0,\boldsymbol{m}_0)}\cup \mathcal{K}$,

$$\begin{aligned}
&\lim_{n\to\infty} \boldsymbol{v}^{(1)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)}\cdot \boldsymbol{v}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}\\
\geq&\|\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}\|_2^2 - 2K\frac{d^3s^3\sqrt{\epsilon}}{2^{d+s}\ell}\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)\|\boldsymbol{m}\|_2\|\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}\|_2\\
&- \left(K\frac{d^3s^3\sqrt{\epsilon}}{2^{d+s}\ell}\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)\|\boldsymbol{m}\|_2\right)^2\\
\geq&\frac{\|\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}\|_2}{2}\Big[\|\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}\|_2 - 4K\frac{d^3s^3\sqrt{\epsilon}}{2^{d+s}\ell}\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)\|\boldsymbol{m}\|_2\Big]+\\
&\frac{1}{2}\Big[\|\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m})}\|_2^2 - 2\left(K\frac{d^3s^3\sqrt{\epsilon}}{2^{d+s}\ell}\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2, \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)\|\boldsymbol{m}\|_2\right)^2\Big]\\
>&0.
\end{aligned}\tag{14}$$

Similarly, we have for all $(\boldsymbol{h},\boldsymbol{m})\notin \mathcal{S}^{(2)}_{4Kd^3s^3\sqrt{\epsilon},(\boldsymbol{h}_0,\boldsymbol{m}_0)}\cup\mathcal{K}$,

$$\lim_{n\to\infty} \boldsymbol{v}^{(2)}_{(\boldsymbol{h}_n,\boldsymbol{m}_n)}\cdot \boldsymbol{v}^{(2)}_{(\boldsymbol{h},\boldsymbol{m})} > 0.\tag{15}$$

So, for all $(\boldsymbol{h},\boldsymbol{m})\notin \left(\mathcal{S}^{(1)}_{4Kd^3s^3\sqrt{\epsilon},(\boldsymbol{h}_0,\boldsymbol{m}_0)}\cap \mathcal{S}^{(2)}_{4Kd^3s^3\sqrt{\epsilon},(\boldsymbol{h}_0,\boldsymbol{m}_0)}\right)\cup\mathcal{K}$, at least (14) or (15) holds. If (14) holds, then we have $D_{-\boldsymbol{g}_{1,(\boldsymbol{h},\boldsymbol{m})}} f(\boldsymbol{h},\boldsymbol{m}) < 0$ and if (15) holds, then we have $D_{-\boldsymbol{g}_{2,(\boldsymbol{h},\boldsymbol{m})}} f(\boldsymbol{h},\boldsymbol{m}) < 0$. Let $\mathcal{S} = \mathcal{S}^{(1)}_{4Kd^3s^3\sqrt{\epsilon},(\boldsymbol{h}_0,\boldsymbol{m}_0)}\cap \mathcal{S}^{(2)}_{4Kd^3s^3\sqrt{\epsilon},(\boldsymbol{h}_0,\boldsymbol{m}_0)}$. Apply Lemma 3 and $38(d^5+s^5)\sqrt{4Kd^3s^3\sqrt{\epsilon}}/\alpha<1$ to get

$$\begin{aligned}
\mathcal{S}\subseteq&\mathcal{A}_{\tilde{K}\frac{d^3s^3\epsilon^{1/4}}{\alpha},(\boldsymbol{h}_0,\boldsymbol{m}_0)}\cup\mathcal{A}_{\tilde{K}\frac{d^8s^3\epsilon^{1/4}}{\alpha},\left(-\rho_d^{(1)}\boldsymbol{h}_0,\boldsymbol{m}_0\right)}\cup\mathcal{A}_{\tilde{K}\frac{d^3s^8\epsilon^{1/4}}{\alpha},\left(\rho_s^{(2)}\boldsymbol{h}_0,-\boldsymbol{m}_0\right)}\\
&\cup\mathcal{A}_{\tilde{K}\frac{d^8s^8\epsilon^{1/4}}{\alpha},\left(-\rho_d^{(1)}\rho_s^{(2)}\boldsymbol{h}_0,-\boldsymbol{m}_0\right)},
\end{aligned}$$

for some absolute constant $\tilde{K}$.

It remains to prove that elements of the set $\mathcal{K}$ are local maximizers. We now show that elements of the set $\{(\boldsymbol{0},\boldsymbol{m})\in\mathbb{R}^{n\times p}|\boldsymbol{m}\in\mathbb{R}^p\}$ are local maximizers. Fix a direction $(\boldsymbol{x},\boldsymbol{y})\in\mathbb{R}^{n\times p}$ and let

$$\tilde{\boldsymbol{h}}_0 = \boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}_0}\boldsymbol{h}_0,\quad \tilde{\boldsymbol{x}} = \boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x},\quad \tilde{\boldsymbol{m}} = \boldsymbol{\Lambda}^{(2)}_{s-1,+,\boldsymbol{m}}\boldsymbol{m},\quad \tilde{\boldsymbol{m}}_0 = \boldsymbol{\Lambda}^{(2)}_{s-1,+,\boldsymbol{m}_0}\boldsymbol{m}_0,$$

$\bar{\theta}_i^{(k)} = g(\bar{\theta}_{i-1}^{(k)})$ for g given in (8), $\bar{\theta}_0^{(1)} = \angle(\boldsymbol{x}, \boldsymbol{h}_0)$ and $\bar{\theta}_0^{(2)} = \angle(\boldsymbol{m}, \boldsymbol{m}_0)$. We compute

$$
\begin{aligned}
& - D_{(\boldsymbol{x},\boldsymbol{y})} f(\boldsymbol{h}, \boldsymbol{m}) \cdot \|(\boldsymbol{x}, \boldsymbol{y})\|_2 \\
&= \lim_{t \to 0^+} \frac{f((\boldsymbol{h}, \boldsymbol{m}) + t(\boldsymbol{x}, \boldsymbol{y})) - f(\boldsymbol{h}, \boldsymbol{m})}{t} \\
&= \left\langle \operatorname{diag}\left(\boldsymbol{\Lambda}_{s,+,\boldsymbol{m}}^{(2)} \boldsymbol{m}\right) \boldsymbol{\Lambda}_{d,+,\boldsymbol{x}}^{(1)} \boldsymbol{x}, \operatorname{diag}\left(\boldsymbol{\Lambda}_{s,+,\boldsymbol{m}_0}^{(2)} \boldsymbol{m}_0\right) \boldsymbol{\Lambda}_{d,+,\boldsymbol{h}_0}^{(1)} \boldsymbol{h}_0 \right\rangle \\
&= \left\langle \tilde{\boldsymbol{x}}, \left(\boldsymbol{W}_{d-1,+,\boldsymbol{m}}^{(2)}\right)^{\mathsf{T}} \operatorname{diag}\left(\boldsymbol{W}_{d-1,+,\boldsymbol{x}}^{(1)} \tilde{\boldsymbol{m}} \odot \boldsymbol{W}_{d-1,+,\boldsymbol{x}_0}^{(1)} \tilde{\boldsymbol{m}}_0\right) \boldsymbol{W}_{d-1,+,\boldsymbol{h}_0}^{(1)} \tilde{\boldsymbol{h}}_0 \right\rangle \\
&= \Big\langle \tilde{\boldsymbol{x}}, \Big(\left(\boldsymbol{W}_{d-1,+,\boldsymbol{m}}^{(2)}\right)^{\mathsf{T}} \operatorname{diag}\left(\boldsymbol{W}_{d-1,+,\boldsymbol{x}}^{(1)} \tilde{\boldsymbol{m}} \odot \boldsymbol{W}_{d-1,+,\boldsymbol{x}_0}^{(1)} \tilde{\boldsymbol{m}}_0\right) \boldsymbol{W}_{d-1,+,\boldsymbol{h}_0}^{(1)} \\
&\quad - \frac{\alpha}{n} \boldsymbol{Q}_{\tilde{\boldsymbol{x}},\tilde{\boldsymbol{h}}_0} \tilde{\boldsymbol{m}}^{\mathsf{T}} \boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{m}}_0} \tilde{\boldsymbol{m}}_0 \Big) \tilde{\boldsymbol{h}}_0 \Big\rangle + \frac{\alpha}{n} \tilde{\boldsymbol{m}}^{\mathsf{T}} \boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{m}}_0} \tilde{\boldsymbol{m}}_0 \cdot \tilde{\boldsymbol{x}}^{\mathsf{T}} \boldsymbol{Q}_{\tilde{\boldsymbol{x}},\tilde{\boldsymbol{h}}_0} \tilde{\boldsymbol{h}}_0 \\
&\geq - \Big\| \left(\boldsymbol{W}_{d-1,+,\boldsymbol{m}}^{(2)}\right)^{\mathsf{T}} \operatorname{diag}\left(\boldsymbol{W}_{d-1,+,\boldsymbol{x}}^{(1)} \tilde{\boldsymbol{m}} \odot \boldsymbol{W}_{d-1,+,\boldsymbol{x}_0}^{(1)} \tilde{\boldsymbol{m}}_0\right) \boldsymbol{W}_{d-1,+,\boldsymbol{h}_0}^{(1)} \\
&\quad - \frac{\alpha}{n} \boldsymbol{Q}_{\tilde{\boldsymbol{x}},\tilde{\boldsymbol{h}}_0} \tilde{\boldsymbol{m}}^{\mathsf{T}} \boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{m}}_0} \tilde{\boldsymbol{m}}_0 \Big\| \|\tilde{\boldsymbol{x}}\|_2 \|\tilde{\boldsymbol{h}}_0\|_2 + \frac{\alpha}{4n} \left(\frac{(\pi - \bar{\theta}_{d-1}^{(1)}) \cos \bar{\theta}_{d-1}^{(1)} + \sin \bar{\theta}_{d-1}^{(1)}}{\pi} \right) \\
&\quad \left(\frac{(\pi - \bar{\theta}_{d-1}^{(2)}) \cos \bar{\theta}_{d-1}^{(2)} + \sin \bar{\theta}_{d-1}^{(2)}}{\pi} \right) \|\tilde{\boldsymbol{x}}\| \|\tilde{\boldsymbol{h}}_0\| \|\tilde{\boldsymbol{m}}\| \|\tilde{\boldsymbol{m}}_0\| \\
&\geq - \frac{4\epsilon}{n} \|\tilde{\boldsymbol{m}}\|_2 \|\tilde{\boldsymbol{m}}_0\|_2 \|\tilde{\boldsymbol{x}}\|_2 \|\tilde{\boldsymbol{h}}_0\|_2 + \frac{\alpha}{4n} \|\tilde{\boldsymbol{x}}\| \|\tilde{\boldsymbol{h}}_0\| \|\tilde{\boldsymbol{m}}\| \|\tilde{\boldsymbol{m}}_0\| \cos \bar{\theta}_d^{(1)} \cos \bar{\theta}_d^{(2)} \\
&\geq \left(- \frac{4\epsilon}{n} + \frac{\alpha}{4\pi^2 n} \right) \|\tilde{\boldsymbol{x}}\| \|\tilde{\boldsymbol{h}}_0\| \|\tilde{\boldsymbol{m}}\| \|\tilde{\boldsymbol{m}}_0\|.
\end{aligned}
$$

By Lemma 1, we have $\|\tilde{\boldsymbol{x}}\| \geq \frac{1}{4\pi} \frac{1}{2^{d-1}} \|\boldsymbol{x}\|$ and $\|\tilde{\boldsymbol{m}}\| \geq \frac{1}{4\pi} \frac{1}{2^{s-1}} \|\boldsymbol{m}\|$. So, if $4\pi^2 \epsilon / \alpha < 1$, then $D_{(\boldsymbol{x},\boldsymbol{y})} f(\boldsymbol{h}, \boldsymbol{m}) \cdot \|(\boldsymbol{x}, \boldsymbol{y})\|_2 < 0$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{n \times p}$ with $\boldsymbol{x} \neq \boldsymbol{0}$ and $D_{(\boldsymbol{x},\boldsymbol{y})} f(\boldsymbol{h}, \boldsymbol{m}) \cdot \|(\boldsymbol{x}, \boldsymbol{y})\|_2 = 0$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{n \times p}$ with $\boldsymbol{x} = \boldsymbol{0}$. Thus, elements of the set $\{(\boldsymbol{0}, \boldsymbol{m}) \in \mathbb{R}^{n \times p} | \boldsymbol{m} \in \mathbb{R}^p\}$ are local maximizers. Similarly, elements of the set $\{(\boldsymbol{h}, \boldsymbol{0}) \in \mathbb{R}^{n \times p} | \boldsymbol{h} \in \mathbb{R}^n\}$ are local maximizers. This concludes the proof of Theorem 4. □

5.2 Concentration of terms in $\tilde{\boldsymbol{g}}_{1,(\boldsymbol{h},\boldsymbol{m})}$ and $\tilde{\boldsymbol{g}}_{2,(\boldsymbol{h},\boldsymbol{m})}$

Lemma 1. *Fix $0 < \epsilon < d^{-4}/(16\pi)^2$ and $d \geq 2$. Suppose that $\boldsymbol{W}_i \in \mathbb{R}^{n_i \times n_{i-1}}$ satisfies the WDC with constant ϵ and 1 for $i = 1, \ldots, d$. Define*

$$
\tilde{t}_{\boldsymbol{p},\boldsymbol{q}} = \frac{1}{2^d} \left[\left(\prod_{i=0}^{d-1} \frac{\pi - \bar{\theta}_i}{\pi} \right) \boldsymbol{q} + \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_i}{\pi} \left(\prod_{j=i+1}^{d-1} \frac{\pi - \bar{\theta}_j}{\pi} \right) \frac{\|\boldsymbol{q}\|_2}{\|\boldsymbol{p}\|_2} \boldsymbol{p} \right],
$$

where $\bar{\theta}_i = g(\bar{\theta}_{i-1})$ for g given by (8) and $\bar{\theta}_0 = \angle(\boldsymbol{p}, \boldsymbol{q})$. For all $\boldsymbol{p} \neq 0$ and $\boldsymbol{q} \neq 0$,

$$
\left\| \left(\prod_{i=d}^{1} \boldsymbol{W}_{i,+,\boldsymbol{p}} \right)^{\mathsf{T}} \left(\prod_{i=d}^{1} \boldsymbol{W}_{i,+,\boldsymbol{q}} \right) \boldsymbol{q} - \tilde{t}_{\boldsymbol{p},\boldsymbol{q}} \right\|_2 \leq 24 \frac{d^3 \sqrt{\epsilon}}{2^d} \|\boldsymbol{q}\|_2, \tag{16}
$$

$$
\left\langle \left(\prod_{i=d}^{1} \boldsymbol{W}_{i,+,\boldsymbol{p}} \right) \boldsymbol{p}, \left(\prod_{i=d}^{1} \boldsymbol{W}_{i,+,\boldsymbol{q}} \right) \boldsymbol{q} \right\rangle \geq \frac{1}{4\pi} \frac{1}{2^d} \|\boldsymbol{p}\|_2 \|\boldsymbol{q}\|_2 \tag{17}
$$

We refer the readers to Hand and Voroninski [2017] for proof of Lemma 1. We now state a related Lemma.

Lemma 2. *Fix $0 < \epsilon < 1/((d^4 + s^4)16\pi)^2$, $d \geq 2$ and $s \geq 2$. Suppose that $\boldsymbol{W}_i^{(1)} \in \mathbb{R}^{n_i \times n_{i-1}}$ for $i = 1, \ldots, d-1$ and $\boldsymbol{W}_i^{(2)} \in \mathbb{R}^{p_i \times p_{i-1}}$ for $i = 1, \ldots, s-1$ satisfy the WDC with constant ϵ and 1. Suppose $\boldsymbol{W}_d^{(1)} \in \mathbb{R}^{\ell \times n_{d-1}}$ satisfy WDC with constants ϵ and α_1, and $\boldsymbol{W}_s^{(2)} \in \mathbb{R}^{\ell \times p_{s-1}}$ satisfy WDC with constants ϵ and α_2. Also, suppose $\left(\boldsymbol{W}_d^{(1)}, \boldsymbol{W}_s^{(2)}\right)$ satisfy pair-WDC with constants ϵ,*

$\alpha = \alpha_1 \cdot \alpha_2$. *Define*

$$\tilde{t}^{(k)}_{\boldsymbol{p},\boldsymbol{q}} = \frac{1}{2^{a^{(k)}}} \left[\left(\prod_{i=0}^{a^{(k)}-1} \frac{\pi - \bar{\theta}^{(k)}_i}{\pi} \right) \boldsymbol{q} + \sum_{i=0}^{a^{(k)}-1} \frac{\sin \bar{\theta}^{(k)}_i}{\pi} \left(\prod_{j=i+1}^{a^{(k)}-1} \frac{\pi - \bar{\theta}^{(k)}_j}{\pi} \right) \frac{\|\boldsymbol{q}\|_2}{\|\boldsymbol{p}\|_2} \boldsymbol{p} \right],$$

where $\bar{\theta}^{(k)}_i = g(\bar{\theta}^{(k)}_{i-1})$ *for* g *given by* (8)*,* $\bar{\theta}^{(k)}_0 = \angle(\boldsymbol{p}, \boldsymbol{q})$*,* $a^{(1)} = d$*, and* $a^{(2)} = s$*. For all* $\boldsymbol{h} \neq 0$*,* $\boldsymbol{x} \neq 0$*,* $\boldsymbol{m} \neq 0$ *and* $\boldsymbol{y} \neq 0$*,*

$$\left\| \left(\boldsymbol{\Lambda}^{(1)}_{d,+,\boldsymbol{h}} \right)^{\mathsf{T}} diag \left(\boldsymbol{\Lambda}^{(2)}_{s,+,\boldsymbol{m}} \boldsymbol{m} \odot \boldsymbol{\Lambda}^{(2)}_{s,+,\boldsymbol{y}} \boldsymbol{y} \right) \boldsymbol{\Lambda}^{(1)}_{d,+,\boldsymbol{x}} \boldsymbol{x} - \frac{\alpha}{n} \left(\boldsymbol{m}^{\mathsf{T}} \tilde{\boldsymbol{t}}^{(2)}_{\boldsymbol{m},\boldsymbol{y}} \right) \tilde{\boldsymbol{t}}^{(1)}_{\boldsymbol{h},\boldsymbol{x}} \right\|_2 \leq \frac{208 d^3 s^3 \sqrt{\epsilon}}{2^{d+s} \ell} \|\boldsymbol{x}\|_2 \|\boldsymbol{m}\|_2 \|\boldsymbol{y}\|_2, \tag{18}$$

$$\left\| \left(\boldsymbol{\Lambda}^{(2)}_{s,+,\boldsymbol{m}} \right)^{\mathsf{T}} diag \left(\boldsymbol{\Lambda}^{(1)}_{d,+,\boldsymbol{h}} \boldsymbol{h} \odot \boldsymbol{\Lambda}^{(1)}_{d,+,\boldsymbol{x}} \boldsymbol{x} \right) \boldsymbol{\Lambda}^{(2)}_{s,+,\boldsymbol{y}} \boldsymbol{y} - \frac{\alpha}{n} \left(\boldsymbol{h}^{\mathsf{T}} \tilde{\boldsymbol{t}}^{(1)}_{\boldsymbol{h},\boldsymbol{x}} \right) \tilde{\boldsymbol{t}}^{(2)}_{\boldsymbol{m},\boldsymbol{y}} \right\|_2 \leq \frac{208 d^3 s^3 \sqrt{\epsilon}}{2^{d+s} \ell} \|\boldsymbol{y}\|_2 \|\boldsymbol{h}\|_2 \|\boldsymbol{x}\|_2. \tag{19}$$

