A. Additional Notations

Given a sub-exponential random variable X, let $\|X\|_{\psi_1} = \inf\{t > 0 \ : \ \mathbb{E}[\exp(|X|/t)] \leq 2\}$. Similarly, for a sub-gaussian random variable, $\|X\|_{\psi_2} = \inf\{t > 0 \ : \ \mathbb{E}[\exp(X^2/t^2)] \leq 2\}$.

B. Proof of Theorem 3.2

Let us first get some useful estimates from the data. By Assumptions 2.1 and 2.2, we have $\|x_i\|_2^2 = \Theta(d)$ for all $i \in [N]$ w.p. $\geq 1 - Ne^{-\Omega(d)}$. For a given pair $i \neq j$, let x_j be fixed and x_i be random, then $\langle x_i, x_j \rangle$ is Lipschitz continuous w.r.t. x_i, where the Lipschitz constant is given by $\|x_j\|_2 = \mathcal{O}(\sqrt{d})$. Thus, it follows from Assumption 2.2 that $\mathbb{P}(|\langle x_i, x_j \rangle| > t) \leq 2e^{-t^2/\mathcal{O}(d)}$. By picking $t = dN^{-1/(r-0.5)}$ and doing a union bound over all data pairs, we get $\max_{i \neq j} |\langle x_i, x_j \rangle|^r \leq dN^{-1/(r-0.5)}$ w.p. at least $1 - N^2 e^{-\Omega\left(dN^{-2/(r-0.5)}\right)}$. Combining these two events, we obtain that the following hold

$$\begin{aligned} &\|x_i\|_2^2 = \Theta(d), \ \forall\, i \in [N], \\ &|\langle x_i, x_j \rangle|^r \leq dN^{-1/(r-0.5)}, \ \forall\, i \neq j \end{aligned} \tag{36}$$

with the same probability as stated in the theorem.

We have from Lemma 3.1 that

$$K^{(L)} = \sum_{l=1}^{L} G^{(l)} \circ \dot{G}^{(l+1)} \circ \dot{G}^{(l+2)} \circ \ldots \circ \dot{G}^{(L)}.$$

One also observes that all the matrices $G^{(l)}, \dot{G}^{(l)}, G^{(l)}$ are positive semidefinite. Recall that, for two p.s.d. matrices $P, Q \in \mathbb{R}^{n \times n}$, one has $\lambda_{\min}(P \circ Q) \geq \lambda_{\min}(P) \min_{i \in [n]} Q_{ii}$ (Schur, 1911). Thus, it holds

$$\lambda_{\min}\left(K^{(L)}\right) \geq \sum_{l=1}^{L} \lambda_{\min}\left(G^{(l)}\right) \min_{i \in [N]} \prod_{p=l+1}^{L} (\dot{G}^{p})_{ii} = \sum_{l=1}^{L} \lambda_{\min}\left(G^{(l)}\right),$$

where the last equality follows from the fact that $(\dot{G}^{(p)})_{ii} = 1$ for all $p \in [2, L], i \in [N]$. From here, it suffices to bound $\lambda_{\min}\left(G^{(2)}\right)$. Let $D = \operatorname{diag}([\|x_i\|_2]_{i=1}^N)$ and $\hat{X} = D^{-1}X$. Then, by the homogeneity of σ, we have $\sigma(Xw) = \sigma(D\hat{X}w) = D\sigma(\hat{X}w)$, and thus

$$\begin{aligned} \lambda_{\min}\left(G^{(2)}\right) &= \lambda_{\min}\left(D\mathbb{E}\left[\sigma(\hat{X}w)\sigma(\hat{X}w)^T\right]D\right) \\ &= \lambda_{\min}\left(D\left[\mu_0(\sigma)^2 1_N 1_N^T + \sum_{s=1}^{\infty} \mu_s(\sigma)^2 (\hat{X}^{*s})(\hat{X}^{*s})^T\right]D\right) \\ &\geq \mu_r(\sigma)^2 \lambda_{\min}\left(D(\hat{X}^{*r})(\hat{X}^{*r})^T D\right) \\ &= \mu_r(\sigma)^2 \lambda_{\min}\left(D^{-(r-1)}(X^{*r})(X^{*r})^T D^{-(r-1)}\right) \\ &\geq \mu_r(\sigma)^2 \frac{\lambda_{\min}\left((X^{*r})(X^{*r})^T\right)}{\max_{i \in [N]} \|x_i\|_2^{2(r-1)}}, \end{aligned} \tag{37}$$

where the second step uses the Hermite expansion of σ (for the proof see Lemma D.3 of (Nguyen & Mondelli, 2020)). By Gershgorin circle theorem, one has

$$\lambda_{\min}\left((X^{*r})(X^{*r})^T\right) \geq \min_{i \in [N]} \|x_i\|_2^{2r} - (N-1) \max_{i \neq j} |\langle x_i, x_j \rangle|^r \geq \Omega(d),$$

where the last estimate follows from (36). Plugging this and the estimate of (36) into the inequality (37) proves the lower bound on the smallest eigenvalue of the NTK. For the upper bound, note that

$$\lambda_{\min}\left(K^{(L)}\right) \leq \frac{\operatorname{tr}(K^{(L)})}{N} = \frac{1}{N} \sum_{i=1}^{N} \sum_{l=1}^{L} (G^{(l)})_{ii} \prod_{p=l+1}^{L} (\dot{G}^{p})_{ii}.$$

One observes that $(G^{(l)})_{ii} = 2\mathbb{E}_{g\sim\mathcal{N}(0,(G_{l-1})_{ii})}[\sigma(g)^2] = (G^{(l-1)})_{ii}$. Iterating this argument gives $(G^{(l)})_{ii} = (G^{(1)})_{ii} = \|x_i\|_2^2$. Thus, it follows that

$$\lambda_{\min}\left(K^{(L)}\right) \leq \frac{L}{N}\operatorname{tr}(G^{(1)}) = \frac{L}{N}\sum_{i=1}^{N}\|x_i\|_2^2 = L\,\mathcal{O}(d),$$

where we used again (36) in the last estimate.

C. Some Useful Estimates

Lemma C.1 *Fix any $0 \leq k \leq L-1$ and $x \sim P_X$. Then, we have*

$$\|f_k(x)\|_2^2 = \Theta\left(d\prod_{l=1}^{k} n_l\beta_l^2\right)$$

w.p. at least $1 - \sum_{l=1}^{k}\exp\left(-\Omega\left(n_l\right)\right) - \exp(-\Omega\left(d\right))$ over $(W_l)_{l=1}^k$ and x. Moreover,

$$\mathbb{E}_x\|f_k(x)\|_2^2 = \Theta\left(d\prod_{l=1}^{k} n_l\beta_l^2\right)$$

w.p. $1 - \sum_{l=1}^{k}\exp\left(-\Omega\left(n_l\right)\right)$ over $(W_l)_{l=1}^k$.

Lemma C.2 *Fix any $k \in [L-1]$. Then, we have*

$$\|\mathbb{E}_x[f_k(x)]\|_2^2 = \Theta\left(d\prod_{l=1}^{k} n_l\beta_l^2\right)$$

w.p. at least $1 - \sum_{l=1}^{k}\exp\left(-\Omega\left(n_l\right)\right)$ over $(W_l)_{l=1}^k$.

Lemma C.3 *Fix any $k \in [L-1]$. Assume $\prod_{l=1}^{k-1}\log(n_l) = o\left(\min_{l\in[0,k]} n_l\right)$. Then, we have*

$$\|f_k(x_i) - \mathbb{E}_x[f_k(x)]\|_2^2 = \Theta\left(d\prod_{l=1}^{k} n_l\beta_l^2\right), \quad \forall\, i \in [N] \tag{38}$$

w.p. at least

$$1 - N\exp\left(-\Omega\left(\frac{\min_{l\in[0,k]} n_l}{\prod_{l=1}^{k-1}\log(n_l)}\right)\right) - \sum_{l=1}^{k}\exp(-\Omega\left(n_l\right)).$$

Lemma C.4 *Fix any $k \in [L-1]$. Then, we have*

$$\mathbb{E}_x\|f_k(x) - \mathbb{E}_x[f_k(x)]\|_2^2 = \Theta\left(d\prod_{l=1}^{k} n_l\beta_l^2\right)$$

w.p. at least $1 - \sum_{l=1}^{k}\exp\left(-\Omega\left(n_l\right)\right)$ over $(W_l)_{l=1}^k$.