Proof. We will prove (18). Proof of (19) is identical to proof of (18). Define $\boldsymbol{h}_0 = \boldsymbol{h}$, $\boldsymbol{x}_0 = \boldsymbol{x}$, $\boldsymbol{m}_0 = \boldsymbol{m}$, $\boldsymbol{y}_0 = \boldsymbol{y}$,

$$\begin{aligned} \boldsymbol{h}_d := \left(\prod_{i=d}^{1} \boldsymbol{W}^{(1)}_{i,+,\boldsymbol{h}} \right) \boldsymbol{h} &= \left(\boldsymbol{W}^{(1)}_{d,+,\boldsymbol{h}} \boldsymbol{W}^{(1)}_{d-1,+,\boldsymbol{h}} \dots \boldsymbol{W}^{(1)}_{1,+,\boldsymbol{h}} \right) \boldsymbol{h} \\ &= \boldsymbol{W}^{(1)}_{d,+,\boldsymbol{h}} \boldsymbol{h}_{d-1} \\ &= (\boldsymbol{W}^{(1)}_d)_{+,\boldsymbol{h}_{d-1}} \boldsymbol{h}_{d-1}, \end{aligned}$$

and analogously $\boldsymbol{x}_d = \left(\prod_{i=d}^{1} \boldsymbol{W}^{(1)}_{i,+,\boldsymbol{x}} \right) \boldsymbol{x}$, $\boldsymbol{m}_s = \left(\prod_{i=s}^{1} \boldsymbol{W}^{(2)}_{i,+,\boldsymbol{m}} \right) \boldsymbol{m}$, and $\boldsymbol{y}_s = \left(\prod_{i=s}^{1} \boldsymbol{W}^{(2)}_{i,+,\boldsymbol{y}} \right) \boldsymbol{y}$. By the WDC, we have for all $\boldsymbol{h} \neq \boldsymbol{0}$, $\boldsymbol{m} \neq \boldsymbol{0}$,

$$\left\| \left(\boldsymbol{W}^{(1)}_i \right)^{\mathsf{T}}_{+,\boldsymbol{h}} \left(\boldsymbol{W}^{(1)}_i \right)_{+,\boldsymbol{h}} - \frac{1}{2} \boldsymbol{I}_{n_{i-1}} \right\| \leq \epsilon \text{ for all } i = 1, \dots, d-1, \text{ and} \tag{20}$$

$$\left\| \left(\boldsymbol{W}^{(2)}_i \right)^{\mathsf{T}}_{+,\boldsymbol{m}} \left(\boldsymbol{W}^{(2)}_i \right)_{+,\boldsymbol{m}} - \frac{1}{2} \boldsymbol{I}_{p_{i-1}} \right\| \leq \epsilon \text{ for all } i = 1, \dots, s-1. \tag{21}$$

In particular, $\left\| \left(\boldsymbol{W}^{(1)}_{i,+,\boldsymbol{h}} \right)^{\mathsf{T}} \boldsymbol{W}^{(1)}_{i,+,\boldsymbol{h}} - \frac{1}{2} \boldsymbol{I}_{n_{i-1}} \right\| \leq \epsilon$ and $\left\| \left(\boldsymbol{W}^{(2)}_{i,+,\boldsymbol{m}} \right)^{\mathsf{T}} \boldsymbol{W}^{(2)}_{i,+,\boldsymbol{m}} - \frac{1}{2} \boldsymbol{I}_{p_{i-1}} \right\| \leq \epsilon$. and consequently,

$$\begin{aligned} \frac{1}{2} - \epsilon \leq \left\| \boldsymbol{W}^{(1)}_{i,+,\boldsymbol{h}} \right\|^2 \leq \frac{1}{2} + \epsilon \\ \frac{1}{2} - \epsilon \leq \left\| \boldsymbol{W}^{(2)}_{i,+,\boldsymbol{m}} \right\|^2 \leq \frac{1}{2} + \epsilon. \end{aligned}$$

Hence,

$$\left\| \prod_{i=d-1}^{1} \boldsymbol{W}^{(1)}_{i,+,\boldsymbol{h}} \right\| \left\| \prod_{i=d-1}^{1} \boldsymbol{W}^{(1)}_{i,+,\boldsymbol{x}} \right\| \leq \frac{1}{2^{d-1}} (1 + 2\epsilon)^{d-1} = \frac{1}{2^{d-1}} e^{(d-1)\log(1+2\epsilon)} \leq \frac{1 + 4\epsilon(d-1)}{2^{d-1}}, \tag{22}$$

where we used that $\log(1+z) \leq z$, $e^z \leq 1 + 2z$ for $z < 1$, and $2(d-1)\epsilon \leq 1$. Similarly,

$$\left\| \prod_{i=s-1}^{1} \boldsymbol{W}^{(2)}_{i,+,\boldsymbol{m}} \right\| \left\| \prod_{i=s-1}^{1} \boldsymbol{W}^{(2)}_{i,+,\boldsymbol{y}} \right\| \leq \frac{1 + 4\epsilon(s-1)}{2^{s-1}}. \tag{23}$$

Let

$$\tilde{\boldsymbol{h}} = \boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\boldsymbol{h}, \quad \tilde{\boldsymbol{x}} = \boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x}, \quad \tilde{\boldsymbol{m}} = \boldsymbol{\Lambda}^{(2)}_{s-1,+,\boldsymbol{m}}\boldsymbol{m}, \quad \tilde{\boldsymbol{y}} = \boldsymbol{\Lambda}^{(2)}_{s-1,+,\boldsymbol{y}}\boldsymbol{y},$$

and consider

$$\begin{aligned}
&\left\| \left(\boldsymbol{\Lambda}^{(1)}_{d,+,\boldsymbol{h}}\right)^{\mathsf{T}} \operatorname{diag}\left(\boldsymbol{\Lambda}^{(2)}_{s,+,\boldsymbol{m}}\boldsymbol{m} \odot \boldsymbol{\Lambda}^{(2)}_{s,+,\boldsymbol{y}}\boldsymbol{y}\right)\boldsymbol{\Lambda}^{(1)}_{d,+,\boldsymbol{x}}\boldsymbol{x} - \frac{\alpha}{\ell}\left(\boldsymbol{m}^{\mathsf{T}}\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\right)\tilde{t}^{(1)}_{\boldsymbol{h},\boldsymbol{x}} \right\|_2 \\
\leq &\Big\| \left(\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right)^{\mathsf{T}} \Big(\left(\boldsymbol{W}^{(1)}_{d,+,\boldsymbol{h}}\right)^{\mathsf{T}} \operatorname{diag}\Big(\boldsymbol{W}^{(2)}_{d,+,\boldsymbol{m}}\tilde{\boldsymbol{m}} \odot \boldsymbol{W}^{(2)}_{d,+,\boldsymbol{y}}\tilde{\boldsymbol{y}}\Big)\boldsymbol{W}^{(1)}_{d,+,\boldsymbol{x}} - \frac{\alpha}{\ell}\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}} \\
&\quad \tilde{\boldsymbol{m}}^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{y}}}\tilde{\boldsymbol{y}}\Big)\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x}\Big\|_2 + \Big\| \frac{\alpha}{\ell}\left(\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right)^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}}\tilde{\boldsymbol{m}}^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{y}}}\tilde{\boldsymbol{y}}\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x} \\
&\quad - \frac{\alpha}{\ell}\left(\boldsymbol{m}^{\mathsf{T}}\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\right)\tilde{t}^{(1)}_{\boldsymbol{h},\boldsymbol{x}}\Big\|_2 \\
\leq &\Big\| \left(\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right)^{\mathsf{T}} \Big(\left(\boldsymbol{W}^{(1)}_{d,+,\boldsymbol{h}}\right)^{\mathsf{T}} \operatorname{diag}\Big(\boldsymbol{W}^{(2)}_{d,+,\boldsymbol{m}}\tilde{\boldsymbol{m}} \odot \boldsymbol{W}^{(2)}_{d,+,\boldsymbol{y}}\tilde{\boldsymbol{y}}\Big)\boldsymbol{W}^{(1)}_{d,+,\boldsymbol{x}} - \frac{\alpha}{\ell}\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}} \\
&\quad \tilde{\boldsymbol{m}}^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{y}}}\tilde{\boldsymbol{y}}\Big)\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x}\Big\|_2 \\
&\quad + \frac{\alpha}{\ell}\Big\|\tilde{\boldsymbol{m}}^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{y}}}\tilde{\boldsymbol{y}}\left(\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right)^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}}\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x} - \boldsymbol{m}^{\mathsf{T}}\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\left(\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right)^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}}\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x}\Big\|_2 \\
&\quad + \frac{\alpha}{\ell}\Big\|\boldsymbol{m}^{\mathsf{T}}\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\left(\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right)^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}}\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x} - \boldsymbol{m}^{\mathsf{T}}\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\tilde{t}^{(1)}_{\boldsymbol{h},\boldsymbol{x}}\Big\|_2 \qquad (24)
\end{aligned}$$

where both the first and second inequality holds because of triangle inequality. We bound the terms in the inequality above separately. First consider

$$\begin{aligned}
&\Big\| \left(\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right)^{\mathsf{T}} \Big(\left(\boldsymbol{W}^{(1)}_{d,+,\boldsymbol{h}}\right)^{\mathsf{T}} \operatorname{diag}\Big(\boldsymbol{W}^{(2)}_{d,+,\boldsymbol{m}}\tilde{\boldsymbol{m}} \odot \boldsymbol{W}^{(2)}_{d,+,\boldsymbol{y}}\tilde{\boldsymbol{y}}\Big)\boldsymbol{W}^{(1)}_{d,+,\boldsymbol{x}} - \frac{\alpha}{\ell}\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}} \\
&\quad \tilde{\boldsymbol{m}}^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{y}}}\tilde{\boldsymbol{y}}\Big)\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x}\Big\|_2 \\
\leq &\Big\| \left(\boldsymbol{W}^{(1)}_{d,+,\boldsymbol{h}}\right)^{\mathsf{T}} \operatorname{diag}\Big(\boldsymbol{W}^{(2)}_{d,+,\boldsymbol{m}}\tilde{\boldsymbol{m}} \odot \boldsymbol{W}^{(2)}_{d,+,\boldsymbol{y}}\tilde{\boldsymbol{y}}\Big)\boldsymbol{W}^{(1)}_{d,+,\boldsymbol{x}} - \frac{\alpha}{\ell}\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}}\tilde{\boldsymbol{m}}^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{y}}}\tilde{\boldsymbol{y}}\Big)\Big\| \\
&\quad \left\|\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right\| \left\|\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\right\| \|\boldsymbol{x}\|_2 \\
\leq &\left(\frac{1+4\epsilon(d-1)}{2^{d-1}}\right)\frac{4\epsilon}{\ell}\|\boldsymbol{x}\|_2\|\tilde{\boldsymbol{m}}\|_2\|\tilde{\boldsymbol{y}}\|_2 \\
= &\frac{(1+4\epsilon(d-1))}{2^d}\frac{8\epsilon}{\ell}\|\boldsymbol{x}\|_2\|\boldsymbol{\Lambda}^{(2)}_{s-1,+,\boldsymbol{m}}\boldsymbol{m}\|_2\|\boldsymbol{\Lambda}^{(2)}_{s-1,+,\boldsymbol{y}}\boldsymbol{y}\|_2 \\
\leq &\frac{(1+4\epsilon(d-1))}{2^d}\frac{(1+4\epsilon(s-1))}{2^s}\frac{16\epsilon}{\ell}\|\boldsymbol{x}\|_2\|\boldsymbol{m}\|_2\|\boldsymbol{y}\|_2 \\
\leq &\frac{64\epsilon}{2^{d+s}\ell}\|\boldsymbol{x}\|_2\|\boldsymbol{m}\|_2\|\boldsymbol{y}\|_2. \qquad (25)
\end{aligned}$$

where the first inequality holds because spectral norm is a sub-multiplicative norm. The second inequality holds because of (22) and joint-WDC. The last inequality holds if $4\epsilon(d-1) < 1$ and $4\epsilon(s-1) < 1$.

Second, consider

$$\begin{aligned}
&\Big\|\tilde{\boldsymbol{m}}^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{y}}}\tilde{\boldsymbol{y}}\left(\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right)^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}}\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x} - \boldsymbol{m}^{\mathsf{T}}\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\left(\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right)^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}}\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x}\Big\|_2 \\
= &\Big\|\left(\tilde{\boldsymbol{m}}^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{y}}}\tilde{\boldsymbol{y}} - \boldsymbol{m}^{\mathsf{T}}\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\right)\left(\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right)^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}}\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x}\Big\|_2 \\
\leq &\frac{1+4\epsilon(d-1)}{2^{d-1}}\|\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}}\|\Big\|\left(\boldsymbol{\Lambda}^{(2)}_{s-1,+,\boldsymbol{m}}\right)^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{y}}}\boldsymbol{\Lambda}^{(2)}_{s-1,+,\boldsymbol{y}}\boldsymbol{y} - \tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\Big\|_2\|\boldsymbol{x}\|_2\|\boldsymbol{m}\|_2 \\
= &\frac{1+4\epsilon(d-1)}{2^d}\Big\|\left(\boldsymbol{\Lambda}^{(2)}_{s-1,+,\boldsymbol{m}}\right)^{\mathsf{T}}\left(\boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{y}}} - \frac{1}{\alpha_2}\left(\boldsymbol{W}^{(2)}_{d,+,\boldsymbol{m}}\right)^{\mathsf{T}}\boldsymbol{W}^{(2)}_{d,+,\boldsymbol{y}}\right)\boldsymbol{\Lambda}^{(2)}_{s-1,+,\boldsymbol{y}}\boldsymbol{y} \\
&\quad + \frac{1}{\alpha_2}\left(\boldsymbol{\Lambda}^{(2)}_{s,+,\boldsymbol{m}}\right)^{\mathsf{T}}\boldsymbol{\Lambda}^{(2)}_{s,+,\boldsymbol{y}}\boldsymbol{y} - \tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\Big\|_2\|\boldsymbol{x}\|_2\|\boldsymbol{m}\|_2 \\
\leq &\frac{1+4\epsilon(d-1)}{2^d}\left(\left\|\boldsymbol{\Lambda}^{(2)}_{s-1,+,\boldsymbol{m}}\right\|\left\|\boldsymbol{\Lambda}^{(2)}_{s-1,+,\boldsymbol{y}}\right\|\left\|\left(\boldsymbol{Q}_{\tilde{\boldsymbol{m}},\tilde{\boldsymbol{y}}} - \frac{1}{\alpha_2}\left(\boldsymbol{W}^{(2)}_{d,+,\boldsymbol{m}}\right)^{\mathsf{T}}\boldsymbol{W}^{(2)}_{d,+,\boldsymbol{y}}\right)\right\|\|\boldsymbol{y}\|_2\right.
\end{aligned}$$

$$
\begin{aligned}
&+\Big\|\frac{1}{\alpha_2}\left(\boldsymbol{\Lambda}^{(2)}_{s,+,\boldsymbol{m}}\right)^{\mathsf{T}}\boldsymbol{\Lambda}^{(2)}_{s,+,\boldsymbol{y}}\boldsymbol{y}-\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\Big\|_2\Big)\|\boldsymbol{x}\|_2\|\boldsymbol{m}\|_2\\
\leq&\frac{1+4\epsilon(d-1)}{2^{d+s}}\Big(6(1+4\epsilon(s-1))\epsilon/\alpha_2+24s^3\sqrt{\epsilon/\alpha_2}\Big)\|\boldsymbol{x}\|_2\|\boldsymbol{m}\|_2\|\boldsymbol{y}\|_2\\
\leq&\frac{2}{2^{d+s}}\Big(12\epsilon/\alpha_2+24s^3\sqrt{\epsilon/\alpha_2}\Big)\|\boldsymbol{x}\|_2\|\boldsymbol{m}\|_2\|\boldsymbol{y}\|_2\\
\leq&\frac{72s^3\sqrt{\epsilon}}{2^{2d}\alpha_2}\|\boldsymbol{x}\|_2\|\boldsymbol{m}\|_2\|\boldsymbol{y}\|_2. \qquad (26)
\end{aligned}
$$

where the first inequality holds because of (22). The second inequality holds because of triangle inequality. The third inequality holds because of (23), $\frac{1}{\sqrt{\alpha_2}}\boldsymbol{W}^{(2)}_d$ satisfy WDC with constant ϵ/α_2 and 1, and Lemma 1.

Third, consider

$$
\begin{aligned}
&\Big\|\boldsymbol{m}^{\mathsf{T}}\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\left(\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right)^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}}\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x}-\boldsymbol{m}^{\mathsf{T}}\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\tilde{t}^{(1)}_{\boldsymbol{h},\boldsymbol{x}}\Big\|_2\\
=&|\boldsymbol{m}^{\mathsf{T}}\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}|\Big\|\left(\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right)^{\mathsf{T}}\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}}\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x}-\tilde{t}^{(1)}_{\boldsymbol{h},\boldsymbol{x}}\Big\|_2\\
\leq&\|\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\|_2\Big\|\left(\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\right)^{\mathsf{T}}\left(\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}}-\frac{1}{\alpha_1}\left(\boldsymbol{W}^{(1)}_{d,+,\boldsymbol{h}}\right)^{\mathsf{T}}\boldsymbol{W}^{(1)}_{d,+,\boldsymbol{x}}\right)\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\boldsymbol{x}\\
&+\frac{1}{\alpha_1}\left(\boldsymbol{\Lambda}^{(1)}_{d,+,\boldsymbol{h}}\right)^{\mathsf{T}}\boldsymbol{\Lambda}^{(1)}_{d,+,\boldsymbol{x}}-\tilde{t}^{(1)}_{\boldsymbol{h},\boldsymbol{x}}\Big\|_2\|\boldsymbol{m}\|_2\\
\leq&\frac{1+s}{2^s}\Big(\Big\|\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{h}}\Big\|\Big\|\boldsymbol{\Lambda}^{(1)}_{d-1,+,\boldsymbol{x}}\Big\|\Big\|\boldsymbol{Q}_{\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}}}-\frac{1}{\alpha_1}\left(\boldsymbol{W}^{(1)}_{d,+,\boldsymbol{h}}\right)^{\mathsf{T}}\boldsymbol{W}^{(1)}_{d,+,\boldsymbol{x}}\Big\|\|\boldsymbol{x}\|_2\\
&+\Big\|\frac{1}{\alpha_1}\left(\boldsymbol{\Lambda}^{(1)}_{d,+,\boldsymbol{h}}\right)^{\mathsf{T}}\boldsymbol{\Lambda}^{(1)}_{d,+,\boldsymbol{x}}\boldsymbol{x}-\tilde{t}^{(1)}_{\boldsymbol{h},\boldsymbol{x}}\Big\|_2\Big)\|\boldsymbol{m}\|_2\|\boldsymbol{y}\|_2\\
\leq&\frac{2s}{2^{d+s}}\Big(6(1+4\epsilon(d-1))\epsilon/\alpha_1+24d^3\sqrt{\epsilon/\alpha_1}\Big)\|\boldsymbol{x}\|_2\|\boldsymbol{m}\|_2\|\boldsymbol{y}\|_2\\
\leq&\frac{72sd^3\sqrt{\epsilon}}{2^{d+s}\alpha_1}\|\boldsymbol{x}\|_2\|\boldsymbol{m}\|_2\|\boldsymbol{y}\|_2. \qquad (27)
\end{aligned}
$$

where the first inequality holds because of Cauchy-Schwartz inequality. The second inequality holds because of triangle inequality along with $\|\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\|_2\leq\frac{1+s}{2^s}$. The third inequality holds because of (22), $\frac{1}{\sqrt{\alpha_1}}\boldsymbol{W}^{(1)}_d$ satisfy WDC with constant ϵ/α_1 and 1, and Lemma 1.

Hence, combining (24), (25), (26), and (27), we get

$$
\begin{aligned}
&\Big\|\left(\boldsymbol{\Lambda}^{(1)}_{d,+,\boldsymbol{h}}\right)^{\mathsf{T}}\mathrm{diag}\left(\boldsymbol{\Lambda}^{(2)}_{s,+,\boldsymbol{m}}\boldsymbol{m}\odot\boldsymbol{\Lambda}^{(2)}_{s,+,\boldsymbol{y}}\boldsymbol{y}\right)\boldsymbol{\Lambda}^{(1)}_{d,+,\boldsymbol{x}}\boldsymbol{x}-\frac{\alpha_1\alpha_2}{\ell}\left(\boldsymbol{m}^{\mathsf{T}}\tilde{t}^{(2)}_{\boldsymbol{m},\boldsymbol{y}}\right)\tilde{t}^{(1)}_{\boldsymbol{h},\boldsymbol{x}}\Big\|_2\\
\leq&\Big(\frac{64\epsilon}{2^{d+s}\ell}+\frac{72s^3\sqrt{\epsilon}}{2^{d+s}\ell}+\frac{72sd^3\sqrt{\epsilon}}{2^{2d}\ell}\Big)\|\boldsymbol{x}\|_2\|\boldsymbol{m}\|_2\|\boldsymbol{y}\|_2\\
\leq&\frac{208s^3d^3\sqrt{\epsilon}}{2^{d+s}\ell}\|\boldsymbol{x}\|_2\|\boldsymbol{m}\|_2\|\boldsymbol{y}\|_2.
\end{aligned}
$$

□

5.3 Zeros of $t_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}$

Lemma 3. *Fix* $0<\epsilon<1$ *and* $0<\alpha\leq 1$ *such that* $38(d^5+s^5)\sqrt{\epsilon}/\alpha<1$. *Let* $\mathcal{K}=\{(\boldsymbol{h},\boldsymbol{0})\in\mathbb{R}^{n\times p}\,|\boldsymbol{h}\in\mathbb{R}^n\}\cup\{(\boldsymbol{0},\boldsymbol{m})\in\mathbb{R}^{n\times p}\big|\boldsymbol{m}\in\mathbb{R}^p\}$. *Let*

$$
S^{(1)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)}=\left\{(\boldsymbol{h},\boldsymbol{x})\in\mathbb{R}^{n\times p}\backslash\mathcal{K}\,\middle|\,\frac{\|\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}\|_2}{\|\boldsymbol{m}\|_2}\leq\frac{\epsilon\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2,\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)}{2^{d+s}\ell}\right\},
$$

$$
S^{(2)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)}=\left\{(\boldsymbol{h},\boldsymbol{x})\in\mathbb{R}^{n\times p}\backslash\mathcal{K}\,\middle|\,\frac{\|\boldsymbol{t}^{(2)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}\|_2}{\|\boldsymbol{h},\|_2}\leq\frac{\epsilon\max(\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2,\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2)}{2^{d+s}\ell}\right\},
$$

where d and s are an integers greater than 1. Let

$$\tilde{\boldsymbol{t}}^{(k)}_{\boldsymbol{m},\boldsymbol{y}} = \frac{1}{2^{a^{(k)}}}\left(\prod_{i=0}^{a^{(k)}-1}\frac{\pi-\bar{\theta}_i^{(k)}}{\pi}\boldsymbol{y} + \sum_{i=0}^{a^{(k)}-1}\frac{\sin\bar{\theta}_i^{(k)}}{\pi}\left(\prod_{j=i+1}^{a^{(k)}-1}\frac{\pi-\bar{\theta}_j^{(k)}}{\pi}\right)\frac{\|\boldsymbol{y}\|_2}{\|\boldsymbol{m}\|_2}\boldsymbol{m}\right), \tag{28}$$

where $\bar{\theta}_i^{(k)} = g(\bar{\theta}_{i-1}^{(k)})$ for g given in (8)*, $\bar{\theta}_0^{(k)} = \angle(\boldsymbol{m},\boldsymbol{y})$, $a^{(1)} = d$, and $a^{(2)} = s$. Let $\boldsymbol{t}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)} = \begin{bmatrix} \boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)} \\ \boldsymbol{t}^{(2)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)} \end{bmatrix}$ where*

$$\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)} = \frac{\alpha}{2^{d+s}\ell}\|\boldsymbol{m}\|_2^2\boldsymbol{h} - \frac{\alpha}{\ell}\boldsymbol{m}^{\mathsf{T}}\tilde{\boldsymbol{t}}^{(2)}_{\boldsymbol{m},\boldsymbol{m}_0}\tilde{\boldsymbol{t}}^{(1)}_{\boldsymbol{h},\boldsymbol{h}_0}, \tag{29}$$

$$\boldsymbol{t}^{(2)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)} = \frac{\alpha}{2^{d+s}\ell}\|\boldsymbol{h}\|_2^2\boldsymbol{m} - \frac{\alpha}{\ell}\boldsymbol{h}^{\mathsf{T}}\tilde{\boldsymbol{t}}^{(1)}_{\boldsymbol{h},\boldsymbol{h}_0}\tilde{\boldsymbol{t}}^{(2)}_{\boldsymbol{m},\boldsymbol{m}_0}. \tag{30}$$

Define

$$\rho^{(k)}_{a^{(k)}} := \sum_{i=1}^{a^{(k)}-1}\frac{\sin\check{\theta}_i^{(k)}}{\pi}\left(\prod_{j=i+1}^{a^{(k)}-1}\frac{\pi-\check{\theta}_j^{(k)}}{\pi}\right),$$

where $\check{\theta}_0^{(k)} = \pi$ and $\check{\theta}_i^{(k)} = g(\check{\theta}_{i-1}^{(k)})$. If $(\boldsymbol{h},\boldsymbol{m}) \in S^{(1)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)} \cap S^{(2)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)}$ then one of the following holds:

- $\left|\bar{\theta}_0^{(1)}\right| \le 2\sqrt{\epsilon},\ \left|\bar{\theta}_0^{(2)}\right| \le 2\sqrt{\epsilon}$ *and*

$$|\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2 - \|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2| \le 145\frac{ds\sqrt{\epsilon}}{\alpha}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2,$$

- $\left|\bar{\theta}_0^{(1)} - \pi\right| \le 12\pi^2 d^3\sqrt{\epsilon}/\alpha,\ \left|\bar{\theta}_0^{(2)}\right| \le 1.5\sqrt{\epsilon}$ *and*

$$\left|\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2 - \rho_d^{(1)}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2\right| \le 532\frac{d^6 s\sqrt{\epsilon}}{\alpha}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2,$$

- $\left|\bar{\theta}_0^{(1)}\right| \le 2\sqrt{\epsilon},\ \left|\bar{\theta}_0^{(2)} - \pi\right| \le 12\pi^2 s^3\sqrt{\epsilon}/\alpha$ *and*

$$\left|\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2 - \rho_d^{(1)}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2\right| \le 532\frac{ds^6\sqrt{\epsilon}}{\alpha}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2,$$

- $\left|\bar{\theta}_0^{(1)} - \pi\right| \le 12\pi^2 d^3\sqrt{\epsilon}\alpha,\ \left|\bar{\theta}_0^{(2)} - \pi\right| \le 12\pi^2 d^3\sqrt{\epsilon}/\alpha$ *and*

$$\left|\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2 - \rho_d^{(1)}\rho_s^{(2)}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2\right| \le 3915\frac{d^6 s^6\sqrt{\epsilon}}{\alpha}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2.$$

In particular,

$$S^{(1)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)} \cap S^{(2)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)} \subseteq \mathcal{A}_{437\frac{ds\sqrt{\epsilon}}{\alpha},(\boldsymbol{h}_0,\boldsymbol{m}_0)} \cup \mathcal{A}_{2436\pi\frac{d^6 s\sqrt{\epsilon}}{\alpha},\left(-\rho_d^{(1)}\boldsymbol{h}_0,\boldsymbol{m}_0\right)} \cup \mathcal{A}_{2436\pi\frac{ds^6\sqrt{\epsilon}}{\alpha},\left(\rho_s^{(2)}\boldsymbol{h}_0,-\boldsymbol{m}_0\right)} \cup \mathcal{A}_{16767\pi^2\frac{d^6 s^6\sqrt{\epsilon}}{\alpha},\left(-\rho_d^{(1)}\rho_s^{(2)}\boldsymbol{h}_0,-\boldsymbol{m}_0\right)},$$

where $\mathcal{A}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)}$ is defined in (3)*.*

Proof. Without loss of generality, let $\boldsymbol{h}_0 = \boldsymbol{e}_1$, $\boldsymbol{m}_0 = \boldsymbol{e}_1$, $\hat{\boldsymbol{h}} = \cos\bar{\theta}_0^{(1)} + \sin\bar{\theta}_0^{(1)}$ and $\hat{\boldsymbol{m}} = \cos\bar{\theta}_0^{(2)} + \sin\bar{\theta}_0^{(2)}$ for some $\bar{\theta}_i^{(1)}, \bar{\theta}_0^{(2)} \in [0,\pi]$. First we introduce some notation for convenience. Let

$$\xi^{(k)} = \prod_{i=0}^{a^{(k)}-1}\frac{\pi-\bar{\theta}_i^{(k)}}{\pi}, \quad \zeta^{(k)} = \sum_{i=1}^{a^{(k)}-1}\frac{\sin\bar{\theta}_i^{(k)}}{\pi}\prod_{j=i+1}^{a^{(k)}-1}\frac{\pi-\bar{\theta}_j^{(k)}}{\pi},$$

$r^{(1)} = \|\boldsymbol{h}\|_2$, $r^{(2)} = \|\boldsymbol{m}\|_2$, and $M = \max(r^{(1)}r^{(2)}, 1)$. Using these notation, we can rewrite $\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}$ as

$$\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}$$

$$
\begin{aligned}
=&\frac{\alpha}{2^{d+s}\ell}\left(\|\boldsymbol{m}\|_2\boldsymbol{h}-\left(\xi^{(2)}\cos\bar{\theta}_0^{(2)}+\zeta^{(2)}\right)\left(\xi^{(1)}\frac{\boldsymbol{h}_0}{\|\boldsymbol{h}_0\|_2}+\zeta^{(1)}\frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right)\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2\right)\|\boldsymbol{m}\|_2\\
=&\frac{\alpha}{2^{d+s}\ell}\left(\|\boldsymbol{m}\|_2\boldsymbol{h}-\cos\bar{\theta}_s^{(2)}\left(\xi^{(1)}\frac{\boldsymbol{h}_0}{\|\boldsymbol{h}_0\|_2}+\zeta^{(1)}\frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right)\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2\right)\|\boldsymbol{m}\|_2\\
=&\frac{\alpha}{2^{d+s}\ell}\Bigg(r^{(1)}r^{(2)}\left(\cos\bar{\theta}_0^{(1)}\boldsymbol{e}_1+\sin\bar{\theta}_0^{(1)}\boldsymbol{e}_2)\right)\\
&\quad-\cos\bar{\theta}_s^{(2)}\Big(\xi^{(1)}\boldsymbol{e}_1+\zeta^{(1)}\left(\cos\bar{\theta}_0^{(1)}\boldsymbol{e}_1+\sin\bar{\theta}_0^{(1)}\boldsymbol{e}_2\right)\Big)\Bigg)r^{(2)}.
\end{aligned}
$$

By inspecting the components of $\boldsymbol{t}^{(1)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}$, we have that $(\boldsymbol{h},\boldsymbol{m})\in S^{(1)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)}$ implies

$$
\left|r^{(1)}r^{(2)}\cos\bar{\theta}_0^{(1)}-\cos\bar{\theta}_s^{(2)}\left(\xi^{(1)}+\zeta^{(1)}\cos\bar{\theta}_0^{(1)}\right)\right|\leq\frac{\epsilon M}{\alpha} \tag{31}
$$

$$
\left|r^{(1)}r^{(2)}\sin\bar{\theta}_0^{(1)}-\cos\bar{\theta}_s^{(2)}\zeta^{(1)}\sin\bar{\theta}_0^{(1)}\right|\leq\frac{\epsilon M}{\alpha} \tag{32}
$$

Similarly, by inspecting the components of $\boldsymbol{t}^{(2)}_{(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{h}_0,\boldsymbol{m}_0)}$, we have that $(\boldsymbol{h},\boldsymbol{m})\in S^{(2)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)}$ implies

$$
\left|r^{(1)}r^{(2)}\cos\bar{\theta}_0^{(2)}-\cos\bar{\theta}_d^{(1)}\left(\xi^{(2)}+\zeta^{(2)}\cos\bar{\theta}_0^{(2)}\right)\right|\leq\frac{\epsilon M}{\alpha} \tag{33}
$$

$$
\left|r^{(1)}r^{(2)}\sin\bar{\theta}_0^{(2)}-\cos\bar{\theta}_d^{(1)}\zeta^{(2)}\sin\bar{\theta}_0^{(2)}\right|\leq\frac{\epsilon M}{\alpha} \tag{34}
$$

Now, we record several properties. We have:

$$
\bar{\theta}_i^{(k)}\in[0,\pi/2]\text{ for }i\geq1 \tag{35}
$$

$$
\bar{\theta}_i^{(k)}\leq\bar{\theta}_{i-1}^{(k)}\text{ for }i\geq1 \tag{36}
$$

$$
|\xi^{(k)}|\leq1 \tag{37}
$$

$$
\check{\theta}_i^{(k)}\leq\frac{3\pi}{i+3}\text{ for }i\geq0 \tag{38}
$$

$$
\check{\theta}_i^{(k)}\geq\frac{\pi}{i+1}\text{ for }i\geq0 \tag{39}
$$

$$
\xi^{(k)}=\prod_{i=1}^{a^{(k)}-1}\frac{\pi-\bar{\theta}_i^{(k)}}{\pi}\geq\frac{\pi-\bar{\theta}_0^{(k)}}{\pi}{a^{(k)}}^{-3} \tag{40}
$$

$$
\bar{\theta}_0^{(k)}=\pi+O_1(\delta)\Rightarrow\bar{\theta}_i^{(k)}=\check{\theta}_i^{(k)}+O_1(i\delta) \tag{41}
$$

$$
\bar{\theta}_0^{(k)}=\pi+O_1(\delta)\Rightarrow|\xi^{(k)}|\leq\frac{\delta}{\pi} \tag{42}
$$

$$
\bar{\theta}_0^{(k)}=\pi+O_1(\delta)\Rightarrow\zeta^{(k)}=\rho_d^{(k)}+O_1(3{a^{(k)}}^3\delta)\text{ if }\frac{{a^{(k)}}^2\delta}{\pi}\leq1 \tag{43}
$$

$$
|\zeta^{(k)}|=|\xi^{(k)}\cos\bar{\theta}_0^{(k)}-\cos\bar{\theta}_{a^{(k)}}^{(k)}|\leq2 \tag{44}
$$

$$
\cos\bar{\theta}_i^{(k)}\geq\frac{1}{\pi}\text{ for }i\geq2 \tag{45}
$$

For a proof of (37)-(43), we refer the readers to Lemma 8 of Hand and Voroninski [2017]. Also, we note that (45) follows directly from (44).

We first show that if $(\boldsymbol{h},\boldsymbol{m})\in S^{(1)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)}$ then $r^{(1)}r^{(2)}\leq6$, and thus $M\leq6$. Suppose $r^{(1)}r^{(2)}>1$. At least one of the following holds:$|\sin\bar{\theta}_0^{(1)}|\geq1/\sqrt{2}$ or $|\cos\bar{\theta}_0^{(1)}|\geq1/\sqrt{2}$. If $|\sin\bar{\theta}_0^{(1)}|\geq1/\sqrt{2}$ then (32) implies that $\left|r^{(1)}r^{(2)}-\cos\bar{\theta}_s^{(2)}\zeta^{(1)}\right|\leq\sqrt{2}\epsilon r^{(1)}r^{(2)}/\alpha$. Using (44), we get $r^{(1)}r^{(2)}\leq\frac{2}{1-\sqrt{2}\epsilon/\alpha}\leq4$ if $\epsilon/\alpha<1/4$. If $|\cos\bar{\theta}_0^{(1)}|\geq1/\sqrt{2}$, then (31) implies $\left|r^{(1)}r^{(2)}-\cos\bar{\theta}_s^{(2)}\zeta^{(1)}\right|\leq\sqrt{2}\left(\epsilon r^{(1)}r^{(2)}/\alpha+|\xi^{(1)}|\right)$. Using (37), (44), and $\epsilon/\alpha<1/4$, we get

$r^{(1)}r^{(2)} \le \frac{\sqrt{2}|\xi^{(k)}|+\cos\bar{\theta}_s^{(2)}\zeta^{(1)}}{1-\sqrt{2}\epsilon/\alpha} \le \frac{2+\sqrt{2}}{1-\sqrt{2}\epsilon/\alpha} \le 6$. Thus, we have $(\boldsymbol{h},\boldsymbol{m}) \in S^{(1)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)} \Rightarrow r^{(1)}r^{(2)} \le 6 \Rightarrow M \le 6$. Similarly, we have $(\boldsymbol{h},\boldsymbol{m}) \in S^{(2)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)} \Rightarrow r^{(1)}r^{(2)} \le 6 \Rightarrow M \le 6$.