Lemma C.5 *Fix any $k \in [L-1]$, and $x \sim P_X$. Then, we have that $\|\Sigma_k(x)\|_F^2 = \Theta\left(n_k\right)$ w.p. at least $1 - \sum_{l=1}^{k}\exp\left(-\Omega\left(n_l\right)\right) - \exp(-\Omega\left(d\right))$ over $(W_l)_{l=1}^k$ and x.*

Lemma C.6 *Fix any $k \in [L-1], k \leq p \leq L-1$, and $x \sim P_X$. Then, we have that*

$$\left\|\Sigma_k(x)\prod_{l=k+1}^{p} W_l\Sigma_l(x)\right\|_F^2 = \Theta\left(n_k\prod_{l=k+1}^{p} n_l\beta_l^2\right)$$

w.p. at least $1 - \sum_{l=1}^{p}\exp\left(-\Omega\left(n_l\right)\right) - \exp(-\Omega\left(d\right))$ over $(W_l)_{l=1}^p$ and x.

C.1. Proof of Lemma C.1

The proof works by induction over k. Note that the statement holds for $k = 0$ due to Assumptions 2.1 and 2.2. Assume that the lemma holds for some $k-1$, i.e. $\|f_{k-1}(x)\|_2^2 = \Theta\left(d\prod_{l=1}^{k-1} n_l\beta_l^2\right)$ w.p. at least $1 - \sum_{l=1}^{k-1} N\exp\left(-\Omega\left(n_l\right)\right) - N\exp\left(-\Omega\left(d\right)\right)$. Let us condition on this event of $(W_l)_{l=1}^{k-1}$ and study probability bounds over W_k. Let $W_k = [w_1, \ldots, w_{n_k}]^T$ where $w_j \sim \mathcal{N}(0, \beta_k^2\,\mathbb{I}_{n_{k-1}})$. Note that

$$\|f_k(x)\|_2^2 = \sum_{j=1}^{n_k} f_{k,j}(x)^2, \tag{39}$$

and that

$$\mathbb{E}_{W_k}\|f_k(x)\|_2^2 = \sum_{j=1}^{n_k}\mathbb{E}_{w_j}[f_{k,j}(x)^2] = \frac{n_k\beta_k^2}{2}\|f_{k-1}(x)\|_2^2 = \Theta\left(d\prod_{l=1}^{k} n_l\beta_l^2\right),$$

where the last equality follows from the induction assumption. Furthermore,

$$\left\|f_{k,j}(x)^2\right\|_{\psi_1} = \|f_{k,j}(x)\|_{\psi_2}^2 \le c\beta_k^2\|f_{k-1}(x)\|_2^2 = \mathcal{O}\left(\beta_k^2 d\prod_{l=1}^{k-1} n_l\beta_l^2\right),$$

where c is an absolute constant. Thus, by applying Bernstein's inequality (see Theorem 2.8.1 of (Vershynin, 2018)) to the sum of i.i.d. random variables in (39), we have

$$\frac{1}{2}\mathbb{E}_{W_k}\|f_k(x)\|_2^2 \le \|f_k(x)\|_2^2 \le \frac{3}{2}\mathbb{E}_{W_k}\|f_k(x)\|_2^2$$

w.p. at least $1 - \exp\left(-\Omega\left(n_k\right)\right)$. Taking the intersection of the two events finishes the proof for $\|f_k(x)\|_2^2$. The proof for $\mathbb{E}_x\|f_k(x)\|_2^2$ can be done by following similar passages and using that $\left\|\mathbb{E}_x[f_{k,j}(x)^2]\right\|_{\psi_1} \le \mathbb{E}_x\left\|f_{k,j}(x)^2\right\|_{\psi_1}$.

C.2. Proof of Lemma C.2

The upper bound follows from Lemma C.1 via Jensen's inequality. The proof for the lower bound works by induction on k. Assume it holds for $k-1$ that $\|\mathbb{E}_x[f_{k-1}(x)]\|_2^2 = \Omega\left(d\prod_{l=1}^{k-1} n_l\beta_l^2\right)$ w.p. at least $1 - \sum_{l=1}^{k-1}\exp\left(-\Omega\left(n_l\right)\right)$ over $(W_l)_{l=1}^{k-1}$. Let us condition on the intersection of this event and that of Lemma C.1 for $(W_l)_{l=1}^{k-1}$. Let $W_k = [w_1, \ldots, w_{n_k}]$ where $w_j \sim \mathcal{N}(0, \beta_k^2\,\mathbb{I}_{n_{k-1}})$. For every $j \in [n_k]$,

$$\left\|(\mathbb{E}_x[f_{k,j}(x)])^2\right\|_{\psi_1} = \|\mathbb{E}_x[f_{k,j}(x)]\|_{\psi_2}^2 \le \mathbb{E}_x\|[f_{k,j}(x)]\|_{\psi_2}^2 \le c\beta_k^2\,\mathbb{E}_x\|f_{k-1}(x)\|_2^2 = \mathcal{O}\left(d\beta_k^2\prod_{l=1}^{k-1} n_l\beta_l^2\right),$$

where c is an absolute constant and the last equality follows from the above conditional event from Lemma C.1. Moreover,

$$\begin{aligned}\mathbb{E}_{W_k}\|\mathbb{E}_x[f_k(x)]\|_2^2 &= \sum_{j=1}^{n_k}\mathbb{E}_{w_j}(\mathbb{E}_x[f_{k,j}(x)])^2 \ge \sum_{j=1}^{n_k}(\mathbb{E}_x\mathbb{E}_{w_j}[f_{k,j}(x)])^2 = \frac{n_k\beta_k^2}{2\pi}\left(\mathbb{E}_x\|f_{k-1}(x)\|\right)^2 \\ &\ge \frac{n_k\beta_k^2}{2\pi}\|\mathbb{E}_x[f_{k-1}(x)]\|_2^2 = \Omega\left(d\prod_{l=1}^{k} n_l\beta_l^2\right),\end{aligned}$$

where the last estimate follows from our induction assumption. By Bernstein's inequality (see Theorem 2.8.1 of (Vershynin, 2018)), we have

$$\|\mathbb{E}_x[f_k(x)]\|_2^2 \ge \frac{1}{2}\mathbb{E}_{W_k}\|\mathbb{E}_x[f_k(x)]\|_2^2 = \Omega\left(d\prod_{l=1}^{k} n_l\beta_l^2\right)$$

w.p. at least $1 - \exp\left(-n_k\right)$ over W_k. Taking the intersection of all these events finishes the proof.

C.3. Proof of Lemma C.3

Let $Z : \mathbb{R}^d \to \mathbb{R}$ be a random function over x_i defined as $Z(x_i) = \|f_k(x_i) - \mathbb{E}_x[f_k(x)]\|_2$. It follows from Theorem 6.2 that w.p. at least $1 - \sum_{l=1}^k \exp\left(-\Omega\left(n_l\right)\right)$ over $(W_l)_{l=1}^k$,

$$\|Z\|_{\text{Lip}}^2 = \mathcal{O}\left(\frac{\prod_{l=0}^k n_l}{\min_{l\in[0,k]} n_l} \prod_{l=1}^{k-1} \log(n_l) \prod_{l=1}^{k} \beta_l^2\right) = o\left(d\prod_{l=1}^{k} n_l\beta_l^2\right). \tag{40}$$

Below, let us denote the shorthand

$$\mathbb{E}[Z] = \mathbb{E}_{x_i}[Z(x_i)] = \int_{\mathbb{R}^d} Z(x_i) dP_X(x_i).$$

It holds

$$\begin{aligned}
\mathbb{E}[Z]^2 &= \mathbb{E}[Z^2] - \mathbb{E}[|Z - \mathbb{E}Z|^2] \\
&\geq \mathbb{E}[Z^2] - \int_0^\infty \mathbb{P}(|Z - \mathbb{E}Z| > \sqrt{t}) dt \\
&\geq \mathbb{E}[Z^2] - \int_0^\infty 2\exp\left(-\frac{c\,t}{\|Z\|_{\text{Lip}}^2}\right) dt \\
&= \mathbb{E}[Z^2] - \frac{2}{c}\|Z\|_{\text{Lip}}^2,
\end{aligned} \tag{41}$$

where the 2nd inequality follows from Assumption 2.2. By Lemma C.4, we have w.p. at least $1 - \sum_{l=1}^k \exp\left(-\Omega\left(n_l\right)\right)$ over $(W_l)_{l=1}^k$ that

$$\mathbb{E}[Z^2] = \Theta\left(d\prod_{l=1}^{k} n_l\beta_l^2\right). \tag{42}$$

By combining (40), (41) and (42), we obtain that $\mathbb{E}[Z] = \Omega\left(\sqrt{d\prod_{l=1}^k n_l\beta_l^2}\right)$. Moreover, $\mathbb{E}[Z] \leq \sqrt{\mathbb{E}[Z^2]} = \mathcal{O}\left(\sqrt{d\prod_{l=1}^k n_l\beta_l^2}\right)$. As a result, we have that $\mathbb{E}[Z] = \Theta\left(\sqrt{d\prod_{l=1}^k n_l\beta_l^2}\right)$ w.p. at least $1 - \sum_{l=1}^k \exp(-\Omega\left(n_l\right))$ over $(W_l)_{l=1}^k$. Let us condition on this event and study probability bounds over the samples. Using Assumption 2.2, we have $\frac{1}{2}\mathbb{E}[Z] \leq Z \leq \frac{3}{2}\mathbb{E}[Z]$, hence $Z = \Theta\left(\sqrt{d\prod_{l=1}^k n_l\beta_l^2}\right)$, w.p. at least

$$1 - \exp\left(-\Omega\left(\frac{\min_{l\in[0,k]} n_l}{\prod_{l=1}^{k-1}\log(n_l)}\right)\right).$$

Taking the union bound over N samples, followed by an intersection with the above event over the weights, finishes the proof.

C.4. Proof of Lemma C.4

The proof works by induction on k. Note that the statement holds for $k = 0$ due to Assumption 2.1. Let us assume for now that the result holds for the first k layers. To prove it for layer k, we condition on the intersection of this event and the event of Lemma C.1 for $(W_l)_{l=1}^{k-1}$, and study probability bounds over W_k. Define $W_k = [w_1, \ldots, w_{n_k}] \in \mathbb{R}^{n_{k-1}\times n_k}$ where $w_j \sim \mathcal{N}(0, \beta_k^2\mathbb{I}_{n_{k-1}})$. Recall that by definition, $f_{k,j}(x) = \sigma(\langle w_j, f_{k-1}(x)\rangle)$ for $j \in [n_k]$. We have that

$$\mathbb{E}_x\|f_k(x) - \mathbb{E}_x[f_k(x)]\|_2^2 = \sum_{j=1}^{n_k} \mathbb{E}_x\Big(f_{k,j}(x) - \mathbb{E}_x[f_{k,j}(x)]\Big)_2^2.$$

Taking the expectation over W_k, we have

$$\begin{aligned}
&\mathbb{E}_{W_k}\mathbb{E}_x \left\|f_k(x) - \mathbb{E}_x[f_k(x)]\right\|_2^2 \\
&= \mathbb{E}_{W_k}\mathbb{E}_x \left\|f_k(x)\right\|_2^2 - \mathbb{E}_{W_k} \left\|\mathbb{E}_x[f_k(x)]\right\|_2^2 \\
&= \frac{n_k\beta_k^2}{2}\,\mathbb{E}_x \left\|f_{k-1}(x)\right\|_2^2 - \mathbb{E}_x\mathbb{E}_y \sum_{j=1}^{n_k} \mathbb{E}_{w_j}\sigma\left(\langle w_j, f_{k-1}(x)\rangle\right)\sigma\left(\langle w_j, f_{k-1}(y)\rangle\right) \\
&= \frac{n_k\beta_k^2}{2}\,\mathbb{E}_x \left\|f_{k-1}(x)\right\|_2^2 - n_k\beta_k^2\,\mathbb{E}_x\mathbb{E}_y \left\|f_{k-1}(x)\right\|_2 \left\|f_{k-1}(y)\right\|_2 \sum_{r=0}^{\infty} \mu_r(\sigma)^2 \left\langle \frac{f_{k-1}(x)}{\left\|f_{k-1}(x)\right\|_2}, \frac{f_{k-1}(y)}{\left\|f_{k-1}(y)\right\|_2} \right\rangle^r \\
&\geq \frac{n_k\beta_k^2}{2}\,\mathbb{E}_x \left\|f_{k-1}(x)\right\|_2^2 - \mu_1(\sigma)^2\, n_k\beta_k^2\, \left\|\mathbb{E}_x[f_{k-1}(x)]\right\|_2^2 - n_k\beta_k^2 \sum_{\substack{r=0 \\ r\neq 1}}^{\infty} \mu_r(\sigma)^2 (\mathbb{E}_x \left\|f_{k-1}(x)\right\|)^2 \\
&= \frac{n_k\beta_k^2}{2}\,\mathbb{E}_x \left\|f_{k-1}(x)\right\|_2^2 - \frac{n_k\beta_k^2}{4}\, \left\|\mathbb{E}_x[f_{k-1}(x)]\right\|_2^2 - \frac{n_k\beta_k^2}{4}\,(\mathbb{E}_x \left\|f_{k-1}(x)\right\|)^2,
\end{aligned}$$

where in the last step we use that $\mu_1(\sigma)^2 = 1/4$ and that $\sum_{\substack{r=0 \\ r\neq 1}}^{\infty} \mu_r(\sigma)^2 = 1/4$. Furthermore, the RHS of the last expression can be lower bounded by

$$\frac{n_k\beta_k^2}{4}\left(\mathbb{E}_x \left\|f_{k-1}(x)\right\|_2^2 - \left\|\mathbb{E}_x[f_{k-1}(x)]\right\|_2^2\right) = \frac{n_k\beta_k^2}{4}\,\mathbb{E}_x \left\|f_{k-1}(x) - \mathbb{E}_x[f_{k-1}(x)]\right\|_2^2 = \Omega\left(d\prod_{l=1}^{k} n_l\beta_l^2\right),$$

where the last step follows by induction assumption. Moreover, it follows from above that

$$\mathbb{E}_{W_k}\mathbb{E}_x \left\|f_k(x) - \mathbb{E}_x[f_k(x)]\right\|_2^2 \leq \frac{n_k\beta_k^2}{2}\,\mathbb{E}_x \left\|f_{k-1}(x)\right\|_2^2 = \mathcal{O}\left(d\prod_{l=1}^{k} n_l\beta_l^2\right),$$

where the last estimate follows from Lemma C.1. For every $j \in [n_k]$,

$$\begin{aligned}
\left\|\mathbb{E}_x\Big(f_{k,j}(x) - \mathbb{E}_x[f_{k,j}(x)]\Big)^2\right\|_{\psi_1} &\leq \mathbb{E}_x \left\|\Big(f_{k,j}(x) - \mathbb{E}_x[f_{k,j}(x)]\Big)^2\right\|_{\psi_1} \\
&= \mathbb{E}_x \left\|f_{k,j}(x) - \mathbb{E}_x[f_{k,j}(x)]\right\|_{\psi_2}^2 \\
&\leq c\,\mathbb{E}_x \left\|f_{k,j}(x)\right\|_{\psi_2}^2 \\
&\leq c\,\mathbb{E}_x \left(\left\|f_{k,j}(x) - \mathbb{E}_{w_j}[f_{k,j}(x)]\right\|_{\psi_2}^2 + \left|\mathbb{E}_{w_j}[f_{k,j}(x)]\right|^2\right) \\
&\leq c\,\mathbb{E}_x \left(\beta_k^2 \left\|f_{k,j}(x)\right\|_{\mathrm{Lip}}^2 + \frac{\beta_k^2}{2\pi} \left\|f_{k-1}(x)\right\|_2^2\right) \\
&\leq c\beta_k^2\,\mathbb{E}_x \left\|f_{k-1}(x)\right\|_2^2 \\
&= \mathcal{O}\left(\beta_k^2 d\prod_{l=1}^{k-1} \beta_l^2 n_l\right),
\end{aligned}$$

where c is an absolute constant (which is allowed to change from line to line) and the last step uses Lemma C.1. By Bernstein's inequality (see Theorem 2.8.1 of (Vershynin, 2018)),

$$\frac{1}{2}\mathbb{E}_{W_k}\mathbb{E}_x \left\|f_k(x) - \mathbb{E}_x[f_k(x)]\right\|_2^2 \leq \mathbb{E}_x \left\|f_k(x) - \mathbb{E}_x[f_k(x)]\right\|_2^2 \leq \frac{3}{2}\mathbb{E}_{W_k}\mathbb{E}_x \left\|f_k(x) - \mathbb{E}_x[f_k(x)]\right\|_2^2,$$

w.p. at least $1 - \exp\left(-\Omega\left(n_k\right)\right)$ over W_k. Thus, with that probability, we have that

$$\mathbb{E}_x \left\|f_k(x) - \mathbb{E}_x[f_k(x)]\right\|_2^2 = \Theta\left(d\prod_{l=1}^{k} n_l\beta_l^2\right).$$