Next we establish that we only need to consider the small angle case and the large angle case (i.e. $\bar{\theta}_0^{(k)} \approx 0$ or π) if $(\boldsymbol{h},\boldsymbol{m}) \in S^{(1)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)} \cap S^{(2)}_{\epsilon,(\boldsymbol{h}_0,\boldsymbol{m}_0)}$. Exactly one of the following holds: $\left|r^{(1)}r^{(2)} - \cos\bar{\theta}_s^{(2)}\zeta^{(1)}\right| \ge \sqrt{\epsilon}M/\alpha$ or $\left|r^{(1)}r^{(2)} - \cos\bar{\theta}_s^{(2)}\zeta^{(1)}\right| < \sqrt{\epsilon}M/\alpha$. If $\left|r^{(1)}r^{(2)} - \cos\bar{\theta}_s^{(2)}\zeta^{(1)}\right| \ge \sqrt{\epsilon}M/\alpha$, then by (32), we have $|\sin\bar{\theta}_0^{(1)}| \le \sqrt{\epsilon}$. Hence $\bar{\theta}_0^{(1)} = O_1(2\sqrt{\epsilon})$ or $\bar{\theta}_0^{(1)} = \pi + O_1(2\sqrt{\epsilon})$, as $\epsilon < 1$. If $\left|r^{(1)}r^{(2)} - \cos\bar{\theta}_s^{(2)}\zeta^{(1)}\right| < \sqrt{\epsilon}M/\alpha$, then by (31) and (45) we have $\left|\xi^{(1)}\right| \le 2\pi\sqrt{\epsilon}M/\alpha$. Using (40), we get $\bar{\theta}_0^{(1)} = \pi + O_1(2\pi^2d^3\sqrt{\epsilon}M/\alpha)$. Thus, we only need to consider the small angle case, $\bar{\theta}_0^{(1)} = O_1(2\sqrt{\epsilon})$ and the large angle case $\bar{\theta}_0^{(1)} = \pi + O_1(12\pi^2d^3\sqrt{\epsilon}/\alpha)$, where we have used $M \le 6$. Similarly, we only need to consider the small angle case, $\bar{\theta}_0^{(2)} = O_1(2\sqrt{\epsilon})$ and the large angle case $\bar{\theta}_0^{(2)} = \pi + O_1(12\pi^2s^3\sqrt{\epsilon}/\alpha)$.

Case 1: $\bar{\boldsymbol{\theta}}_{\mathbf{0}}^{(\mathbf{1})} \approx \mathbf{0}$ and $\bar{\boldsymbol{\theta}}_{\mathbf{0}}^{(\mathbf{2})} \approx \mathbf{0}$. Assume $\bar{\theta}_0^{(k)} = O_1(2\sqrt{\epsilon})$. As $\bar{\theta}_i^{(k)} \le \bar{\theta}_0^{(k)} \le 2\sqrt{\epsilon}$ for all i, we have $\xi^{(k)} \ge \left(1 - \frac{2\sqrt{\epsilon}}{\pi}\right)^{a^{(k)}} = 1 + O_1(\frac{4a^{(k)}\sqrt{\epsilon}}{\pi})$ provided $2a^{(k)}\sqrt{\epsilon} \le 1/2$. By (44), we also have $\zeta^{(k)} = O_1(\frac{a^{(k)}}{\pi}2\sqrt{\epsilon}) = O_1(a^{(k)}\sqrt{\epsilon})$. By (31), we have

$$\left|r^{(1)}r^{(2)}\cos\bar{\theta}_0^{(1)} - \left(\xi^{(2)}\cos\bar{\theta}_0^{(2)} + \zeta^{(2)}\right)\left(\xi^{(1)} + \zeta^{(1)}\cos\bar{\theta}_0^{(1)}\right)\right| \le \frac{\epsilon M}{\alpha}$$

where we used $\cos\bar{\theta}_{a^{(k)}}^{(k)} = \xi^{(k)}\cos\bar{\theta}_0^{(k)} + \zeta^{(k)}$. As $\cos\bar{\theta}_0^{(k)} = 1 + O_1((\bar{\theta}_0^{(k)})^2/2) = 1 + O_1(2\epsilon)$,

$$\begin{aligned}\xi^{(2)}\cos\bar{\theta}_0^{(2)} + \zeta^{(2)} =& 1 + O_1(8s\epsilon\sqrt{\epsilon} + 4s\sqrt{\epsilon} + 2\epsilon + s\sqrt{\epsilon}) = 1 + O_1(15s\sqrt{\epsilon}),\\ \xi^{(1)} + \zeta^{(1)}\cos\bar{\theta}_0^{(1)} =& 1 + O_1(4d\sqrt{\epsilon} + 2d\epsilon\sqrt{\epsilon} + d\sqrt{\epsilon}) = 1 + O_1(7d\sqrt{\epsilon}).\end{aligned}$$

Thus,

$$r^{(1)}r^{(2)} = 1 + O_1(12\epsilon + 6\epsilon/\alpha + 105ds\epsilon + 7d\sqrt{\epsilon} + 15s\sqrt{\epsilon}) = 1 + O_1(145ds\sqrt{\epsilon}/\alpha). \quad (46)$$

We now show $(\boldsymbol{h},\boldsymbol{m})$ is close to $\left(c\boldsymbol{h}_0, \frac{1}{c}\boldsymbol{m}_0\right)$, where $c = \frac{\|\boldsymbol{m}_0\|_2}{\|\boldsymbol{m}\|_2}$. Consider

$$\begin{aligned}&\left\|\boldsymbol{h} - \frac{\|\boldsymbol{m}_0\|_2}{\|\boldsymbol{m}\|_2}\boldsymbol{h}_0\right\|_2\\ \le&\frac{1}{\|\boldsymbol{m}\|_2}\left(\left|\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2 - \|\boldsymbol{m}_0\|_2\|\boldsymbol{h}_0\|_2\right| + \left(\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2 + \left|\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2 - \|\boldsymbol{m}_0\|_2\|\boldsymbol{h}_0\|_2\right|\right)\bar{\theta}_0^{(1)}\right)\\ \le&\frac{1}{\|\boldsymbol{m}\|_2}\left(145ds\sqrt{\epsilon}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2/\alpha + \left(\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2 + 145ds\sqrt{\epsilon}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2/\alpha\right)2\sqrt{\epsilon}\right)\\ \le&437\frac{ds\sqrt{\epsilon}}{\alpha}\frac{\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2}{\|\boldsymbol{m}\|_2}.\end{aligned}$$

Similarly,

$$\left\|\boldsymbol{m} - \frac{\|\boldsymbol{m}\|_2}{\|\boldsymbol{m}_0\|_2}\boldsymbol{m}_0\right\|_2 \le \left(\left|\|\boldsymbol{m}\|_2 - \|\boldsymbol{m}\|_2\right| + \left(\|\boldsymbol{m}\|_2 + \left|\|\boldsymbol{m}\|_2 - \|\boldsymbol{m}\|_2\right|\right)\bar{\theta}_0^{(2)}\right) \le 2\sqrt{\epsilon}\|\boldsymbol{m}\|_2.$$

Hence,

$$\left\|(\boldsymbol{h},\boldsymbol{m}) - \left(c\boldsymbol{h}_0, \frac{1}{c}\boldsymbol{m}_0\right)\right\|_2 \le 437\frac{ds\sqrt{\epsilon}}{\alpha}\left\|\left(c\boldsymbol{h}_0, \frac{1}{c}\boldsymbol{m}_0\right)\right\|_2.$$

Case 2: $\bar{\boldsymbol{\theta}}_{\mathbf{0}}^{(\mathbf{1})} \approx \boldsymbol{\pi}$ and $\bar{\boldsymbol{\theta}}_{\mathbf{0}}^{(\mathbf{2})} \approx \mathbf{0}$ Assume $\bar{\theta}_0^{(1)} = \pi + O_1(\delta)$ where $\delta = 12\pi^2d^3\sqrt{\epsilon}/\alpha$. By (42) and (43), we have $\xi^{(1)} = O_1(\delta/\pi)$, and we have $\zeta^{(1)} = \rho_d^{(1)} + O_1(3d^3\delta)$ if $38d^5\sqrt{\epsilon}/\alpha \le 1$. Also, assume

$\bar{\theta}_0^{(2)} = O_1(2\sqrt{\epsilon})$. As $\bar{\theta}_i^{(2)} \le \bar{\theta}_0^{(2)} \le 2\sqrt{\epsilon}$ for all i, we have $\xi^{(2)} \ge \left(1 - \frac{2\sqrt{\epsilon}}{\pi}\right)^s = 1 + O_1(\frac{4s\sqrt{\epsilon}}{\pi})$ provided $2s\sqrt{\epsilon} \le 1/2$. By (44), we also have $\zeta^{(2)} = O_1(\frac{s}{\pi}2\sqrt{\epsilon}) = O_1(s\sqrt{\epsilon})$. By (33), we have

$$\left| r^{(1)} r^{(2)} \cos\bar{\theta}_0^{(2)} - \left(\xi^{(1)} \cos\bar{\theta}_0^{(1)} + \zeta^{(1)}\right)\left(\xi^{(2)} + \zeta^{(2)} \cos\bar{\theta}_0^{(2)}\right)\right| \le \frac{\epsilon M}{\alpha}$$

where we used $\cos\bar{\theta}_{a^{(k)}}^{(k)} = \xi^{(k)} \cos\bar{\theta}_0^{(k)} + \zeta^{(k)}$. As $\cos\bar{\theta}_0^{(1)} = -1 + O((\bar{\theta}_0^{(1)} - \pi)^2/2) = -1 + O_1(\delta^2/2)$ provided $\delta < 1$ and $\cos\bar{\theta}_0^{(2)} = 1 + O_1((\bar{\theta}_0^{(2)})^2/2) = 1 + O_1(2\epsilon)$,

$$\begin{aligned}
\xi^{(1)} \cos\bar{\theta}_0^{(1)} + \zeta^{(1)} =& \rho_d^{(1)} + O_1(\frac{\delta^3}{2\pi} + \frac{\delta}{\pi} + 3d^3\delta) = \rho_d^{(1)} + O_1(4\delta d^3),\\
\xi^{(2)} + \zeta^{(2)} \cos\bar{\theta}_0^{(2)} =& 1 + O_1(4s\sqrt{\epsilon} + 2s\epsilon\sqrt{\epsilon} + s\sqrt{\epsilon}) = 1 + O_1(7s\sqrt{\epsilon}).
\end{aligned}$$

Thus,

$$\begin{aligned}
r^{(1)} r^{(2)} =& \rho_d^{(1)} + O_1(12\epsilon + 6\epsilon/\alpha + 4\delta d^3 + 7s\sqrt{\epsilon} + 28d^3 s\delta\sqrt{\epsilon})\\
=& \rho_d^{(1)} + O_1(30d\sqrt{\epsilon}/\alpha + 4\delta d^3 + 28d^3 s\sqrt{\epsilon})\\
=& \rho_d^{(1)} + O_1(532d^6 s\sqrt{\epsilon}/\alpha).
\end{aligned}$$

where, in the second equality, we use $\delta < 1$. We now show $(\boldsymbol{h}, \boldsymbol{m})$ is close to $\left(-c\rho_d^{(1)}\boldsymbol{h}_0, \frac{1}{c}\boldsymbol{m}_0\right)$, where $c = \frac{\|\boldsymbol{m}_0\|_2}{\|\boldsymbol{m}\|_2}$. Consider

$$\begin{aligned}
&\left\|\boldsymbol{h} + \frac{\|\boldsymbol{m}_0\|_2}{\|\boldsymbol{m}\|_2}\rho_d^{(1)}\boldsymbol{h}_0\right\|_2\\
\le& \frac{1}{\|\boldsymbol{m}\|_2}\Big(\left|\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2 - \rho_d^{(1)}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2\right| + \Big(\rho_d^{(1)}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2 + \left|\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2\right.\\
&\left. - \rho_d^{(1)}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2\right|\Big)\bar{\theta}_0^{(1)}\Big)\\
\le& \frac{1}{\|\boldsymbol{m}\|_2}\left(532d^6 s\sqrt{\epsilon}\|\boldsymbol{m}_0\|_2\|\boldsymbol{h}_0\|_2/\alpha + \left(2\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2 + 532d^6 s\sqrt{\epsilon}\|\boldsymbol{m}_0\|_2\|\boldsymbol{h}_0\|_2/\alpha\right)119d^3\sqrt{\epsilon}/\alpha\right)\\
\le& \frac{1}{\|\boldsymbol{m}\|_2}\left(532d^6 s\sqrt{\epsilon}\|\boldsymbol{m}_0\|_2\|\boldsymbol{h}_0\|_2/\alpha + \left(2\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2 + 14s\|\boldsymbol{m}_0\|_2\|\boldsymbol{h}_0\|_2\right)119d^3\sqrt{\epsilon}/\alpha\right)\\
\le& 2436\pi\frac{d^6 s\sqrt{\epsilon}}{\alpha}\rho_d^{(1)}\frac{\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2}{\|\boldsymbol{m}\|_2}.
\end{aligned}$$

Similarly,

$$\left\|\boldsymbol{m} - \frac{\|\boldsymbol{m}\|_2}{\|\boldsymbol{m}_0\|_2}\boldsymbol{m}_0\right\|_2 \le \left(\left|\|\boldsymbol{m}\|_2 - \|\boldsymbol{m}\|_2\right| + \left(\|\boldsymbol{m}\|_2 + \left|\|\boldsymbol{m}\|_2 - \|\boldsymbol{m}\|_2\right|\right)\bar{\theta}_0^{(2)}\right) \le 2\sqrt{\epsilon}\|\boldsymbol{m}\|_2.$$

Hence,

$$\left\|(\boldsymbol{h}, \boldsymbol{m}) - \left(-c\rho_d^{(1)}\boldsymbol{h}_0, \frac{1}{c}\boldsymbol{m}_0\right)\right\|_2 \le 2436\pi\frac{d^6 s\sqrt{\epsilon}}{\alpha}\left\|\left(-c\rho_d^{(1)}\boldsymbol{h}_0, \frac{1}{c}\boldsymbol{m}_0\right)\right\|_2.$$

Case 3: $\bar{\theta}_0^{(1)} \approx 0$ and $\bar{\theta}_0^{(2)} \approx \pi$. The analysis is similar to case 2. Using (31), we get

$$r^{(1)} r^{(2)} = \rho_s^{(2)} + O_1(532ds^6\sqrt{\epsilon}/\alpha).$$

Again, similar to case 2, we can show $(\boldsymbol{h}, \boldsymbol{m})$ is close to $\left(c\rho_s^{(2)}\boldsymbol{h}_0, -\frac{1}{c}\boldsymbol{m}_0\right)$, where $c = \frac{\|\boldsymbol{h}\|_2}{\|\boldsymbol{h}_0\|_2}$. We get,

$$\left\|(\boldsymbol{h}, \boldsymbol{m}) - \left(c\rho_s^{(2)}\boldsymbol{h}_0, -\frac{1}{c}\boldsymbol{m}_0\right)\right\|_2 \le 2436\pi\frac{ds^6\sqrt{\epsilon}}{\alpha}\left\|\left(c\rho_s^{(2)}\boldsymbol{h}_0, -\frac{1}{c}\boldsymbol{m}_0\right)\right\|_2.$$

Case 4: $\bar{\theta}_0^{(1)} \approx \pi$ and $\bar{\theta}_0^{(2)} \approx \pi$. Assume $\bar{\theta}_0^{(k)} = \pi + O_1(\delta^{(k)})$ where $\delta^{(k)} = 12\pi^2 {a^{(k)}}^3\sqrt{\epsilon}/\alpha$. By (42) and (43), we have $\xi^{(k)} = O_1(\delta^{(k)}/\pi)$, and we have $\zeta^{(k)} = \rho_d^{(k)} + O_1(3{a^{(k)}}^3\delta^{(k)})$ if $\frac{{a^{(k)}}^2\delta}{\pi} \le 1$. By (31), we have

$$\left| r^{(1)}r^{(2)}\cos\bar{\theta}_0^{(1)} - \left(\xi^{(2)}\cos\bar{\theta}_0^{(2)} + \zeta^{(2)}\right)\left(\xi^{(1)} + \zeta^{(1)}\cos\bar{\theta}_0^{(1)}\right)\right| \le \frac{\epsilon M}{\alpha}$$

where we used $\cos\bar{\theta}_{a^{(k)}}^{(k)} = \xi^{(k)}\cos\bar{\theta}_0^{(k)} + \zeta^{(k)}$. As $\cos\bar{\theta}_0^{(k)} = -1 + O((\bar{\theta}_0^{(k)} - \pi)^2/2) = -1 + O_1((\delta^{(k)})^2/2)$,

$$\begin{aligned}
\xi^{(2)}\cos\bar{\theta}_0^{(2)} + \zeta^{(2)} =& \rho_s^{(2)} + O_1(\frac{(\delta^{(2)})^3}{2\pi} + \frac{\delta^{(2)}}{\pi} + 3s^3\delta^{(2)}) = \rho_s^{(2)} + O_1(4\delta^{(2)}s^3),\\
\xi^{(1)} + \zeta^{(1)}\cos\bar{\theta}_0^{(1)} =& -\rho_d^{(1)} + O_1(\frac{\delta^{(1)}}{\pi} + \frac{3}{2}d^3(\delta^{(1)})^3 + 3\delta^{(1)}d^3) = -\rho_d^{(1)} + O_1(5\delta^{(1)}d^3).
\end{aligned}$$

Thus,

$$\begin{aligned}
r^{(1)}r^{(2)} =& \rho_d^{(1)}\rho_s^{(2)} + O_1(6\epsilon/\alpha + 4(\delta^{(1)})^2 + 4\delta^{(2)}s^3 + 5\delta^{(1)}d^3 + 20\delta^{(1)}\delta^{(2)}d^3s^3)\\
=& \rho_d^{(1)}\rho_s^{(2)} + O_1(6\epsilon/\alpha + 4\delta^{(1)} + 4\delta^{(2)}s^3 + 5\delta^{(1)}d^3 + 20\delta^{(2)}d^3s^3)\\
=& \rho_d^{(1)}\rho_s^{(2)} + O_1(6\epsilon/\alpha + 3909d^6s^6\sqrt{\epsilon}/\alpha)\\
=& \rho_d^{(1)}\rho_s^{(2)} + O_1(3915d^6s^6\sqrt{\epsilon}/\alpha),
\end{aligned}$$

where, in the second equality, we used $\delta^{(1)} \le \frac{\pi}{d^2} < 1$. We now show $(\boldsymbol{h}, \boldsymbol{m})$ is close to $\left(-c\rho_d^{(1)}\rho_s^{(2)}\boldsymbol{h}_0, -\frac{1}{c}\boldsymbol{m}_0\right)$, where $c = \frac{\|\boldsymbol{m}_0\|_2}{\|\boldsymbol{m}\|_2}$. Consider

$$\begin{aligned}
&\left\|\boldsymbol{h} + \frac{\|\boldsymbol{m}_0\|_2}{\|\boldsymbol{m}\|_2}\rho_d^{(1)}\rho_s^{(2)}\boldsymbol{h}_0\right\|_2\\
\le& \frac{1}{\|\boldsymbol{m}\|_2}\Big(\left|\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2 - \rho_d^{(1)}\rho_s^{(2)}\|\boldsymbol{m}_0\|_2\|\boldsymbol{h}_0\|_2\right| + \Big(\rho_d^{(1)}\rho_s^{(2)}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2 +\\
&\quad\left|\|\boldsymbol{h}\|_2\|\boldsymbol{m}\|_2 - \rho_d^{(1)}\rho_s^{(2)}\|\boldsymbol{m}_0\|_2\|\boldsymbol{h}_0\|_2\right|\Big)\bar{\theta}_0^{(1)}\Big)\\
\le& \frac{1}{\|\boldsymbol{m}\|_2}\big(3915d^6s^6\sqrt{\epsilon}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2/\alpha + \big(4\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2\\
&\quad + 3915d^6s^6\sqrt{\epsilon}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2/\alpha\big)119d^3\sqrt{\epsilon}/\alpha\big)\\
\le& \frac{1}{\|\boldsymbol{m}\|_2}\left(3915d^6s^6\sqrt{\epsilon}\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2/\alpha + \left(4\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2 + 104ds^6\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2\right)119d^3\sqrt{\epsilon}/\alpha\right)\\
\le& 16767\pi^2\frac{d^6s^6\sqrt{\epsilon}}{\alpha}\rho_d^{(1)}\rho_s^{(2)}\frac{\|\boldsymbol{h}_0\|_2\|\boldsymbol{m}_0\|_2}{\|\boldsymbol{m}\|_2}.
\end{aligned}$$