Taking the intersection of all the events finishes the proof.

C.5. Proof of Lemma C.5

Proof: By Lemma C.1, we have $f_{k-1}(x) \neq 0$ w.p. at least $1 - \sum_{l=1}^{k-1} \exp\left(-\Omega\left(n_l\right)\right) - \exp(-\Omega\left(d\right))$ over $(W_l)_{l=1}^{k-1}$ and x. Let us condition on this event and derive probability bounds over W_k. Let $W_k = [w_1, \ldots, w_{n_k}]$. Then, $\|\Sigma_k(x)\|_F^2 = \sum_{j=1}^{n_k} \sigma'(\langle f_{k-1}(x), w_j\rangle)$. Thus,

$$\mathbb{E}_{W_k} \|\Sigma_k(x)\|_F^2 = n_k \mathbb{E}_{w_1}[\sigma'(-\langle f_{k-1}(x), w_1\rangle))] = n_k \mathbb{E}_{w_1}[(1 - \sigma'(\langle f_{k-1}(x), w_1\rangle))] = n_k - \mathbb{E}_{W_k} \|\Sigma_k(x)\|_F^2 ,$$

where we used the fact that w_j has a symmetric distribution, $\sigma'(t) = 1 - \sigma'(-t)$ for $t \neq 0$, and the set of $w_1 \in \mathbb{R}^{n_{k-1}}$ for which $\langle f_{k-1}(x), w_j\rangle = 0$ has measure zero. This implies that $\mathbb{E}_{W_k} \|\Sigma_k(x)\|_F^2 = n_k/2$. By Hoeffding's inequality on bounded random variables (see Theorem 2.2.6 of (Vershynin, 2018)), we have

$$\mathbb{P}\left(\left|\|\Sigma_k(x)\|_F^2 - \mathbb{E}_{W_k} \|\Sigma_k(x)\|_F^2\right| > t\right) \leq 2\exp\left(-\frac{2t^2}{n_k}\right).$$

Picking $t = n_k/4$ finishes the proof. □

C.6. Proof of Lemma C.6

The proof works by induction on p. First, Lemma C.5 implies that the statement holds for $p = k$. Suppose it holds for some $p - 1$. Note that this implies $f_{p-1}(x) \neq 0$ because otherwise $\Sigma_{p-1}(x) = 0$, which contradicts the induction assumption. Let $S_p = \Sigma_k(x) \prod_{l=k+1}^{p} W_l \Sigma_l(x)$. Then, $S_p = S_{p-1} W_p \Sigma_p(x)$. Let $W_p = [w_1, \ldots, w_{n_p}]$. Then,

$$\|S_p\|_F^2 = \sum_{j=1}^{n_p} \|S_{p-1} w_j\|_2^2 \, \sigma'(g_{p,j}(x)) = \sum_{j=1}^{n_p} \|S_{p-1} w_j\|_2^2 \, \sigma'(\langle f_{p-1}(x), w_j\rangle).$$

We have

$$\begin{aligned}
\mathbb{E}_{W_p} \|S_p\|_F^2 &= n_p \mathbb{E}_{w_1} \|S_{p-1} w_1\|_2^2 \, \sigma'(\langle f_{p-1}(x), w_1\rangle) \\
&= n_p \mathbb{E}_{w_1} \|S_{p-1}(-w_1)\|_2^2 \, \sigma'(\langle f_{p-1}(x), (-w_1)\rangle) \\
&= n_p \mathbb{E}_{w_1} \|S_{p-1} w_1\|_2^2 \, (1 - \sigma'(\langle f_{p-1}(x), w_1\rangle)) \\
&= n_p \mathbb{E}_{w_1} \|S_{p-1} w_1\|_2^2 - \mathbb{E}_{W_p} \|S_p\|_F^2 \\
&= n_p \beta_p^2 \|S_{p-1}\|_F^2 - \mathbb{E}_{W_p} \|S_p\|_F^2 ,
\end{aligned}$$

where the second step uses that w_1 has a symmetric distribution, the third step uses the fact that $\sigma'(t) = 1 - \sigma'(-t)$ for $t \neq 0$ and the set of w_1 for which $\langle f_{p-1}(x), w_1\rangle) = 0$ has measure zero. Thus,

$$\mathbb{E}_{W_p} \|S_p\|_F^2 = \frac{n_p}{2} \beta_p^2 \|S_{p-1}\|_F^2 = \Theta\left(n_k \prod_{l=k+1}^{p} n_l \beta_l^2\right),$$

where the last equality holds by induction assumption. Moreover,

$$\left\| \|S_{p-1} w_j\|_2^2 \, \sigma'(\langle f_{p-1}(x), w_j\rangle) \right\|_{\psi_1} \leq c \left\| \|S_{p-1} w_j\|_2 \right\|_{\psi_2}^2 \leq c \beta_p^2 \|S_{p-1}\|_F^2 ,$$

where c is an absolute constant (which is allowed to change from passage to passage). By Bernstein's inequality (see Theorem 2.8.1 of (Vershynin, 2018)), we have

$$\frac{1}{2} \mathbb{E}_{W_p} \|S_p\|_F^2 \leq \|S_p\|_F^2 \leq \frac{3}{2} \mathbb{E}_{W_p} \|S_p\|_F^2$$

w.p. at least $1 - e^{-\Omega(n_p)}$. Taking the intersection of all the events finishes the proof.

D. Missing Proofs from Section 4

D.1. Proof of Corollary 4.2

Let $p = \sum_{l=1}^{L} n_l n_{l-1}$. Let $\frac{\partial F_L}{\partial \theta} \in \mathbb{R}^{N \times p}$ denote the true Jacobian of F_L (without the convention that $\sigma'(0) = 0$) at a differentiable point θ. Note that, by Lemma B.2 of (Nguyen & Mondelli, 2020), $F_L(\theta)$ is locally Lipschitz, thus a.e. differentiable. Let $J(\theta) \in \mathbb{R}^{N \times p}$ be the Jacobian matrix defined in (2) (with the convention that $\sigma'(0) = 0$). Let

$$\Omega_1 = \{\theta \in \mathbb{R}^p \mid \operatorname{rank}(J(\theta)) = N\}$$

and

$$\Omega_0 = \{\theta \in \mathbb{R}^p \mid \exists l \in [L-1], j \in [n_l], i \in [N] : g_{lj}(x_i) = 0\}.$$

Let λ_p denote the Lebesgue measure in $\mathbb{R}^p$. Pick an even integer r s.t. $r \geq 0.1 + 2/\delta'$. Then, Theorem 4.1 implies that, with high probability (as stated in the corollary) over the training data, we have $\lambda_p(\Omega_1) > 0$. For every $\theta \in \Omega_1$, it holds that $f_l(\theta, x_i) \neq 0$ for all $0 \leq l \leq L-2, i \in [N]$, because otherwise $J(\theta)_{i:} = 0$ (which leads to a contradiction). Thus, every $\theta \in \Omega_1 \cap \Omega_0$ must satisfy $0 = g_{lj}(\theta, x_i) = \langle f_{l-1}(\theta, x_i), (W_l)_{:j} \rangle$ for some $l \in [L-1], j \in [n_l], i \in [N]$. The set of W_l which satisfies this equation has measure zero, and thus it holds $\lambda_p(\Omega_1 \cap \Omega_0) = 0$. Combining these facts, we get $\lambda_p(\Omega_1 \setminus \Omega_0) > 0$. Pick some $\theta_0 \in \Omega_1 \setminus \Omega_0$. Then clearly, we have the following: *(i)* $J(\theta_0) = \frac{\partial F_L}{\partial \theta}\Big|_{\theta=\theta_0}$ and *(ii)* $\operatorname{rank}(J(\theta_0)) = N$. This implies that there exists $\theta' \in \mathbb{R}^p$ such that $\left(\frac{\partial F_L}{\partial \theta}\Big|_{\theta=\theta_0}\right)\theta' = Y$ and thus,

$$y_i = \left(\left(\frac{\partial F_L}{\partial \theta}\Big|_{\theta=\theta_0}\right)\theta'\right)_i = \left\langle \frac{\partial f_L(\theta, x_i)}{\partial \theta}\Big|_{\theta_0}, \theta' \right\rangle = \lim_{\epsilon \to 0} \underbrace{\frac{f_L(\theta_0 + \epsilon\theta', x_i) - f_L(\theta_0, x_i)}{\epsilon}}_{=:h_\epsilon(x_i)}, \quad \forall\, i \in [N].$$

The result follows by noting that $h_\epsilon(x_i)$ can be implemented by a network of the same depth with twice more neurons at every hidden layer.

D.2. Proof of Lemma 4.3

By a change of index $k+1 \to k$, it is equivalent to prove the following:

$$\left\| \Sigma_k(x) \left(\prod_{l=k+1}^{L-1} W_l \Sigma_l(x) \right) W_L \right\|_2^2 = \Theta\left(\beta_L^2 \, n_k \prod_{l=k+1}^{L-1} n_l \beta_l^2 \right).$$

Let $B = \Sigma_k(x)\left(\prod_{l=k+1}^{L-1} W_l \Sigma_l(x)\right)$. By Lemma C.6, $\|B\|_F^2 = \Theta\left(n_k \prod_{l=k+1}^{L-1} n_l \beta_l^2\right)$ w.p. at least $1 - \sum_{l=1}^{L-1} \exp\left(-\Omega\left(n_l\right)\right) - \exp\left(-\Omega\left(d\right)\right)$. Moreover, one can also show that with a similar probability,

$$\|B\|_{\text{op}}^2 = \mathcal{O}\left(\frac{n_k}{\min_{l \in [k, L-1]} n_l} \prod_{l=k+1}^{L-1} n_l \beta_l^2 \right).$$

The proof of this is postponed below. Let us condition on the intersection of these two events of $(W_l)_{l=1}^{L-1}$. Then, by Hanson-Wright inequality (see Theorem 6.2.1 of (Vershynin, 2018)), we have

$$\frac{1}{2}\mathbb{E}_{W_L} \|BW_L\|_2^2 \leq \|BW_L\|_2^2 \leq \frac{3}{2}\mathbb{E}_{W_L} \|BW_L\|_2^2.$$

w.p. at least $1 - e^{-\Omega\left(\|B\|_F^2 / \|B\|_{\text{op}}^2\right)}$ over W_L. Plugging the above bounds leads to the desired result.

In the remainder of this proof, we verify the above bound of $\|B\|_{\text{op}}^2$. Concretely, we want to show that for every $p, q \in [L-1]$, the following holds w.p. at least $1 - \sum_{l=p-1}^{q} \exp\left(-\Omega\left(n_l\right)\right)$

$$\left\| \prod_{l=p}^{q} W_l \Sigma_l(x) \right\|_{\text{op}}^2 = \mathcal{O}\left(\frac{\prod_{l=p-1}^{q} n_l}{\min_{l \in [p-1, q]} n_l} \prod_{l=p}^{q} \beta_l^2 \right). \tag{43}$$

Given that, the bound of $\|B\|_{\text{op}}^2$ follows immediately by letting $p = k+1, q = L-1$, and noting $\|\Sigma_k(x)\|_{\text{op}} \leq 1$. The proof of (43) is by induction over the length $s = q - p$. First, (43) holds for $s = 0$ since $\|W_p \Sigma_p(x)\|_{\text{op}}^2 \leq \|W_p\|_{\text{op}}^2 = \mathcal{O}\left(\beta_p^2 \max(n_p, n_{p-1})\right)$ where the last estimate follows from the standard bounds on the operator norm of Gaussian matrices (see Theorem 2.12 of (Davidson & Szarek, 2001)). Suppose that (43) holds for p, q such that $q - p \leq s - 1$, and we want to prove it for all pairs p, q with $q - p = s$. It suffices to provide bound for one pair of (p, q) and then do a union bound over all possible pairs. In the following, let

$$j = \underset{l \in [p-1,q]}{\arg\min}\, n_l, \quad t = \underset{l \in [p-1,q] \setminus \{j\}}{\arg\min}\, n_l.$$

We analyze three cases below. In the first case, namely $j \in [p, q-1]$, then

$$\begin{aligned}\left\|\prod_{l=p}^{q} W_l \Sigma_l(x)\right\|_{\text{op}}^2 &\leq \left\|\prod_{l=p}^{j} W_l \Sigma_l(x)\right\|_{\text{op}}^2 \left\|\prod_{l=j+1}^{q} W_l \Sigma_l(x)\right\|_{\text{op}}^2 = \mathcal{O}\left(\frac{\prod_{l=p-1}^{j} n_l}{\min_{l\in[p-1,j]} n_l} \frac{\prod_{l=j}^{q} n_l}{\min_{l\in[j,q]} n_l} \prod_{l=p}^{q} \beta_l^2\right) \\ &= \mathcal{O}\left(\frac{\prod_{l=p-1}^{q} n_l}{n_j} \prod_{l=p}^{q} \beta_l^2\right) = \mathcal{O}\left(\frac{\prod_{l=p-1}^{q} n_l}{\min_{l\in[p-1,q]} n_l} \prod_{l=p}^{q} \beta_l^2\right),\end{aligned}$$

where the first equality follows from our induction assumption, the second equality follows from the current choice of j. In the second case, if $j = q$ and $t \in [p, q-1]$, then similarly one has

$$\begin{aligned}\left\|\prod_{l=p}^{q} W_l \Sigma_l(x)\right\|_{\text{op}}^2 &\leq \left\|\prod_{l=p}^{t} W_l \Sigma_l(x)\right\|_{\text{op}}^2 \left\|\prod_{l=t+1}^{q} W_l \Sigma_l(x)\right\|_{\text{op}}^2 = \mathcal{O}\left(\frac{\prod_{l=p-1}^{t} n_l}{\min_{l\in[p-1,t]} n_l} \frac{\prod_{l=t}^{q} n_l}{\min_{l\in[t,q]} n_l} \prod_{l=p}^{q} \beta_l^2\right) \\ &= \mathcal{O}\left(\frac{\prod_{l=p-1}^{t} n_l}{n_t} \frac{\prod_{l=t}^{q} n_l}{n_q} \prod_{l=p}^{q} \beta_l^2\right) = \mathcal{O}\left(\frac{\prod_{l=p-1}^{q} n_l}{\min_{l\in[p-1,q]} n_l} \prod_{l=p}^{q} \beta_l^2\right).\end{aligned}$$

It remains to handle the case in which either $(j = p-1)$ or $(j = q$ and $t = p-1)$. To do so, we use an ϵ-net argument. Since $\|\Sigma_q(x)\|_{\text{op}} \leq 1$, it holds that

$$\left\|\prod_{l=p}^{q} W_l \Sigma_l(x)\right\|_{\text{op}}^2 \leq \left\|\left(\prod_{l=p}^{q-1} W_l \Sigma_l(x)\right) W_q\right\|_{\text{op}}^2. \tag{44}$$