Similarly,

$$\left\|\boldsymbol{m} + \frac{\|\boldsymbol{m}\|_2}{\|\boldsymbol{m}_0\|_2}\boldsymbol{m}_0\right\|_2 \le \left(|\|\boldsymbol{m}\|_2 - \|\boldsymbol{m}\|_2| + (\|\boldsymbol{m}\|_2 + |\|\boldsymbol{m}\|_2 - \|\boldsymbol{m}\|_2|)\,\bar{\theta}_0^{(2)}\right) \le 119s^3\sqrt{\epsilon}\|\boldsymbol{m}\|_2/\alpha.$$

Hence,

$$\left\|(\boldsymbol{h}, \boldsymbol{m}) - \left(-c\rho_d^{(1)}\rho_s^{(2)}\boldsymbol{h}_0, -\frac{1}{c}\boldsymbol{m}_0\right)\right\|_2 \le 16767\pi^2\frac{d^6s^6\sqrt{\epsilon}}{\alpha}\left\|\left(c\rho_d^{(1)}\rho_s^{(2)}\boldsymbol{h}_0, \frac{1}{c}\boldsymbol{m}_0\right)\right\|_2.$$

□

5.4 Proof of WDC condition

We first state a lemma that shows that the weight $\boldsymbol{W} \in \mathbb{R}^{\ell\times n}$ of a layer of a neural network layer with i.i.d. $\mathcal{N}(0, 1/\ell)$ entries satisfies the WDC with constant ϵ and 1, and we refer the readers to Hand and Voroninski [2017] for a proof of the lemma.

Lemma 4. *Fix $0 < \epsilon < 1$. Let $\boldsymbol{W} \in \mathbb{R}^{\ell \times n}$ have i.i.d. $\mathcal{N}(0, 1/\ell)$ entries. If $\ell > cn \log n$, then with probability at least $1 - 8\ell e^{-\gamma n}$, $\boldsymbol{W}$ satisfies the WDC with constant ϵ and 1. Here c, γ^{-1} are constants that depend only polynomially on ϵ^{-1}.*

We now state a lemma similar to Lemma 4 which applies to truncated random variable. The proof follows the proof of lemma 4 in Hand and Voroninski [2017].

Lemma 5. *Fix $0 < \epsilon < 1$. Let $\boldsymbol{W} \in \mathbb{R}^{\ell \times n}$ where ith row of $\boldsymbol{W}$ satisfy $\boldsymbol{w}_i^\intercal = \boldsymbol{w}^\intercal \cdot \mathbf{1}_{\|\boldsymbol{w}\|_2 \le 3\sqrt{n/\ell}}$ and $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\ell}\boldsymbol{I}_n)$. If $\ell > cn \log n$, then with probability at least $1 - 8ne^{-\gamma n}$, $\boldsymbol{W}$ satisfies the WDC with constant ϵ and α. Here c, γ^{-1} are constants that depend only polynomially on ϵ^{-1} and*

$$\alpha = \frac{\Gamma\left(\frac{n+2}{2}\right) - \Gamma\left(\frac{n+2}{2}, \frac{9n}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)}, \tag{47}$$

where Γ is the Gamma function.

The WDC condition with constant ϵ and α can be written as

$$\|\boldsymbol{W}_{+,\boldsymbol{x}}^\intercal \boldsymbol{W}_{+,\boldsymbol{y}} - \alpha \boldsymbol{Q}_{\boldsymbol{x},\boldsymbol{y}}\| \le \epsilon$$

for all nonzero $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. We note that

$$\boldsymbol{W}_{+,\boldsymbol{x}}^\intercal \boldsymbol{W}_{+,\boldsymbol{y}} = \sum_{i=1}^{\ell} \mathbf{1}_{\boldsymbol{w}_i^\intercal \boldsymbol{x}} \mathbf{1}_{\boldsymbol{w}_i^\intercal \boldsymbol{y}} \boldsymbol{w}_i \boldsymbol{w}_i^\intercal$$

and it is not continuous in $\boldsymbol{x}$ and $\boldsymbol{y}$. So, we consider an arbitrarily good continuous approximation of $\boldsymbol{W}_{+,\boldsymbol{x}}^\intercal \boldsymbol{W}_{+,\boldsymbol{y}}$. Let

$$t_{-\epsilon}(z) = \begin{cases} 0 & z \le -\epsilon, \\ 1 + \frac{z}{\epsilon} & -\epsilon \le z \le 0, \\ 1 & z \le 0, \end{cases} \quad \text{and} \quad t_{\epsilon}(z) = \begin{cases} 0 & z \le 0, \\ \frac{z}{\epsilon} & 0 \le z \le \epsilon, \\ 1 & z \ge \epsilon. \end{cases}$$

and define

$$H_{-\epsilon}(\boldsymbol{x}\boldsymbol{y}) := \sum_{i=1}^{\ell} t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{x}) t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{y}) \boldsymbol{w}_i \boldsymbol{w}_i^\intercal,$$

$$H_{\epsilon}(\boldsymbol{x}, \boldsymbol{y}) := \sum_{i=1}^{\ell} t_{\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{x}) t_{\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{y}) \boldsymbol{w}_i \boldsymbol{w}_i^\intercal.$$

The proof of Lemma 5 follows from the follow two lemmas. We first provide an upper bound on the singular values of $H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y})$.

Lemma 6. *Fix $0 < \epsilon < 1$. Let $\boldsymbol{W} \in \mathbb{R}^{\ell \times n}$ where ith row of $\boldsymbol{W}$ satisfy $\boldsymbol{w}_i^\intercal = \boldsymbol{w}^\intercal \cdot \mathbf{1}_{\|\boldsymbol{w}\|_2 \le 3\sqrt{n}}$ and $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$. If $\ell > cn \log n$, then with probability at least $1 - 4ne^{-\gamma n}$,*

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \ne (\mathbf{0}, \mathbf{0}), \quad H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y}) \preceq \alpha \ell \boldsymbol{Q}_{\boldsymbol{x},\boldsymbol{y}} + 3\ell\epsilon I_n.$$

Here, c and γ^{-1} are constants that depend only polynomially on ϵ^{-1} and α is

$$\alpha = \frac{\Gamma\left(\frac{n+2}{2}\right) - \Gamma\left(\frac{n+2}{2}, \frac{9n}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)}, \tag{48}$$

where Γ is the Gamma function.

Proof. First we bound $\mathbb{E}[H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y})]$ for fixed $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^{n-1}$. Noting that $t_{-\epsilon}(z) \le \mathbf{1}_{z \ge -\epsilon}(z) = \mathbf{1}_{z>0}(z) + \mathbf{1}_{-\epsilon \le z \le 0}(z)$, we have

$$\begin{aligned} &\mathbb{E}\left[H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y})\right] \\ \preceq &\mathbb{E}\left[\sum_{i=1}^{\ell} \mathbf{1}_{\boldsymbol{w}_i^\intercal \boldsymbol{x} \ge -\epsilon} \mathbf{1}_{\boldsymbol{w}_i^\intercal \boldsymbol{y} \ge -\epsilon} \boldsymbol{w}_i \boldsymbol{w}_i^\intercal\right] \end{aligned}$$

$$
\begin{aligned}
&= \ell \mathbb{E}\left[\mathbf{1}_{\boldsymbol{w}_i^\intercal \boldsymbol{x} \geq -\epsilon} \mathbf{1}_{\boldsymbol{w}_i^\intercal \boldsymbol{y} \geq -\epsilon} \boldsymbol{w}_i \boldsymbol{w}_i^\intercal\right] \\
&= \ell \mathbb{E}\Big[\left(\mathbf{1}_{\boldsymbol{w}_i^\intercal \boldsymbol{x} \geq 0} \mathbf{1}_{\boldsymbol{w}_i^\intercal \boldsymbol{y} \geq 0} \boldsymbol{w}_i \boldsymbol{w}_i^\intercal\right)\Big] + 2\ell \mathbb{E}\left[\left(\mathbf{1}_{-\epsilon \leq \boldsymbol{w}_i^\intercal \boldsymbol{x} \leq 0} \boldsymbol{w}_i \boldsymbol{w}_i^\intercal\right)\right].
\end{aligned}
$$

We first note that $\mathbb{E}\left[\mathbf{1}_{\boldsymbol{w}_i^\intercal \boldsymbol{x} \geq 0} \mathbf{1}_{\boldsymbol{w}_i^\intercal \boldsymbol{y} \geq 0} \boldsymbol{b}_i \boldsymbol{b}_i^\intercal\right] = \alpha \boldsymbol{Q}_{\boldsymbol{x},\boldsymbol{y}}$ where α satisfies $0.97 < \alpha < 1$. Also, we have $\mathbb{E}\left[\mathbf{1}_{-\epsilon \leq \boldsymbol{w}_i^\intercal \boldsymbol{x} \leq 0} \boldsymbol{w}_i \boldsymbol{w}_i^\intercal\right] \preceq \frac{\epsilon\alpha}{2} \boldsymbol{I}_n$. Thus,

$$
\begin{aligned}
\mathbb{E}\left[H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y})\right] &\preceq \alpha\ell \cdot \boldsymbol{m}^\intercal \boldsymbol{Q}_{\boldsymbol{x},\boldsymbol{y}} \boldsymbol{y} + \epsilon\alpha\ell \boldsymbol{I}_n \\
&\preceq \alpha\ell \cdot \boldsymbol{m}^\intercal \boldsymbol{Q}_{\boldsymbol{x},\boldsymbol{y}} + \epsilon\ell \boldsymbol{I}_n
\end{aligned} \tag{49}
$$

Second, we show concentration of $H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y})$ for fixed $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^{n-1}$. Let

$$
\boldsymbol{\xi}_i = \sqrt{t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{x}) t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{y})}\, \boldsymbol{w}_i.
$$

We have

$$
\begin{aligned}
&H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y}) - \mathbb{E}\left[H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y})\right] \\
=&\sum_{i=1}^{\ell} \Big(t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{x}) t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{y}) \boldsymbol{w}_i \boldsymbol{w}_i^\intercal - \mathbb{E}\left[t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{x}) t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{y}) \boldsymbol{w}_i \boldsymbol{w}_i^\intercal\right]\Big) \\
=&\sum_{i=1}^{\ell} \left(\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\intercal - \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\intercal\right]\right).
\end{aligned}
$$

Note that $\boldsymbol{\xi}_i$ is sub-Gaussian for all i and that the sub-Gaussian norm of $\boldsymbol{\xi}_i$ is bounded from above by an absolute constant which we call K. By first part of Remark 5.40 in Vershynin [2012], there exists a c_K and γ_K such that for all $t \geq 0$, with probability at least $1 - 2e^{-\gamma_K t^2}$,

$$
\|H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y}) - \mathbb{E}\left[H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y})\right]\| \leq \max(\delta, \delta^2)\ell, \quad \text{where } \delta = c_K \sqrt{\frac{n}{\ell}} + \frac{t}{\sqrt{\ell}}.
$$

If $\ell > (2c_K/\epsilon)^2 n, t = \epsilon\sqrt{\ell}/2$, and $\epsilon < 1$, we have

$$
\|H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y}) - \mathbb{E}\left[H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y})\right]\| \leq \epsilon\ell \tag{50}
$$

with probability at least $1 - 2e^{-\gamma_K \frac{\epsilon^2 \ell}{4}}$.

Third, we bound the Lipschitz constant of $H_{-\epsilon}$. For $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \in \mathbb{R}^n$ we have

$$
\begin{aligned}
&H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y}) - H_{-\epsilon}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \\
=&\sum_{i=1}^{\ell} \Big[t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{x}) t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{y}) - t_{-\epsilon}(\boldsymbol{w}_i^\intercal \tilde{\boldsymbol{x}}) t_{-\epsilon}(\boldsymbol{w}_i^\intercal \tilde{\boldsymbol{y}})\Big] \boldsymbol{w}_i \boldsymbol{w}_i^\intercal \\
=&\sum_{i=1}^{\ell} \Big[t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{x}) \left(t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{y}) - t_{-\epsilon}(\boldsymbol{w}_i^\intercal \tilde{\boldsymbol{y}})\right) \\
&\quad + t_{-\epsilon}(\boldsymbol{w}_i^\intercal \tilde{\boldsymbol{y}}) \left(t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{x}) - t_{-\epsilon}(\boldsymbol{w}_i^\intercal \tilde{\boldsymbol{x}})\right)\Big] \boldsymbol{w}_i \boldsymbol{w}_i^\intercal \\
=&\boldsymbol{W}^\intercal \Big[\operatorname{diag}\left(t_{-\epsilon}(\boldsymbol{W}\boldsymbol{x})\right) \operatorname{diag}((\boldsymbol{W}\boldsymbol{y})_+ - (\boldsymbol{W}\tilde{\boldsymbol{y}})_+) \\
&\quad + \operatorname{diag}\left(t_{-\epsilon}(\boldsymbol{W}\tilde{\boldsymbol{y}})\right) \operatorname{diag}\left((\boldsymbol{W}\boldsymbol{x})_+ - (\boldsymbol{W}\tilde{\boldsymbol{x}})_+\right)\Big] \boldsymbol{W}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\|H_{-\epsilon}(\boldsymbol{x}, \boldsymbol{y}) - H_{-\epsilon}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})\| \\
\leq&\|\boldsymbol{W}\|^2 \Big[\|t_{-\epsilon}(\boldsymbol{W}\boldsymbol{y}) - t_{-\epsilon}(\boldsymbol{W}\tilde{\boldsymbol{y}})\|_\infty + \|t_{-\epsilon}(\boldsymbol{W}\boldsymbol{x}) - t_{-\epsilon}(\boldsymbol{W}\tilde{\boldsymbol{x}})\|_\infty\Big] \\
\leq&\|\boldsymbol{W}\|^2 \left[\max_{i \in [\ell]} |t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{y}) - t_{-\epsilon}(\boldsymbol{w}_i^\intercal \tilde{\boldsymbol{y}})| + \max_{i \in [\ell]} |t_{-\epsilon}(\boldsymbol{w}_i^\intercal \boldsymbol{x}) - t_{-\epsilon}(\boldsymbol{w}_i^\intercal \tilde{\boldsymbol{x}})|\right]
\end{aligned}
$$

$$
\begin{aligned}
\leq & \|\boldsymbol{W}\|^2 \left[\max_{i\in[\ell]} \frac{1}{\epsilon} \left|\boldsymbol{w}_i^\intercal (\boldsymbol{y}-\tilde{\boldsymbol{y}})\right| + \max_{i\in[\ell]} \frac{1}{\epsilon} \left|\boldsymbol{w}_i^\intercal (\boldsymbol{x}-\tilde{\boldsymbol{x}})\right| \right] \\
\leq & \|\boldsymbol{W}\|^2 \left[\frac{1}{\epsilon} \max_{i\in[\ell]} \|\boldsymbol{w}_i\|_2 \|\boldsymbol{y}-\tilde{\boldsymbol{y}}\| + \frac{1}{\epsilon} \max_{i\in[\ell]} \|\boldsymbol{w}_i\|_2 \|\boldsymbol{x}-\tilde{\boldsymbol{x}}\| \right] \\
\leq & \|\boldsymbol{W}\|^2 \left[\frac{9}{\epsilon}\sqrt{n} \|\boldsymbol{x}-\tilde{\boldsymbol{x}}\| + \frac{9}{\epsilon}\sqrt{n} \left\|\boldsymbol{h}-\tilde{\boldsymbol{h}}\right\| \right]
\end{aligned}
$$

where the first inequality follows because $|t_{-\epsilon}(z)| \leq 1$ for all z, and the third inequality follows because $t_{-\epsilon}(z)$ is $1/\epsilon$-Lipschitz. Let E_1 be the event that $\|\boldsymbol{W}\| \leq 3\sqrt{\ell}$. By Corollary 5.35 in Vershynin [2012], for $\boldsymbol{A} \in \mathbb{R}^{\ell\times n}$ with rows of $\boldsymbol{A}$ following $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, we have $\mathbb{P}(\|\boldsymbol{A}\| \leq 3\sqrt{\ell}) \geq 1 - 2e^{-\ell/2}$, if $\ell \geq n$. As rows of $\boldsymbol{W}$ are truncated, we have $\mathbb{P}(E_1) \geq 1 - 2e^{-\ell/2}$, if $\ell \geq n$ as well. On E_1, we have

$$
\begin{aligned}
& \|H_{-\epsilon}(\boldsymbol{x},\boldsymbol{y}) - H_{-\epsilon}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\| \\
\leq & \frac{27\ell\sqrt{n}}{\epsilon} \left[\|\boldsymbol{x}-\tilde{\boldsymbol{y}}\| + \|\boldsymbol{y}-\tilde{\boldsymbol{y}}\|\right]
\end{aligned} \tag{51}
$$

for all $\boldsymbol{x}, \boldsymbol{y}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \in \mathcal{S}^{n-1}$.