Furthermore, by using Lemma 4.4.1 of (Vershynin, 2018),

$$\left\|\left(\prod_{l=p}^{q-1} W_l \Sigma_l(x)\right) W_q\right\|_{\text{op}}^2 \leq 4 \sup_{y \in \mathsf{N}_{1/2}^{p-1}} \left\| \underbrace{y^T \left(\prod_{l=p}^{q-1} W_l \Sigma_l(x)\right)}_{=:z^T} W_q \right\|_2^2, \tag{45}$$

where $\mathsf{N}_{1/2}^{p-1}$ is a $\frac{1}{2}$-net of the unit sphere in $\mathbb{R}^{n_{p-1}}$. Fix $y \in \mathsf{N}_{1/2}^{p-1}$, and let z be defined as above, then clearly z is independent of W_q, and it holds by induction assumption

$$\|z\|_2^2 = \mathcal{O}\left(\frac{\prod_{l=p-1}^{q-1} n_l}{\min_{l\in[p-1,q-1]} n_l} \prod_{l=p}^{q-1} \beta_l^2\right) \tag{46}$$

w.p. at least $1 - \sum_{l=p}^{q-1} \exp\left(-\Omega\left(n_l\right)\right)$ over $(W_l)_{l=1}^{q-1}$. Conditioned on this event of the first $q-1$ layers, let us study concentration bound for $\left\|z^T W_q\right\|_2^2$ where the only randomness is over W_q. Note that $\left\|z^T W_q\right\|_2^2 = \sum_{j=1}^{n_q} \left\langle z, (W_q)_{:j} \right\rangle^2$ and

$\left\|\langle z,(W_q)_{:j}\rangle^2\right\|_{\psi_1} \le c_1\beta_q^2\|z\|_2^2$. Thus by Bernstein's inequality (see Theorem 2.8.1 of (Vershynin, 2018)),

$$\mathbb{P}\left(\left|\left\|z^TW_q\right\|_2^2 - \mathbb{E}_{W_q}\left\|z^TW_q\right\|_2^2\right| > t\right) \le \exp\left(-c_2\min\left(\frac{t}{c_1\beta_q^2\|z\|_2^2}, \frac{t^2}{n_q c_1^2\beta_q^4\|z\|_2^4}\right)\right),$$

for some constant c_2. By plugging $t = Cc_1\max(n_q,n_{p-1})\beta_q^2\|z\|_2^2/c_2$ for some $C > \max(c_2,\log 5)$, and $\mathbb{E}_{W_q}\left\|z^TW_q\right\|_2^2 = n_q\beta_q^2\|z\|_2^2$, one obtains $\left\|z^TW_q\right\|_2^2 = \mathcal{O}\left(\max(n_q,n_{p-1})\beta_q^2\|z\|_2^2\right)$ w.p. at least $1-e^{-C\max(n_q,n_{p-1})}$. Taking the union bound over $y\in\mathsf{N}_{1/2}^{p-1}$, we get

$$\sup_{y\in\mathsf{N}_{1/2}^{p-1}}\left\|z^TW_q\right\|_2^2 = \sup_{y\in\mathsf{N}_{1/2}^{p-1}}\left\|y^T\left(\prod_{l=p}^{q-1}W_l\Sigma_l(x)\right)W_q\right\|_2^2 = \mathcal{O}\left(\max(n_q,n_{p-1})\beta_q^2\|z\|^2\right)$$

w.p. at least $1-\left|\mathsf{N}_{1/2}^{p-1}\right|e^{-C\max(n_q,n_{p-1})} = 1-e^{-\Omega(\max(n_q,n_{p-1}))}$, where we used the fact that $\left|\mathsf{N}_{1/2}^{p-1}\right| \le 5^{n_{p-1}}$ and $C>\log 5$. This combined with (44),(45) and (46) implies

$$\left\|\prod_{l=p}^{q}W_l\Sigma_l(x)\right\|_{\mathrm{op}}^2 = \mathcal{O}\left(\max(n_q,n_{p-1})\beta_q^2\frac{\prod_{l=p-1}^{q-1}n_l}{\min_{l\in[p-1,q-1]}n_l}\prod_{l=p}^{q-1}\beta_l^2\right) = \mathcal{O}\left(\frac{\prod_{l=p-1}^{q}n_l}{\min_{l\in[p-1,q]}n_l}\prod_{l=p}^{q}\beta_l^2\right),$$

where the last estimate follows from the current conditions on (j,t). To summarize, we have shown that (43) holds for every given pair (p,q) such that $q-p=s$. Taking the union bound over all these pairs finishes the proof. Finally, note that doing the union bound above does not affect the probability of the final result since the number of all possible pairs is only a constant.

E. Missing Proofs from Section 5

E.1. Proof of Lemma 5.2

For a subgaussian random variable Z, recall that $\mathbb{P}(Z>t)\le\exp(-ct^2/\|Z\|_{\psi_2}^2)$, where c is an absolute constant. In the following, let $t = \frac{4\beta_k\|F_{k-1}\|_F}{c}\sqrt{\max\left(1,\log\frac{8\beta_k^2\|F_{k-1}\|_F^2}{c\,\lambda}\right)}$. Let us denote the shorthand $W_k=[w_1,\dots,w_{n_k}]\in\mathbb{R}^{n_{k-1}\times n_k}$, and denote by $A\in\mathbb{R}^{N\times n_k}$ a matrix such that $A_{:j} = \sigma(F_{k-1}w_j)\,\mathbb{1}_{\|\sigma(F_{k-1}w_j)\|_2\le t}$ for all $j\in[n_k]$. Let

$$G = \mathbb{E}_{w\sim\mathcal{N}(0,\beta_k^2\mathbb{I}_{n_{k-1}})}\left[\sigma(F_{k-1}w)\sigma(F_{k-1}w)^T\right],$$
$$\hat{G} = \mathbb{E}_{w\sim\mathcal{N}(0,\beta_k^2\mathbb{I}_{n_{k-1}})}\left[\sigma(F_{k-1}w)\sigma(F_{k-1}w)^T\,\mathbb{1}_{\|\sigma(F_{k-1}w)\|_2\le t}\right].$$

Note $\lambda=\lambda_{\min}(G)$, $\lambda_{\min}\left(F_kF_k^T\right)\ge\lambda_{\min}\left(AA^T\right)$ and $\lambda_{\max}\left(A_{:j}A_{:j}^T\right)\le t^2$. By Matrix Chernoff inequality (see Theorem 1.1 of (Tropp, 2012)), it holds for every $\epsilon\in[0,1)$

$$\mathbb{P}\Big(\lambda_{\min}\left(AA^T\right)\le(1-\epsilon)\lambda_{\min}\left(\mathbb{E}AA^T\right)\Big)\le N\left[\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right]^{\lambda_{\min}\left(\mathbb{E}AA^T\right)/t^2}.$$

Pick $\epsilon=1/2$. Then,

$$\mathbb{P}\left(\lambda_{\min}\left(AA^T\right)\le n_k\lambda_{\min}\left(\hat{G}\right)/2\right)\le\exp\left(-c_1\,n_k\lambda_{\min}\left(\hat{G}\right)/t^2+\log N\right).$$

Thus, for $n_k \geq \frac{t^2}{c_1 \lambda_{\min}\left(\hat{G}\right)} \log \frac{N}{\delta}$ we have $\lambda_{\min}\left(AA^T\right) \geq \frac{n_k \lambda_{\min}\left(\hat{G}\right)}{2}$ w.p. $\geq 1-\delta$. Moreover,

$$\begin{aligned}
\left\|\hat{G}-G\right\|_2 &\leq \mathbb{E}\left\|\sigma(F_{k-1}w)\sigma(F_{k-1}w)^T\,\mathbb{1}_{\|\sigma(F_{k-1}w)\|_2\leq t} - \sigma(F_{k-1}w)\sigma(F_{k-1}w)^T\right\|_2 \\
&= \mathbb{E}\left[\|\sigma(F_{k-1}w)\|_2^2\;\mathbb{1}_{\|\sigma(F_{k-1}w)\|_2>t}\right] \\
&= \int_{s=0}^{\infty} \mathbb{P}\left(\|\sigma(F_{k-1}w)\|_2\;\mathbb{1}_{\|\sigma(F_{k-1}w)\|_2>t} > \sqrt{s}\right) ds \\
&= \int_{s=0}^{\infty} \mathbb{P}\left(\|\sigma(F_{k-1}w)\|_2 > t\right) \mathbb{P}\left(\|\sigma(F_{k-1}w)\|_2 > \sqrt{s}\right) ds \\
&\leq \int_{s=0}^{\infty} \exp\left(-c\frac{t^2+s}{4\beta_k^2\left\|F_{k-1}\right\|_F^2}\right) ds \\
&\leq \lambda/2,
\end{aligned}$$

where the second inequality uses the fact that $\left\|\|\sigma(F_{k-1}w)\|_2\right\|_{\psi_2} \leq 2\beta_k \left\|F_{k-1}\right\|_F$. It follows that $\lambda_{\min}\left(\hat{G}\right) \geq \lambda/2$. In total, for $n_k \geq \frac{2t^2}{c_1\lambda}\log\frac{N}{\delta}$, it holds w.p. at least $1-\delta$ that

$$\sigma_{\min}\left(F_k\right)^2 = \lambda_{\min}\left(F_kF_k^T\right) \geq \lambda_{\min}\left(AA^T\right) \geq n_k\lambda_{\min}\left(\hat{G}\right)/2 \geq n_k\lambda/4,$$

where we used the condition $n_k \geq N$ in the above equality.