Finally, we complete the proof by a covering argument. Let $\mathcal{N}_\delta$ be a δ-net on $\mathcal{S}^{n-1}$ such that $|\mathcal{N}_\delta| \leq (3/\delta)^n$. Take $\delta = \frac{\epsilon^2}{54\sqrt{n}}$. Combining (49) and (54), we have

$$
\begin{aligned}
\forall (\boldsymbol{x},\boldsymbol{y}), \in \mathcal{N}_\delta, \quad H_{-\epsilon}(\boldsymbol{x},\boldsymbol{y}) \preceq & \mathbb{E} H_{-\epsilon}(\boldsymbol{x},\boldsymbol{y}) + \ell\epsilon I_n \\
\preceq & \alpha\ell \boldsymbol{Q}_{\boldsymbol{x},\boldsymbol{y}} + 2\ell\epsilon I_n.
\end{aligned}
$$

with probability at least

$$
1 - 2|\mathcal{N}_\delta| e^{-\gamma_K \epsilon^2 \ell/4} \geq 1 - 2\left(\frac{3}{\delta}\right)^n e^{-\gamma_K\epsilon^2\ell/4} \geq 1 - 2e^{-\gamma_K\epsilon^2\ell/4 + n\log(3\cdot 54\sqrt{n}/\epsilon^2)}.
$$

If $\ell \geq \tilde{c} n \log(n)$ for some $\tilde{c} = \Omega(\epsilon^2 \log\epsilon)$, then this probability is at least $1 - 2e^{-\tilde{\gamma}\ell}$ for some $\tilde{\gamma} = O(\epsilon^2)$. For $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^{n-1}$, let $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \in \mathcal{N}_\delta$ be such that $\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2 \leq \delta$, and $\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|_2 \leq \delta$. By (51), we have that

$$
\begin{aligned}
\forall \boldsymbol{x}, \boldsymbol{y} \neq \boldsymbol{0}, \quad & H_{-\epsilon}(\boldsymbol{x},\boldsymbol{y}) \\
\preceq & H_{-\epsilon}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) + \frac{27\ell\sqrt{n}}{\epsilon} 2\delta \boldsymbol{I}_n \\
\preceq & \alpha\ell \boldsymbol{Q}_{\boldsymbol{x},\boldsymbol{y}} + 3\ell\epsilon \boldsymbol{I}_n.
\end{aligned}
$$

In conclusion, the result of this lemma holds if $\ell > (2c_K/\epsilon)^2 n$ and $\ell \geq \tilde{c}(n)\log n$, with probability at least $1 - 2e^{-\gamma_K\epsilon^2\ell/4} - 2e^{-\ell/2} - 2e^{-\tilde{\gamma}\ell} > 1 - 6e^{-\gamma\ell}$ for some $\gamma = O(\epsilon^2)$ and $\tilde{c} = \Omega(\epsilon^2\log\epsilon)$. □

Next, we now provide an upper bound on the singular values of $G_\epsilon(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})$.

Lemma 7. *Fix $0 < \epsilon < 1$. Let $\boldsymbol{W} \in \mathbb{R}^{\ell\times n}$ where ith row of $\boldsymbol{W}$ satisfy $\boldsymbol{w}_i^\intercal = \boldsymbol{w}^\intercal \cdot \boldsymbol{1}_{\|\boldsymbol{w}\|_2 \leq 3\sqrt{n}}$ and $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. If $\ell > cn\log n$, then with probability at least $1 - 4ne^{-\gamma n}$,*

$$
\forall (\boldsymbol{x},\boldsymbol{y}) \neq (\boldsymbol{0},\boldsymbol{0}), \quad H_\epsilon(\boldsymbol{x},\boldsymbol{y}) \succeq \alpha\ell \boldsymbol{Q}_{\boldsymbol{x},\boldsymbol{y}} - 3\ell\epsilon I_n.
$$

Here, c and γ^{-1} are constants that depend only polynomially on ϵ^{-1} and α is

$$
\alpha = \frac{\Gamma\left(\frac{n+2}{2}\right) - \Gamma\left(\frac{n+2}{2}, \frac{9n}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)}, \tag{52}
$$

where Γ is the Gamma function.

Proof. First we bound $\mathbb{E}[H_{-\epsilon}(\boldsymbol{x},\boldsymbol{y})]$ for fixed $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^{n-1}$. Noting that $t_\epsilon(z) \geq \boldsymbol{1}_{z>0}(z) - \boldsymbol{1}_{-\epsilon\leq z\leq 0}(z)$ for all z, we have

$$
\mathbb{E}\left[H_\epsilon(\boldsymbol{x},\boldsymbol{y})\right]
$$

$$\succeq \ell \mathbb{E}\Big[\Big(\mathbf{1}_{\boldsymbol{w}_i^\mathsf{T}\boldsymbol{x}\geq 0}\mathbf{1}_{\boldsymbol{w}_i^\mathsf{T}\boldsymbol{y}\geq 0}\boldsymbol{w}_i\boldsymbol{w}_i^\mathsf{T} \Big) \Big] - 2\ell\mathbb{E}\left[\Big(\mathbf{1}_{-\epsilon\leq \boldsymbol{w}_i^\mathsf{T}\boldsymbol{x}\leq 0}\boldsymbol{w}_i\boldsymbol{w}_i^\mathsf{T} \Big) \right].$$

We first note that $\mathbb{E}\left[\mathbf{1}_{\boldsymbol{w}_i^\mathsf{T}\boldsymbol{x}\geq 0}\mathbf{1}_{\boldsymbol{w}_i^\mathsf{T}\boldsymbol{y}\geq 0}\boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T} \right] = \alpha \boldsymbol{Q}_{\boldsymbol{x},\boldsymbol{y}}$ where α satisfies $0.97 < \alpha < 1$. Also, we have $\mathbb{E}\left[\mathbf{1}_{-\epsilon\leq \boldsymbol{w}_i^\mathsf{T}\boldsymbol{x}\leq 0}\boldsymbol{w}_i\boldsymbol{w}_i^\mathsf{T} \right] \preceq \frac{\epsilon\alpha}{2}\boldsymbol{I}_n$. Thus,

$$\begin{aligned}\mathbb{E}\left[H_{-\epsilon}(\boldsymbol{x},\boldsymbol{y})\right] &\succeq \alpha\ell \cdot \boldsymbol{m}^\mathsf{T}\boldsymbol{Q}_{\boldsymbol{x},\boldsymbol{y}}\boldsymbol{y} - \epsilon\alpha\ell\boldsymbol{I}_n \\ &\succeq \alpha\ell \cdot \boldsymbol{m}^\mathsf{T}\boldsymbol{Q}_{\boldsymbol{x},\boldsymbol{y}} - \epsilon\ell\boldsymbol{I}_n \end{aligned} \tag{53}$$

Second, the same argument as in Lemma 6 provides that for fixed $\boldsymbol{x},\boldsymbol{y}\in\mathcal{S}^{n-1}$, if $\ell > (2c_K/\epsilon)^2 n$, then we have with probability at least $1-2e^{-\gamma_K\frac{\epsilon^2\ell}{4}}$,

$$\|H_{-\epsilon}(\boldsymbol{x},\boldsymbol{y}) - \mathbb{E}\left[H_{-\epsilon}(\boldsymbol{x},\boldsymbol{y})\right]\| \leq \epsilon\ell \tag{54}$$

Third, same argument as in Lemma 6 provides on the event E_1, we have

$$\|H_{-\epsilon}(\boldsymbol{x},\boldsymbol{y}) - H_{-\epsilon}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\| \leq \frac{27\ell\sqrt{n}}{\epsilon}\left[\|\boldsymbol{x}-\tilde{\boldsymbol{y}}\| + \|\boldsymbol{y}-\tilde{\boldsymbol{y}}\|\right]$$

for all $\boldsymbol{x},\boldsymbol{y},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}\in\mathcal{S}^{n-1}$.

Finally, we complete the proof by an identical covering argument as in Lemma 6. We have if $\ell \geq c_0 n\log n$ then with probability at least $1-6e^{-\gamma\ell}$,

$$\forall \boldsymbol{x},\boldsymbol{y}\neq\boldsymbol{0},\quad H_\epsilon(\boldsymbol{x},\boldsymbol{y}) \succeq \alpha\ell\boldsymbol{Q}_{\boldsymbol{x},\boldsymbol{y}} - 3\ell\epsilon\boldsymbol{I}_n.$$

□

5.5 Proof of joint-WDC condition

We now state a result that states random gaussian matrices with truncated rows satisfy joint-WDC.

Lemma 8. *Fix $0<\epsilon<1$. Let $\boldsymbol{B}\in\mathbb{R}^{\ell\times n}$ where ith row of $\boldsymbol{B}$ satisfy $\boldsymbol{b}_i^\mathsf{T} = \boldsymbol{b}^\mathsf{T}\cdot\mathbf{1}_{\|\boldsymbol{b}\|_2\leq 3\sqrt{n/\ell}}$ and $\boldsymbol{b}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_n/\ell)$. Similarly, let $\boldsymbol{C}\in\mathbb{R}^{\ell\times p}$ where ith row of $\boldsymbol{C}$ satisfy $\boldsymbol{c}_i^\mathsf{T} = \boldsymbol{c}^\mathsf{T}\cdot\mathbf{1}_{\|\boldsymbol{c}\|_2\leq 3\sqrt{p/\ell}}$ and $\boldsymbol{c}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_p/\ell)$. If $\ell > c((n\log n)^2 + (p\log p)^2)$, then with probability at least $1-8e^{-\gamma\ell/(n\log n)}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ satisfy joint-WDC with constants ϵ and $\alpha=\alpha_1\cdot\alpha_2$. Here, c and γ^{-1} are constants that depend only polynomially on ϵ^{-1} and*

$$\alpha_1 = \frac{\Gamma\left(\frac{n+2}{2}\right) - \Gamma\left(\frac{n+2}{2},\frac{9n}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)} \text{ and } \alpha_2 = \frac{\Gamma\left(\frac{p+2}{2}\right) - \Gamma\left(\frac{p+2}{2},\frac{9p}{2}\right)}{\Gamma\left(\frac{p+2}{2}\right)}, \tag{55}$$

where Γ is the Gamma function.

The proof of Lemma 8 follows directly from Lemmas 9 and 10. Using Corollary 1, we provide a concentration result of $\boldsymbol{B}_{+,\boldsymbol{h}}^\mathsf{T}\mathrm{diag}(\boldsymbol{C}_{+,\boldsymbol{m}}\boldsymbol{m})\mathrm{diag}(\boldsymbol{C}_{+,\boldsymbol{y}}\boldsymbol{y})\boldsymbol{B}_{+,\boldsymbol{x}}$, which is part of the joint-WDC condition. We note that

$$\boldsymbol{B}_{+,\boldsymbol{h}}^\mathsf{T}\mathrm{diag}(\boldsymbol{C}_{+,\boldsymbol{m}}\boldsymbol{m})\mathrm{diag}(\boldsymbol{C}_{+,\boldsymbol{y}}\boldsymbol{y})\boldsymbol{B}_{+,\boldsymbol{x}} = \sum_{i=1}^{\ell}\mathbf{1}_{\boldsymbol{b}_i^\mathsf{T}\boldsymbol{h}>0}\mathbf{1}_{\boldsymbol{b}_i^\mathsf{T}\boldsymbol{x}>0}(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{m})_+(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{y})_+\boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T}$$

and it is not continuous in $\boldsymbol{h}$ and $\boldsymbol{x}$. So, we consider an arbitrarily good continuous approximation of $\boldsymbol{B}_{+,\boldsymbol{h}}^\mathsf{T}\mathrm{diag}(\boldsymbol{C}_{+,\boldsymbol{m}}\boldsymbol{m})\mathrm{diag}(\boldsymbol{C}_{+,\boldsymbol{y}}\boldsymbol{y})\boldsymbol{B}_{+,\boldsymbol{x}}$. Let

$$t_{-\epsilon}(z) = \begin{cases} 0 & z\leq -\epsilon, \\ 1+\frac{z}{\epsilon} & -\epsilon\leq z\leq 0, \\ 1 & z\leq 0, \end{cases} \quad \text{and} \quad t_\epsilon(z) = \begin{cases} 0 & z\leq 0, \\ \frac{z}{\epsilon} & 0\leq z\leq\epsilon, \\ 1 & z\geq\epsilon. \end{cases}$$

and define

$$G_{-\epsilon}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{m},\boldsymbol{y}) := \sum_{i=1}^{\ell} t_{-\epsilon}(\boldsymbol{b}_i^\mathsf{T}\boldsymbol{h})t_{-\epsilon}(\boldsymbol{b}_i^\mathsf{T}\boldsymbol{x})(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{m})_+(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{y})_+\boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T},$$

$$G_{\epsilon}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{m},\boldsymbol{y}) := \sum_{i=1}^{\ell} t_{\epsilon}(\boldsymbol{b}_i^\mathsf{T}\boldsymbol{h})t_{\epsilon}(\boldsymbol{b}_i^\mathsf{T}\boldsymbol{x})(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{m})_+(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{y})_+\boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T}.$$

We now provide an upper bound on the singular values of $G_{-\epsilon}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{m},\boldsymbol{y})$.

Lemma 9. *Fix $0 < \epsilon < 1$. Let $\boldsymbol{B} \in \mathbb{R}^{\ell \times n}$ where ith row of $\boldsymbol{B}$ satisfy $\boldsymbol{b}_i^\intercal = \boldsymbol{b}^\intercal \cdot \mathbf{1}_{\|\boldsymbol{b}\|_2 \le 3\sqrt{n}}$ and $\boldsymbol{b} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$. Similarly, let $\boldsymbol{C} \in \mathbb{R}^{\ell \times p}$ where ith row of $\boldsymbol{C}$ satisfy $\boldsymbol{c}_i^\intercal = \boldsymbol{c}^\intercal \cdot \mathbf{1}_{\|\boldsymbol{c}\|_2 \le 3\sqrt{p}}$ and $\boldsymbol{c} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$. If $\ell > c((n\log n)^2 + (p \log p)^2)$, then with probability at least $1 - 4e^{-\gamma \ell/(n \log n)}$,*

$$\begin{aligned}\forall (\boldsymbol{h}, \boldsymbol{x}) \neq (\mathbf{0}, \mathbf{0}) \text{ and } \boldsymbol{m}, \boldsymbol{y} \in \mathcal{S}^{p-1},\\ G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y}) \preceq \alpha_1 \alpha_2 \ell \boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}} \boldsymbol{m}^\intercal \boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}} \boldsymbol{y} + 4\ell\epsilon I_n.\end{aligned}$$

Here, c and γ^{-1} are constants that depend only polynomially on ϵ^{-1} and α_1 and α_2 is as in (55)*.*

Proof. First we bound $\mathbb{E}[G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})]$ for fixed $\boldsymbol{h}, \boldsymbol{x} \in \mathcal{S}^{n-1}$ and $\boldsymbol{m}, \boldsymbol{y} \in \mathcal{S}^{p-1}$. Noting that $t_{-\epsilon}(z) \le \mathbf{1}_{z \ge -\epsilon}(z) = \mathbf{1}_{z>0}(z) + \mathbf{1}_{-\epsilon \le z \le 0}(z)$, we have

$$\begin{aligned}&\mathbb{E}\left[G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})\right]\\ \preceq&\mathbb{E}\left[\sum_{i=1}^{\ell} \mathbf{1}_{\boldsymbol{b}_i^\intercal \boldsymbol{h} \ge -\epsilon} \mathbf{1}_{\boldsymbol{b}_i^\intercal \boldsymbol{x} \ge -\epsilon} (\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ (\boldsymbol{c}_i^\intercal \boldsymbol{y})_+ \boldsymbol{b}_i \boldsymbol{b}_i^\intercal\right]\\ =&\ell \mathbb{E}\left[\mathbf{1}_{\boldsymbol{b}_i^\intercal \boldsymbol{h} \ge -\epsilon} \mathbf{1}_{\boldsymbol{b}_i^\intercal \boldsymbol{x} \ge -\epsilon} (\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ (\boldsymbol{c}_i^\intercal \boldsymbol{y})_+ \boldsymbol{b}_i \boldsymbol{b}_i^\intercal\right]\\ \preceq&\ell \mathbb{E}\Big[\Big(\mathbf{1}_{\boldsymbol{b}_i^\intercal \boldsymbol{h} \ge 0} \mathbf{1}_{\boldsymbol{b}_i^\intercal \boldsymbol{x} \ge 0} (\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ (\boldsymbol{c}_i^\intercal \boldsymbol{y})_+ + \Big(\mathbf{1}_{-\epsilon \le \boldsymbol{b}_i^\intercal \boldsymbol{h} \le 0} + \mathbf{1}_{-\epsilon \le \boldsymbol{b}_i^\intercal \boldsymbol{x} \le 0}\Big) (\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ (\boldsymbol{c}_i^\intercal \boldsymbol{y})_+\Big) \boldsymbol{b}_i \boldsymbol{b}_i^\intercal\Big]\\ =&\ell \mathbb{E}\Big[\Big(\mathbf{1}_{\boldsymbol{b}_i^\intercal \boldsymbol{h} \ge 0} \mathbf{1}_{\boldsymbol{b}_i^\intercal \boldsymbol{x} \ge 0} \boldsymbol{b}_i \boldsymbol{b}_i^\intercal\Big) (\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ (\boldsymbol{c}_i^\intercal \boldsymbol{y})_+\Big] + 2\ell \mathbb{E}\left[\Big(\mathbf{1}_{-\epsilon \le \boldsymbol{b}_i^\intercal \boldsymbol{h} \le 0} \boldsymbol{b}_i \boldsymbol{b}_i^\intercal\Big) (\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ (\boldsymbol{c}_i^\intercal \boldsymbol{y})_+\right].\end{aligned}$$

We first note that $\mathbb{E}\left[\mathbf{1}_{\boldsymbol{b}_i^\intercal \boldsymbol{h} \ge 0} \mathbf{1}_{\boldsymbol{b}_i^\intercal \boldsymbol{x} \ge 0} \boldsymbol{b}_i \boldsymbol{b}_i^\intercal\right] = \alpha_1 \boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}}$ and $\mathbb{E}\left[(\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ (\boldsymbol{c}_i^\intercal \boldsymbol{y})_+\right] = \alpha_2 \boldsymbol{m}^\intercal \boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}} \boldsymbol{y}$ where α_i satisfies $0.97 < \alpha_i < 1$. Also, we have $\left|\boldsymbol{m}^\intercal \boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}} \boldsymbol{y}\right| \le \frac{1}{2}$ and $\mathbb{E}\left[\mathbf{1}_{-\epsilon \le \boldsymbol{b}_i^\intercal \boldsymbol{h} \le 0} \boldsymbol{b}_i \boldsymbol{b}_i^\intercal\right] \preceq \frac{\epsilon \alpha_1}{2} \boldsymbol{I}_n$. Thus,

$$\begin{aligned}\mathbb{E}\left[G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})\right] &\preceq \alpha_1 \alpha_2 \ell \cdot \boldsymbol{m}^\intercal \boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}} \boldsymbol{y} \cdot \boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}} + 2\ell \cdot \mathbb{E}\left[\mathbf{1}_{-\epsilon \le \boldsymbol{b}_i^\intercal \boldsymbol{h} \le 0} \boldsymbol{b}_i \boldsymbol{b}_i^\intercal\right] \cdot \alpha_2 \boldsymbol{m}^\intercal \boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}} \boldsymbol{y}\\ &\preceq \alpha_1 \alpha_2 \ell \cdot \boldsymbol{m}^\intercal \boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}} \boldsymbol{y} \cdot \boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}} + \frac{\epsilon \alpha_1 \alpha_2 \ell}{2} \boldsymbol{I}_n\\ &\preceq \alpha_1 \alpha_2 \ell \cdot \boldsymbol{m}^\intercal \boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}} \boldsymbol{y} \cdot \boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}} + \frac{\epsilon \ell}{2} \boldsymbol{I}_n \qquad (56)\end{aligned}$$