E.2. Proof of Lemma 5.3

Let $D = \operatorname{diag}(\|(F_k)_{1:}\|_2, \ldots, \|(F_k)_{N:}\|_2)$ and $\hat{F}_k = D^{-1}F_k$. Then, by the homogeneity of σ, we have

$$\begin{aligned}
\lambda_{\min}\left(\mathbb{E}[\sigma(F_kw)\sigma(F_kw)^T]\right) &= \lambda_{\min}\left(D\mathbb{E}\left[\sigma(\hat{F}_kw)\sigma(\hat{F}_kw)^T\right]D\right) \\
&= \beta_{k+1}^2\lambda_{\min}\left(D\left[\mu_0(\sigma)^2 1_N 1_N^T + \sum_{s=1}^{\infty}\mu_s(\sigma)^2(\hat{F}_k^{*s})(\hat{F}_k^{*s})^T\right]D\right) \\
&\geq \beta_{k+1}^2\,\mu_r(\sigma)^2\lambda_{\min}\left(D(\hat{F}_k^{*r})(\hat{F}_k^{*r})^TD\right) \\
&= \beta_{k+1}^2\,\mu_r(\sigma)^2\lambda_{\min}\left(D^{-(r-1)}(F_k^{*r})(F_k^{*r})^TD^{-(r-1)}\right) \\
&\geq \beta_{k+1}^2\,\mu_r(\sigma)^2\frac{\lambda_{\min}\left((F_k^{*r})(F_k^{*r})^T\right)}{\max_{i\in[N]}\|(F_k)_{i:}\|_2^{2(r-1)}},
\end{aligned}$$

where the second equality uses the Hermite expansion of σ (for the proof see Lemma D.3 of (Nguyen & Mondelli, 2020)).

E.3. Proof of Lemma 5.4

We have that

$$(F_k^{*r})(F_k^{*r})^T = \left(F_kF_k^T\right)^{\circ r}. \tag{47}$$

After some manipulations, we obtain

$$\begin{aligned}
F_kF_k^T &= \tilde{F}_k\tilde{F}_k^T + \|\mu\|_2^2\,1_N1_N^T + \Lambda 1_N1_N^T + 1_N1_N^T\Lambda \\
&= \tilde{F}_k\tilde{F}_k^T + \left(\|\mu\|_2\,1_N + \frac{\Lambda 1_N}{\|\mu\|_2}\right)\left(\|\mu\|_2\,1_N + \frac{\Lambda 1_N}{\|\mu\|_2}\right)^T - \frac{\Lambda 1_N1_N^T\Lambda}{\|\mu\|_2^2} \\
&\succeq \tilde{F}_k\tilde{F}_k^T - \frac{\Lambda 1_N1_N^T\Lambda}{\|\mu\|_2^2},
\end{aligned} \tag{48}$$

where in the last passage we use that $\left(\|\mu\|_2 1_N + \frac{\Lambda 1_N}{\|\mu\|_2}\right)\left(\|\mu\|_2 1_N + \frac{\Lambda 1_N}{\|\mu\|_2}\right)^T$ is PSD. Note that the RHS of (48) is also PSD since

$$\tilde{F}_k \tilde{F}_k^T \succeq \frac{\tilde{F}_k \mu\mu^T \tilde{F}_k^T}{\|\mu\|_2^2} = \frac{\Lambda 1_N 1_N^T \Lambda}{\|\mu\|_2^2}.$$

Hence, the r-th Hadamard power of the LHS of (48) is an upper bound (in the PSD sense) of the r-th Hadamard power of the RHS of (48), which concludes the proof.

E.4. Proof of Lemma 5.5

Note that

$$\left(\tilde{F}_k \tilde{F}_k^T - \frac{\Lambda 1_N 1_N^T \Lambda}{\|\mu\|_2^2}\right)^{\circ r} = \left(\tilde{F}_k \tilde{F}_k^T\right)^{\circ r} + \sum_{i=1}^{r} \binom{r}{i} (-1)^i \frac{\Lambda^i \left(\tilde{F}_k \tilde{F}_k^T\right)^{\circ(r-i)} \Lambda^i}{\|\mu\|_2^{2i}}. \tag{49}$$

Thus, an application of Weyl's inequality gives

$$\lambda_{\min}\left(\left(\tilde{F}_k \tilde{F}_k^T - \frac{\Lambda 1_N 1_N^T \Lambda}{\|\mu\|_2^2}\right)^{\circ r}\right) \geq \lambda_{\min}\left(\left(\tilde{F}_k \tilde{F}_k^T\right)^{\circ r}\right) - \left\| \sum_{i=1}^{r} \binom{r}{i} (-1)^i \frac{\Lambda^i \left(\tilde{F}_k \tilde{F}_k^T\right)^{\circ(r-i)} \Lambda^i}{\|\mu\|_2^{2i}} \right\|_{\text{op}}. \tag{50}$$

We start by bounding the term $\lambda_{\min}\left(\left(\tilde{F}_k \tilde{F}_k^T\right)^{\circ r}\right)$. From Gershgorin circle theorem, one obtains

$$\lambda_{\min}\left((\tilde{F}_k^{*r})(\tilde{F}_k^{*r})^T\right) \geq \min_{i\in[N]} \|(\tilde{F}_k)_{i:}\|_2^{2r} - N \max_{i\neq j} |\langle (\tilde{F}_k)_{i:}, (\tilde{F}_k)_{j:}\rangle|^r, \tag{51}$$

$$\lambda_{\min}\left((\tilde{F}_k^{*r})(\tilde{F}_k^{*r})^T\right) \leq \max_{i\in[N]} \|(\tilde{F}_k)_{i:}\|_2^{2r} + N \max_{i\neq j} |\langle (\tilde{F}_k)_{i:}, (\tilde{F}_k)_{j:}\rangle|^r. \tag{52}$$

By Lemma C.3, it holds w.p. at least $1 - N\exp\left(-\Omega\left(\frac{\min_{l\in[0,k]} n_l}{\prod_{l=1}^{k-1}\log(n_l)}\right)\right) - \sum_{l=1}^{k}\exp(-\Omega(n_l))$ that

$$\|(\tilde{F}_k)_{i:}\|_2^{2r} = \Theta\left(\left(d\prod_{l=1}^{k} n_l \beta_l^2\right)^r\right), \quad \forall\, i \in [N]. \tag{53}$$

In the following, we bound the second term on the RHS of (52). For a fixed $j \in [N]$, Lemma C.3 implies that w.p. at least $1 - \exp\left(-\Omega\left(\frac{\min_{l\in[0,k]} n_l}{\prod_{l=1}^{k-1}\log(n_l)}\right)\right) - \sum_{l=1}^{k}\exp(-\Omega(n_l))$ over $(W_l)_{l=1}^k$ and x_j, we have

$$\left\|(\tilde{F}_k)_{j:}\right\|_2^2 = \Theta\left(d\prod_{l=1}^{k} n_l \beta_l^2\right). \tag{54}$$