Second, we show concentration of $G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})$ for fixed $\boldsymbol{h}, \boldsymbol{x} \in \mathcal{S}^{n-1}$ and $\boldsymbol{m}, \boldsymbol{y} \in \mathcal{S}^{p-1}$. Let $\boldsymbol{\xi}_i = \sqrt{t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{h}) t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{x}) (\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ (\boldsymbol{c}_i^\intercal \boldsymbol{y})_+} \boldsymbol{b}_i$. We have

$$\begin{aligned}&G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y}) - \mathbb{E}\left[G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})\right]\\ =&\sum_{i=1}^{\ell} \Big(t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{h}) t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{x}) (\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ (\boldsymbol{c}_i^\intercal \boldsymbol{y})_+ \boldsymbol{b}_i \boldsymbol{b}_i^\intercal - \mathbb{E}\left[t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{h}) t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{x}) (\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ (\boldsymbol{c}_i^\intercal \boldsymbol{y})_+ \boldsymbol{b}_i \boldsymbol{b}_i^\intercal\right]\Big)\\ =&\sum_{i=1}^{\ell} \left(\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\intercal - \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\intercal\right]\right).\end{aligned}$$

Note that $\boldsymbol{\xi}_i$ is sub-Gaussian for all i and that the sub-Gaussian norm of $\boldsymbol{\xi}_i$ is bounded from above by $K = \tilde{K}\sqrt{n}$, where $\tilde{K}$ is an absolute constant. By Corollary 1, there exists a $c = \bar{c}\sqrt{n}\log n$ and $\gamma = \frac{\bar{\gamma}}{n \log n}$ such that for all $t \ge 0$, with probability at least $1 - 2e^{-\gamma t^2}$,

$$\|G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y}) - \mathbb{E}\left[G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})\right]\| \le \max(\delta, \delta^2)\ell, \quad \text{where } \delta = c\sqrt{\frac{n}{\ell}} + \frac{t}{\sqrt{\ell}}.$$

Here, $\bar{c}$ and $\bar{\gamma}$ are absolute constants. If $\ell > (2\bar{c}/\epsilon)^2 n^2 (\log n)^2$, $t = \epsilon\sqrt{\ell}/2$, and $\epsilon < 1$, we have

$$\|G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y}) - \mathbb{E}\left[G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})\right]\| \le \epsilon\ell \qquad (57)$$

with probability at least $1 - 2e^{-\bar{\gamma}\frac{\epsilon^2}{4}\frac{\ell}{n \log n}}$.

Third, we bound the Lipschitz constant of $G_{-\epsilon}$. For $\tilde{\boldsymbol{h}}, \tilde{\boldsymbol{x}} \in \mathbb{R}^n$ and $\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{y}} \in \mathcal{S}^{p-1}$ we have

$$
\begin{aligned}
&G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y}) - G_{-\epsilon}(\tilde{\boldsymbol{h}}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{m}}, \tilde{\boldsymbol{y}}) \\
=&\sum_{i=1}^{\ell} \Big[t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{h}) t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{x}) (\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ (\boldsymbol{c}_i^\intercal \boldsymbol{y})_+ - t_{-\epsilon}(\boldsymbol{b}_i^\intercal \tilde{\boldsymbol{h}}) t_{-\epsilon}(\boldsymbol{b}_i^\intercal \tilde{\boldsymbol{x}}) (\boldsymbol{c}_i^\intercal \tilde{\boldsymbol{m}})_+ (\boldsymbol{c}_i^\intercal \tilde{\boldsymbol{y}})_+ \Big] \boldsymbol{b}_i \boldsymbol{b}_i^\intercal \\
=&\sum_{i=1}^{\ell} \Big[t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{h}) t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{x}) (\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ \left((\boldsymbol{c}_i^\intercal \boldsymbol{y})_+ - (\boldsymbol{c}_i^\intercal \tilde{\boldsymbol{y}})_+ \right) \\
&\quad + t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{h}) t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{x}) (\boldsymbol{c}_i^\intercal \tilde{\boldsymbol{y}})_+ \left((\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ - (\boldsymbol{c}_i^\intercal \tilde{\boldsymbol{m}})_+ \right) \\
&\quad + t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{h}) (\boldsymbol{c}_i^\intercal \tilde{\boldsymbol{m}})_+ (\boldsymbol{c}_i^\intercal \tilde{\boldsymbol{y}})_+ \left(t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{x}) - t_{-\epsilon}(\boldsymbol{b}_i^\intercal \tilde{\boldsymbol{x}}) \right) \\
&\quad + t_{-\epsilon}(\boldsymbol{b}_i^\intercal \tilde{\boldsymbol{x}}) (\boldsymbol{c}_i^\intercal \tilde{\boldsymbol{m}})_+ (\boldsymbol{c}_i^\intercal \tilde{\boldsymbol{y}})_+ \Big(t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{h}) - t_{-\epsilon}(\boldsymbol{b}_i^\intercal \tilde{\boldsymbol{h}}) \Big) \Big] \boldsymbol{b}_i \boldsymbol{b}_i^\intercal \\
=&\boldsymbol{B}^\intercal \Big[\operatorname{diag}\left(t_{-\epsilon}(\boldsymbol{B}\boldsymbol{h}) \odot t_{-\epsilon}(\boldsymbol{B}\boldsymbol{x}) \odot (\boldsymbol{C}\boldsymbol{m})_+ \right) \operatorname{diag}((\boldsymbol{C}\boldsymbol{y})_+ - (\boldsymbol{C}\tilde{\boldsymbol{y}})_+) \\
&\quad + \operatorname{diag}\left(t_{-\epsilon}(\boldsymbol{B}\boldsymbol{h}) \odot t_{-\epsilon}(\boldsymbol{B}\boldsymbol{x}) \odot (\boldsymbol{C}\tilde{\boldsymbol{y}})_+ \right) \operatorname{diag}\left((\boldsymbol{C}\boldsymbol{m})_+ - (\boldsymbol{C}\tilde{\boldsymbol{m}})_+ \right) \\
&\quad + \operatorname{diag}\big(t_{-\epsilon}(\boldsymbol{B}\boldsymbol{h}) \odot (\boldsymbol{C}\tilde{\boldsymbol{m}})_+ \odot (\boldsymbol{C}\tilde{\boldsymbol{y}})_+ \big) \operatorname{diag}\left(t_{-\epsilon}(\boldsymbol{B}\boldsymbol{x}) - t_{-\epsilon}(\boldsymbol{B}\tilde{\boldsymbol{x}}) \right) \\
&\quad + \operatorname{diag}\left(t_{-\epsilon}(\boldsymbol{B}\tilde{\boldsymbol{x}}) \odot (\boldsymbol{C}\tilde{\boldsymbol{m}})_+ \odot (\boldsymbol{C}\tilde{\boldsymbol{y}})_+ \right) \operatorname{diag}\Big(t_{-\epsilon}(\boldsymbol{B}\boldsymbol{h}) - t_{-\epsilon}(\boldsymbol{B}\tilde{\boldsymbol{h}}) \Big) \Big] \boldsymbol{B}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\|G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y}) - G_{-\epsilon}(\tilde{\boldsymbol{h}}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{m}}, \tilde{\boldsymbol{y}})\| \\
\le&\|\boldsymbol{B}\|^2 \Big[\|\boldsymbol{C}\boldsymbol{m}\|_\infty \|(\boldsymbol{C}\boldsymbol{y})_+ - (\boldsymbol{C}\tilde{\boldsymbol{y}})_+\|_\infty + \|\boldsymbol{C}\tilde{\boldsymbol{y}}\|_\infty \|(\boldsymbol{C}\boldsymbol{m})_+ - (\boldsymbol{C}\tilde{\boldsymbol{m}})_+\|_\infty \\
&\quad + \|\boldsymbol{C}\tilde{\boldsymbol{m}}\|_\infty \|\boldsymbol{C}\tilde{\boldsymbol{y}}\|_\infty \|t_{-\epsilon}(\boldsymbol{B}\boldsymbol{x}) - t_{-\epsilon}(\boldsymbol{B}\tilde{\boldsymbol{x}})\|_\infty + \|\boldsymbol{C}\tilde{\boldsymbol{m}}\|_\infty \|\boldsymbol{C}\tilde{\boldsymbol{y}}\|_\infty \|t_{-\epsilon}(\boldsymbol{B}\boldsymbol{h}) - t_{-\epsilon}(\boldsymbol{B}\tilde{\boldsymbol{h}})\|_\infty \Big] \\
\le&\|\boldsymbol{B}\|^2 \Bigg[\max_{i\in[\ell]} \|\boldsymbol{c}_i\|_2 \max_{i\in[\ell]} |(\boldsymbol{c}_i^\intercal \boldsymbol{y})_+ - (\boldsymbol{c}_i^\intercal \tilde{\boldsymbol{y}})_+| + \max_{i\in[\ell]} \|\boldsymbol{c}_i\|_2 \max_{i\in[\ell]} |(\boldsymbol{c}_i^\intercal \boldsymbol{m})_+ - (\boldsymbol{c}_i^\intercal \tilde{\boldsymbol{m}})_+| \\
&\quad + \left(\max_{i\in[\ell]} \|\boldsymbol{c}_i\|_2 \right)^2 \max_{i\in[\ell]} |t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{x}) - t_{-\epsilon}(\boldsymbol{b}_i^\intercal \tilde{\boldsymbol{x}})| + \left(\max_{i\in[\ell]} \|\boldsymbol{c}_i\|_2 \right)^2 \max_{i\in[\ell]} \left| t_{-\epsilon}(\boldsymbol{b}_i^\intercal \boldsymbol{h}) - t_{-\epsilon}(\boldsymbol{b}_i^\intercal \tilde{\boldsymbol{h}}) \right| \Bigg] \\
\le&\|\boldsymbol{B}\|^2 \Bigg[\max_{i\in[\ell]} \|\boldsymbol{c}_i\|_2 \max_{i\in[\ell]} |\boldsymbol{c}_i^\intercal (\boldsymbol{y} - \tilde{\boldsymbol{y}})| + \max_{i\in[\ell]} \|\boldsymbol{c}_i\|_2 \max_{i\in[\ell]} |\boldsymbol{c}_i^\intercal (\boldsymbol{m} - \tilde{\boldsymbol{m}})| \\
&\quad + \left(\max_{i\in[\ell]} \|\boldsymbol{c}_i\|_2 \right)^2 \max_{i\in[\ell]} \frac{1}{\epsilon} |\boldsymbol{b}_i^\intercal (\boldsymbol{x} - \tilde{\boldsymbol{x}})| + \left(\max_{i\in[\ell]} \|\boldsymbol{c}_i\|_2 \right)^2 \max_{i\in[\ell]} \frac{1}{\epsilon} \left| \boldsymbol{b}_i^\intercal (\boldsymbol{h} - \tilde{\boldsymbol{h}}) \right| \Bigg] \\
\le&\|\boldsymbol{B}\|^2 \Bigg[\left(\max_{i\in[\ell]} \|\boldsymbol{c}_i\|_2 \right)^2 \|\boldsymbol{y} - \tilde{\boldsymbol{y}}\| + \left(\max_{i\in[\ell]} \|\boldsymbol{c}_i\|_2 \right)^2 \|\boldsymbol{m} - \tilde{\boldsymbol{m}}\| \\
&\quad + \frac{1}{\epsilon} \left(\max_{i\in[\ell]} \|\boldsymbol{c}_i\|_2 \right)^2 \max_{i\in[\ell]} \|\boldsymbol{b}_i\|_2 \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\| + \frac{1}{\epsilon} \left(\max_{i\in[\ell]} \|\boldsymbol{c}_i\|_2 \right)^2 \max_{i\in[\ell]} \|\boldsymbol{b}_i\|_2 \left\| \boldsymbol{h} - \tilde{\boldsymbol{h}} \right\| \Bigg] \\
\le&\|\boldsymbol{B}\|^2 \Big[9p \|\boldsymbol{y} - \tilde{\boldsymbol{y}}\| + 9p \|\boldsymbol{m} - \tilde{\boldsymbol{m}}\| + \frac{27}{\epsilon} \sqrt{n} p \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\| + \frac{27}{\epsilon} \sqrt{n} p \left\| \boldsymbol{h} - \tilde{\boldsymbol{h}} \right\| \Big]
\end{aligned}
$$

where the first inequality follows because $|t_{-\epsilon}(z)| \le 1$ for all z, and the third inequality follows because $t_{-\epsilon}(z)$ is $1/\epsilon$-Lipschitz and $(z)_+$ is 1-Lipschitz. Let E_1 be the event that $\|\boldsymbol{B}\| \le 3\sqrt{\ell}$. By Corollary 5.35 in Vershynin [2012], for $\boldsymbol{A} \in \mathbb{R}^{\ell \times n}$ with rows of $\boldsymbol{A}$ following $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, we have $\mathbb{P}(\|\boldsymbol{A}\| \le 3\sqrt{\ell}) \ge 1 - 2e^{-\ell/2}$, if $\ell \ge n$. As rows of $\boldsymbol{B}$ are truncated, we have $\mathbb{P}(E_1) \ge 1 - 2e^{-\ell/2}$, if $\ell \ge n$ as well. On E_1, we have

$$
\begin{aligned}
&\|G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y}) - G_{-\epsilon}(\tilde{\boldsymbol{h}}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{m}}, \tilde{\boldsymbol{y}})\| \\
\le&\frac{729 \ell \sqrt{np}}{\epsilon} \left[\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\| + \|\boldsymbol{m} - \tilde{\boldsymbol{m}}\| + \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\| + \left\| \boldsymbol{h} - \tilde{\boldsymbol{h}} \right\| \right]
\end{aligned} \tag{58}
$$

for all $\tilde{\boldsymbol{h}}, \tilde{\boldsymbol{x}} \in \mathcal{S}^{n-1}$ and $\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{y}} \in \mathcal{S}^{p-1}$.

Finally, we complete the proof by a covering argument. Let $\mathcal{N}_\delta$ be a δ-net on $\mathcal{S}^{n-1} \times \mathcal{S}^{p-1}$ such that $|\mathcal{N}_\delta| \le (3/\delta)^{n+p}$. Take $\delta = \frac{\epsilon^2}{2916\sqrt{np}}$. Combining (56) and (57), we have

$$\begin{aligned}\forall (\boldsymbol{h}, \boldsymbol{m}), (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{N}_\delta, \quad G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y}) \preceq & \mathbb{E} G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y}) + \ell\epsilon I_n \\ \preceq & \alpha^2 \ell \boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}} \boldsymbol{m}^\mathsf{T} \boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}} \boldsymbol{y} + 3\ell\epsilon I_n.\end{aligned}$$

with probability at least

$$1 - 2|\mathcal{N}_\delta| e^{-\gamma_K \epsilon^2 \frac{\ell}{4n\log n}} \ge 1 - 2\left(\frac{3}{\delta}\right)^{n+p} e^{-\gamma_K \epsilon^2 \frac{\ell}{4n\log n}} \ge 1 - 2e^{-\gamma_K \epsilon^2 \frac{\ell}{4n\log n} + (n+p)\log(3\cdot 2916\sqrt{np}/\epsilon^2)}.$$

If $\ell \ge \tilde{c} n(n+p)\log(n)\log(np)$ for some $\tilde{c} = \Omega(\epsilon^2 \log \epsilon^{-1})$, then this probability is at least $1 - 2e^{-\tilde{\gamma}\ell/(n\log(n))}$ for some $\tilde{\gamma} = O(\epsilon^2)$. For $(\boldsymbol{h}, \boldsymbol{m}), (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{S}^{n-1} \times \mathcal{S}^{p-1}$, let $(\tilde{\boldsymbol{h}}, \tilde{\boldsymbol{m}}), (\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \in \mathcal{N}_\delta$ be such that $\|\boldsymbol{h} - \tilde{\boldsymbol{h}}\|_2 \le \delta$, $\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2 \le \delta$, $\|\boldsymbol{m} - \tilde{\boldsymbol{m}}\|_2 \le \delta$ and $\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|_2 \le \delta$. By (58), we have that

$$\begin{aligned}\forall (\boldsymbol{h}, \boldsymbol{x}) \neq (\boldsymbol{0}, \boldsymbol{0}) \text{ and } \boldsymbol{m}, \boldsymbol{y} \in \mathcal{S}^{p-1}, \quad & G_{-\epsilon}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y}) \\ & \preceq G_{-\epsilon}(\tilde{\boldsymbol{h}}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{m}}, \tilde{\boldsymbol{y}}) + \frac{729\ell\sqrt{np}}{\epsilon} 4\delta \boldsymbol{I}_n \\ & \preceq \alpha_1\alpha_2 \ell \boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}} \boldsymbol{m}^\mathsf{T} \boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}} \boldsymbol{y} + 4\ell\epsilon \boldsymbol{I}_n.\end{aligned}$$

In conclusion, the result of this lemma holds if $\ell > (2\bar{c}/\epsilon)^2 n^2 (\log n)^2$ and $\ell \ge \tilde{c} n(n+p)\log(n)\log(np)$, with probability at least $1 - 2e^{-\ell/2} - 2e^{-\tilde{\gamma}\ell/(n\log(n))} > 1 - 4e^{-\gamma\ell/(n\log(n))}$ for some $\gamma = O(\epsilon^2)$ and $\tilde{c} = \Omega(\epsilon^2 \log \epsilon)$. □

Next, we now provide an upper bound on the singular values of $G_\epsilon(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})$.

Lemma 10. *Fix $0 < \epsilon < 1$. Let $\boldsymbol{B} \in \mathbb{R}^{\ell \times n}$ where ith row of $\boldsymbol{B}$ satisfy $\boldsymbol{b}_i^\mathsf{T} = \boldsymbol{b}^\mathsf{T} \cdot \mathbf{1}_{\|\boldsymbol{b}\|_2 \le 3\sqrt{n}}$ and $\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Similarly, let $\boldsymbol{C} \in \mathbb{R}^{\ell \times p}$ where ith row of $\boldsymbol{C}$ satisfy $\boldsymbol{c}_i^\mathsf{T} = \boldsymbol{c}^\mathsf{T} \cdot \mathbf{1}_{\|\boldsymbol{c}\|_2 \le 3\sqrt{p}}$ and $\boldsymbol{c} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. If $\ell > c((n\log n)^2 + (p\log p)^2)$, then with probability at least $1 - 4e^{-\gamma\ell/(n\log n)}$,*

$$\begin{aligned}&\forall (\boldsymbol{h}, \boldsymbol{x}) \neq (\boldsymbol{0}, \boldsymbol{0}) \text{ and } \boldsymbol{m}, \boldsymbol{y} \in \mathcal{S}^{p-1}, \\ &\qquad G_\epsilon(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y}) \succeq \alpha_1\alpha_2 \ell \boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}} \boldsymbol{m}^\mathsf{T} \boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}} \boldsymbol{y} - 4\ell\epsilon I_n.\end{aligned}$$

Here, c and γ^{-1} are constants that depend only polynomially on ϵ^{-1} and α_1 and α_2 as in (55).

Proof. First we bound $\mathbb{E}\left[G_\epsilon(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})\right]$ for fixed $\boldsymbol{h}, \boldsymbol{x} \in \mathcal{S}^{n-1}$ and $\boldsymbol{m}, \boldsymbol{y} \in \mathcal{S}^{p-1}$. Noting that $t_\epsilon(z) \ge = \mathbf{1}_{z>0}(z) - \mathbf{1}_{0\le z\le\epsilon}(z)$, we have

$$\begin{aligned}&\mathbb{E}\left[G_\epsilon(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})\right] \\ \succeq & \ell \mathbb{E}\Big[\Big(\mathbf{1}_{\boldsymbol{b}_i^\mathsf{T}\boldsymbol{h}\ge 0}\mathbf{1}_{\boldsymbol{b}_i^\mathsf{T}\boldsymbol{x}\ge 0}(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{m})_+(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{y})_+ - \left(\mathbf{1}_{0\le\boldsymbol{b}_i^\mathsf{T}\boldsymbol{h}\le\epsilon} + \mathbf{1}_{0\le\boldsymbol{b}_i^\mathsf{T}\boldsymbol{x}\le\epsilon}\right)(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{m})_+(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{y})_+\Big)\boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T}\Big] \\ = & \ell \mathbb{E}\Big[\left(\mathbf{1}_{\boldsymbol{b}_i^\mathsf{T}\boldsymbol{h}\ge 0}\mathbf{1}_{\boldsymbol{b}_i^\mathsf{T}\boldsymbol{x}\ge 0}\boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T}\right)(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{m})_+(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{y})_+\Big] - 2\ell\mathbb{E}\left[\left(\mathbf{1}_{0\le\boldsymbol{b}_i^\mathsf{T}\boldsymbol{h}\le\epsilon}\boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T}\right)(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{m})_+(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{y})_+\right].\end{aligned}$$

We first note that $\mathbb{E}\left[\mathbf{1}_{\boldsymbol{b}_i^\mathsf{T}\boldsymbol{h}\ge 0}\mathbf{1}_{\boldsymbol{b}_i^\mathsf{T}\boldsymbol{x}\ge 0}\boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T}\right] = \alpha_1\boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}}$ and $\mathbb{E}\left[(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{m})_+(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{y})_+\right] = \alpha_2\boldsymbol{m}^\mathsf{T}\boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}}\boldsymbol{y}$ where α_i satisfies $0.97 < \alpha_i < 1$. Also, we have $\left|\boldsymbol{m}^\mathsf{T}\boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}}\boldsymbol{y}\right| \le \frac{1}{2}$ and $\mathbb{E}\left[\mathbf{1}_{0\le\boldsymbol{b}_i^\mathsf{T}\boldsymbol{h}\le\epsilon}\boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T}\right] \preceq \frac{\epsilon\alpha_1}{2}\boldsymbol{I}_n$. Thus,

$$\begin{aligned}\mathbb{E}\left[G_\epsilon(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})\right] &\succeq \alpha_1\alpha_2\ell \cdot \boldsymbol{m}^\mathsf{T}\boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}}\boldsymbol{y} \cdot \boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}} - 2\ell \cdot \mathbb{E}\left[\mathbf{1}_{0\le\boldsymbol{b}_i^\mathsf{T}\boldsymbol{h}\le\epsilon}\boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T}\right] \cdot \alpha_2\boldsymbol{m}^\mathsf{T}\boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}}\boldsymbol{y} \\ &\succeq \alpha_1\alpha_2\ell \cdot \boldsymbol{m}^\mathsf{T}\boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}}\boldsymbol{y} \cdot \boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}} - \frac{\epsilon\alpha_1\alpha_2\ell}{2}\boldsymbol{I}_n \\ &\succeq \alpha_1\alpha_2\ell \cdot \boldsymbol{m}^\mathsf{T}\boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}}\boldsymbol{y} \cdot \boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}} - \epsilon\ell\boldsymbol{I}_n \qquad (59)\end{aligned}$$

Second, we show concentration of $G_\epsilon(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})$ for fixed $\boldsymbol{h}, \boldsymbol{x} \in \mathcal{S}^{n-1}$ and $\boldsymbol{m}, \boldsymbol{y} \in \mathcal{S}^{p-1}$ and is similar to the steps shown in proof of Lemma 9. Let $\boldsymbol{\xi}_i = \sqrt{t_\epsilon(\boldsymbol{b}_i^\mathsf{T}\boldsymbol{h})t_\epsilon(\boldsymbol{b}_i^\mathsf{T}\boldsymbol{x})(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{m})_+(\boldsymbol{c}_i^\mathsf{T}\boldsymbol{y})_+}\boldsymbol{b}_i$. If $\ell > (2\bar{c}/\epsilon)^2 n^2 (\log n)^2$, we have

$$\|G_\epsilon(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y}) - \mathbb{E}\left[G_\epsilon(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{y})\right]\| \le \epsilon n \qquad (60)$$

with probability at least $1 - 2e^{-\bar{\gamma}\frac{\epsilon^2}{4}\frac{\ell}{n\log n}}$. Here, $\bar{c}$ and $\bar{\gamma}$ are absolute constants.

Third, we bound the Lipschitz constant of G_ϵ, and is again similar to the steps shown in proof of Lemma 9. If $\ell \geq n$ then we have

$$\begin{aligned}&\|G_\epsilon(\boldsymbol{h},\boldsymbol{x},\boldsymbol{m},\boldsymbol{y}) - G_\epsilon(\tilde{\boldsymbol{h}},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{m}},\tilde{\boldsymbol{y}})\| \\ \leq &\frac{729\ell\sqrt{np}}{\epsilon}\left[\|\boldsymbol{y}-\tilde{\boldsymbol{y}}\| + \|\boldsymbol{m}-\tilde{\boldsymbol{m}}\| + \|\boldsymbol{x}-\tilde{\boldsymbol{x}}\| + \left\|\boldsymbol{h}\boldsymbol{h}^\intercal - \tilde{\boldsymbol{h}}\tilde{\boldsymbol{h}}^\intercal\right\|\right] \end{aligned} \tag{61}$$

for all $\tilde{\boldsymbol{h}}, \tilde{\boldsymbol{x}} \in \mathcal{S}^{n-1}$ and $\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{y}} \in \mathcal{S}^{p-1}$ with probability at least $1 - 2e^{-\ell/2}$.

Finally, we complete the proof by a covering argument. Let $\mathcal{N}_\delta$ be a δ-net on $\mathcal{S}^{n-1} \times \mathcal{S}^{p-1}$ such that $|\mathcal{N}_\delta| \leq (3/\delta)^{n+p}$. Take $\delta = \frac{\epsilon^2}{2916\sqrt{np}}$. Combining (59) and (60), we have

$$\begin{aligned}\forall(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{N}_\delta, \quad G_\epsilon(\boldsymbol{h},\boldsymbol{x},\boldsymbol{m},\boldsymbol{y}) \succeq &\mathbb{E}G_\epsilon(\boldsymbol{h},\boldsymbol{x},\boldsymbol{m},\boldsymbol{y}) - \ell\epsilon I_n \\ \succeq &\alpha^2\ell\boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}}\boldsymbol{m}^\intercal\boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}}\boldsymbol{y} - 3\ell\epsilon I_n\end{aligned}$$

with probability at least

$$1-2|\mathcal{N}_\delta|e^{-\gamma_K\epsilon^2\frac{\ell}{4n\log n}} \geq 1-2\left(\frac{3}{\delta}\right)^{n+p}e^{-\gamma_K\epsilon^2\frac{\ell}{4n\log n}} \geq 1-2e^{-\gamma_K\epsilon^2\frac{\ell}{4n\log n}+(n+p)\log(108\sqrt{np}/\epsilon^2)}.$$

If $\ell \geq \tilde{c}n(n+p)\log(n)\log(np)$ for some $\tilde{c} = \Omega(\epsilon^2\log\epsilon)$, then this probability is at least $1 - 2e^{-\tilde{\gamma}\ell/(n\log n)}$ for some $\tilde{\gamma} = O(\epsilon^2)$. For $(\boldsymbol{h},\boldsymbol{m}),(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{S}^{n-1}\times\mathcal{S}^{p-1}$, let $(\tilde{\boldsymbol{h}},\tilde{\boldsymbol{m}}),(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) \in \mathcal{N}_\delta$ be such that $\|\boldsymbol{h}-\tilde{\boldsymbol{h}}\|_2 \leq \delta$, $\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_2 \leq \delta$, $\|\boldsymbol{m}-\tilde{\boldsymbol{m}}\|_2 \leq \delta$ and $\|\boldsymbol{y}-\tilde{\boldsymbol{y}}\|_2 \leq \delta$. By (61), we have that

$$\begin{aligned}\forall(\boldsymbol{h},\boldsymbol{x}) \neq (\boldsymbol{0},\boldsymbol{0}) \text{ and } \boldsymbol{m},\boldsymbol{y} \in \mathcal{S}^{p-1}, \quad &G_\epsilon(\boldsymbol{h},\boldsymbol{x},\boldsymbol{m},\boldsymbol{y}) \\ &\succeq \alpha_1\alpha_2\ell\boldsymbol{Q}_{\boldsymbol{h},\boldsymbol{x}}\boldsymbol{m}^\intercal\boldsymbol{Q}_{\boldsymbol{m},\boldsymbol{y}}\boldsymbol{y} - 4\ell\epsilon\boldsymbol{I}_n.\end{aligned}$$

In conclusion, the result of this lemma holds if $\ell > (2\bar{c}/\epsilon)^2n^2(\log n)^2$ and $\ell \geq \tilde{c}n(n+p)\log(n)\log(np)$, with probability at least $1 - 2e^{-\ell/2} - 2e^{-\tilde{\gamma}\ell/(n\log n)} > 1 - 4e^{-\gamma\ell/(n\log n)}$ for some $\gamma = O(\epsilon^2)$ and $\tilde{c} = \Omega(\epsilon^2\log\epsilon)$. □

5.6 Concentration of matrices with sub-gaussian rows

The proof of Lemmas 6 and 7 require results from concentration of sub-exponential random variables that has a better dependence on the sub-exponential parameters. To this end, we use the following Bernstein inequality and refer the readers to Jeong et al. [2019] for a proof of the theorem.

Theorem 5. *Let $\boldsymbol{a} = (a_1,\ldots,a_n)$ be a fixed non-zero vector and let $y_1,\ldots,y_m$ be independent, mean zero sub-exponential random variables satisfying $\mathbb{E}|y_i| \leq 2$ and $\|y_i\|_{\psi_1} \leq K_i^2$ ($K_i \geq 2$). Then for every $u \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^m a_iy_i\right| \geq u\right) \leq 2\exp\left[-c\min\left(\frac{u^2}{\sum_{i=1}^m a_i^2K_i^2\log K_i}, \frac{u}{\|\boldsymbol{a}\|_\infty K^2\log K}\right)\right],$$

where $K = \max_i K_i$ and c is an absolute constant.

We now state a theorem that controls the singular values of a random matrix $\boldsymbol{A}$. The Theorem is exactly the same as Theorem 5.39 in Vershynin [2012] with the notable difference in the dependence of the constants to the sub-gaussian parameters. We use Theorem 5 to get this improved dependence.

Theorem 6. *Let $\boldsymbol{A}$ be a $N \times n$ matrix whose rows $\boldsymbol{a}_i$ are independent sub-gaussian isotropic random vectors in $\mathbb{R}^n$. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-ct^2)$ one has*

$$\sqrt{N} - C\sqrt{n} - t \leq s_{\min}(\boldsymbol{A}) \leq s_{\max}(\boldsymbol{A}) \leq \sqrt{N} + C\sqrt{n} + t. \tag{62}$$

Here $C = C_K = K\sqrt{\log K}\sqrt{\frac{\log 9}{c_1}}$, $c = c_K = \frac{c_1}{K^2\log K} > 0$ with c_1 is an absolute constant and $K = \max_i \|\boldsymbol{a}_i\|_{\psi_2}$.

The proof structure of Theorem 6 is exactly the same as the proof of Theorem 5.39 in Vershynin [2012], and so we provide the proof presented in Vershynin [2012] below.

Proof. The proof is a basic version of a covering argunemt, and it has three steps. We need to control $\|\boldsymbol{A}\boldsymbol{x}\|_2$ for all vectors on the unit sphere. To this end, we discretize the sphere using a $\mathcal{N}$ (the approximation step), establish a tight control of $\|\boldsymbol{A}\boldsymbol{x}\|_2$ for every fixed vector $\boldsymbol{x} \in \mathcal{N}$ with high probability (the concentration step), and finish off by taking a union bound over all $\boldsymbol{x}$ in the net.

Step 1: Approximation. Using Lemma 5.36 in Vershynin [2012] for the matrix $\boldsymbol{B} = \boldsymbol{A}/\sqrt{N}$ we see that the conclusion of the theorem is equivalent to

$$\|\tfrac{1}{N}\boldsymbol{A}^\intercal\boldsymbol{A} - \boldsymbol{I}\| \le \max(\delta, \delta^2) =: \epsilon \text{ where } \delta = C\sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}}. \tag{63}$$

Using Lemma 5.34 in Vershynin [2012], we can evaluate the operator norm in (63) on a $\frac{1}{4}$-net $\mathcal{N}$ pf unit sphere $\mathcal{S}^{n-1}$:

$$\|\frac{1}{N}\boldsymbol{A}^\intercal\boldsymbol{A} - \boldsymbol{I}\| \le 2\max_{\boldsymbol{x}\in\mathcal{N}}\left|\left\langle(\frac{1}{N}\boldsymbol{A}^\intercal\boldsymbol{A} - \boldsymbol{I})\boldsymbol{x}, \boldsymbol{x}\right\rangle\right| = 2\max_{\boldsymbol{x}\in\mathcal{N}}|\frac{1}{N}\|\boldsymbol{A}\boldsymbol{x}\|_2^2 - 1|.$$

So to complete the proof it suffices to show that, which high probability,

$$\max_{\boldsymbol{x}\in\mathcal{N}}|\frac{1}{N}\|\boldsymbol{A}\|_2^2 - 1| \le \frac{\epsilon}{2}. \tag{64}$$

By Lemma 5.2 in Vershynin [2012], we can choose the net $\mathcal{N}$ so that it has cardinality $|\mathcal{N}| \le 9^n$.

Step 2: Concentration Let us fix any vector $\boldsymbol{x} \in \mathcal{S}^{n-1}$. We can express $\|\boldsymbol{A}\boldsymbol{x}\|_2^2$ as a sum of independent random variablies

$$\|\boldsymbol{A}\boldsymbol{x}\|_2^2 = \sum_{i=1}^N \langle \boldsymbol{a}_i, \boldsymbol{x}\rangle =: \sum_{i=1}^N z_i^2 \tag{65}$$

where $\boldsymbol{a}_i$ denote the rows of the matrix $\boldsymbol{A}$. By assumption, $z_i = \langle \boldsymbol{a}_i, \boldsymbol{x}\rangle$ are independent sub-gaussian random variables with $\mathbb{E}z_i^2 = 1$ and $\|z_i\|_{\psi_2} \le K$. Therefore, by Remark 5.18 and Lemma 5.14 in Vershynin [2012], $z_i^2 - 1$ are independent centered sub-exponential random variables with $\|z_i^2 - 1\|_{\psi_1} \le 2\|z_i^2\|_{\psi_1} \le 4\|z_i\|_{\psi_2}^2 = 4K^2$.

We can therefore use an exponential deviation inequality, Theorem 5, to control the sum (65).

$$\begin{aligned}
\mathbb{P}\left\{|\frac{1}{N}\|\boldsymbol{A}\|_2^2 - 1| \ge \frac{\epsilon}{2}\right\} &= \mathbb{P}\left\{|\frac{1}{N}\sum_{i=1}^N z_i^2 - 1| \ge \frac{\epsilon}{2}\right\} \\
&\le 2\exp\left[-\tilde{c}_1 \min\left(\frac{\epsilon^2 N^2/4}{\sum_{i=1}^N 4K_i^2 \log 2K_i}, \frac{\epsilon N/2}{4K\log 2K}\right)\right] \\
&\le 2\exp\left[-\frac{\tilde{c}_1}{4K^2\log 2K}\min\left(\epsilon^2, \epsilon\right)N\right] \\
&= 2\exp\left[-\frac{\tilde{c}_1}{4K^2\log 2K}\delta N\right] \\
&\le 2\exp\left[-\frac{c_1}{K^2\log K}\left(C^2 n + t^2\right)\right],
\end{aligned}$$

where the last inequality follows by the definition of δ and using the inequality $(a+b)^2 \ge a^2 + b^2$ for $a, b \ge 0$.

Step 3: Union bound. Taking the union bound over all vectors $\boldsymbol{x}$ in the net $\mathcal{N}$ of cardinality $|\mathcal{N}| \le 9^n$, we obtain

$$\mathbb{P}\left\{\max_{\boldsymbol{x}\in\mathcal{N}}|\frac{1}{N}\|\boldsymbol{A}\boldsymbol{x}\|_2^2 - 1| \ge \frac{\epsilon}{2}\right\} \le 9^n \cdot 2\exp[-\frac{c_1}{K^2\log K}\left(C^2 n + t^2\right)] \le 2\exp[-\frac{c_1}{K^2\log K}t^2],$$

where the second inequality follows for $C = C_K$ sufficiently large, e.g. $C = K\sqrt{\log K}\sqrt{\frac{\log 9}{c_1}}$. $\square$

We now state a corollary of Theorem 6 that applies to general, non-isotropic sub-gaussian distribution.

Corollary 1. *Let $\boldsymbol{A}$ be a $N \times n$ matrix whose rows $\boldsymbol{a}_i$ are independent sub-gaussian random vectors in $\mathbb{R}^n$ with second moment matrix $\boldsymbol{\Sigma}$. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-ct^2)$ one has*

$$\|\tfrac{1}{N}\boldsymbol{A}^\intercal\boldsymbol{A} - \boldsymbol{\Sigma}\| \leq \max(\delta, \delta^2) \text{ where } \delta = C\sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}}. \tag{66}$$

Here $C = C_K = K\sqrt{\log K}\sqrt{\frac{\log 9}{c_1}}$, $c = c_K = \frac{c_1}{K^2 \log K} > 0$ with c_1 is an absolute constant and $K = \max_i \|\boldsymbol{a}_i\|_{\psi_2}$.