Moreover, Theorem 6.2 implies that w.p. at least $1 - \sum_{l=1}^{k}\exp(-\Omega(n_l))$ over $(W_l)_{l=1}^k$,

$$\|f_k(x) - \mathbb{E}_x f_k(x)\|_{\text{Lip}}^2 = \mathcal{O}\left(\frac{\prod_{l=0}^{k} n_l}{\min_{l\in[0,k]} n_l} \prod_{l=1}^{k-1}\log(n_l) \prod_{l=1}^{k}\beta_l^2\right). \tag{55}$$

Let us condition on the intersection of these two events of $(W_l)_{l=1}^k$ and x_j, and derive probability bounds over x_i, for every $i \neq j$. Let $h(x_i) = \left\langle (\tilde{F}_k)_{i:}, (\tilde{F}_k)_{j:}\right\rangle$ be a function of x_i, then

$$\|h\|_{\text{Lip}}^2 \leq \left\|(\tilde{F}_k)_{j:}\right\|_2^2 \|f_k(x_i) - \mathbb{E}_x f_k(x_i)\|_{\text{Lip}}^2 = \mathcal{O}\left(\left(d\prod_{l=1}^{k} n_l\beta_l^2\right)^2 \frac{\prod_{l=1}^{k-1}\log(n_l)}{\min_{l\in[0,k]} n_l}\right),$$

where the last estimate follows from (54) and (55). Using Assumption 2.2, followed by a union bound over $\{x_i\}_{i\neq j}$, we have for every $t>0$ that

$$\mathbb{P}\left(\max_{i\in[N],i\neq j}\left|\left\langle(\tilde{F}_k)_{i:},(\tilde{F}_k)_{j:}\right\rangle\right|\geq t\right)\leq(N-1)\exp\left(-\frac{t^2}{\mathcal{O}\left(\left(d\prod_{l=1}^k n_l\beta_l^2\right)^2\frac{\prod_{l=1}^{k-1}\log(n_l)}{\min_{l\in[0,k]}n_l}\right)}\right). \tag{56}$$

Pick $t=N^{-1/(r-0.1)}\left(d\prod_{l=1}^k n_l\beta_l^2\right)$. Then, taking the intersection bound with (54) and (55) yields

$$N\max_{i\in[N],i\neq j}|\langle(\tilde{F}_k)_{i:},(\tilde{F}_k)_{j:}\rangle|^r\leq N\frac{\left(d\prod_{l=1}^k n_l\beta_l^2\right)^r}{N^{r/(r-0.1)}}=o\left(\left(d\prod_{l=1}^k n_l\beta_l^2\right)^r\right) \tag{57}$$

w.p. at least

$$1-(N-1)\exp\left(-\Omega\left(\frac{\min_{l\in[0,k]}n_l}{N^{2/(r-0.1)}\prod_{l=1}^{k-1}\log(n_l)}\right)\right)-\sum_{l=1}^k\exp\left(-\Omega\left(n_l\right)\right).$$

Since this holds for every given x_j, taking the union bound over $j\in[N]$ yields that

$$N\max_{i\neq j}|\langle(\tilde{F}_k)_{i:},(\tilde{F}_k)_{j:}\rangle|^r=o\left(\left(d\prod_{l=1}^k n_l\beta_l^2\right)^r\right) \tag{58}$$

w.p. at least

$$1-N^2\exp\left(-\Omega\left(\frac{\min_{l\in[0,k]}n_l}{N^{2/(r-0.1)}\prod_{l=1}^{k-1}\log(n_l)}\right)\right)-N\sum_{l=1}^k\exp\left(-\Omega\left(n_l\right)\right). \tag{59}$$

Combining (51), (52), (53), (58) gives that, with probability lower bounded by (59),

$$\lambda_{\min}\left(\left(\tilde{F}_k\tilde{F}_k^T\right)^{\circ r}\right)=\Theta\left(\left(d\prod_{l=1}^k n_l\beta_l^2\right)^r\right). \tag{60}$$

By applying again Gershgorin circle theorem and following similar passages, we also have that

$$\max_{i\in\{1,\dots,r/2\}}\left\|\left(\tilde{F}_k\tilde{F}_k^T\right)^{\circ(r-i)}\right\|_{\text{op}}=\mathcal{O}\left(\left(d\prod_{l=1}^k n_l\beta_l^2\right)^{r-i}\right) \tag{61}$$

w.p. at least

$$1-N^2\exp\left(-\Omega\left(\frac{\min_{l\in[0,k]}n_l}{N^{2/(r/2-0.1)}\prod_{l=1}^{k-1}\log(n_l)}\right)\right)-N\sum_{l=1}^k\exp\left(-\Omega\left(n_l\right)\right). \tag{62}$$

Furthermore, by using (54) and that the Frobenius norm upper bounds the operator norm, we also obtain the following simple bound

$$\left\|\left(\tilde{F}_k\tilde{F}_k^T\right)^{\circ(r-i)}\right\|_{\text{op}}\leq\left\|\left(\tilde{F}_k\tilde{F}_k^T\right)^{\circ(r-i)}\right\|_{F}\leq\mathcal{O}\left(N\left(d\prod_{l=1}^k n_l\beta_l^2\right)^{r-i}\right) \tag{63}$$

holding w.p. at least $1-N\exp\left(-\Omega\left(\frac{\min_{l\in[0,k]}n_l}{\prod_{l=1}^{k-1}\log(n_l)}\right)\right)-\sum_{l=1}^k\exp(-\Omega\left(n_l\right))$.

Next, we upper bound the operator norm in the RHS of (50). As the operator norm is sub-additive and sub-multiplicative, we have that

$$\begin{aligned}\left\|\sum_{i=1}^{r}\binom{r}{i}(-1)^i\frac{\Lambda^i\left(\tilde{F}_k\tilde{F}_k^T\right)^{\circ(r-i)}\Lambda^i}{\|\mu\|_2^{2i}}\right\|_{\mathrm{op}} &\leq \sum_{i=1}^{r}\binom{r}{i}\left\|\frac{\Lambda}{\|\mu\|_2}\right\|_{\mathrm{op}}^{2i}\left\|\left(\tilde{F}_k\tilde{F}_k^T\right)^{\circ(r-i)}\right\|_{\mathrm{op}} \\ &= \sum_{i=1}^{r/2}\binom{r}{i}\left\|\frac{\Lambda}{\|\mu\|_2}\right\|_{\mathrm{op}}^{2i}\left\|\left(\tilde{F}_k\tilde{F}_k^T\right)^{\circ(r-i)}\right\|_{\mathrm{op}} + \sum_{i=r/2+1}^{r}\binom{r}{i}\left\|\frac{\Lambda}{\|\mu\|_2}\right\|_{\mathrm{op}}^{2i}\left\|\left(\tilde{F}_k\tilde{F}_k^T\right)^{\circ(r-i)}\right\|_{\mathrm{op}} := S_1+S_2.\end{aligned} \tag{64}$$

Let $h:\mathbb{R}^d\to\mathbb{R}$ be a function over a random sample x, defined as $h(x)=\langle f_k(x),\mu\rangle$. Then, $\Lambda_{ii}=h(x_i)-\mathbb{E}_x[h(x)]$. Since $\|h\|_{\mathrm{Lip}}^2\leq\|\mu\|_2^2\|f_k\|_{\mathrm{Lip}}^2$, it holds

$$\mathbb{P}\left(|\Lambda_{ii}|\geq t\right)\leq\exp\left(-\frac{t^2}{2\|\mu\|_2^2\|f_k\|_{\mathrm{Lip}}^2}\right). \tag{65}$$

By Lemma C.2, it holds w.p. at least $1-\sum_{l=1}^{k}\exp\left(-\Omega\left(n_l\right)\right)$ over $(W_l)_{l=1}^k$ that

$$\|\mu\|_2^2=\Theta\left(d\prod_{l=1}^{k}n_l\beta_l^2\right). \tag{66}$$

Also, Theorem 6.2 shows that w.p. at least $1-\sum_{l=1}^{k}\exp\left(-\Omega\left(n_l\right)\right)$ over $(W_l)_{l=1}^k$,

$$\|f_k\|_{\mathrm{Lip}}^2=\mathcal{O}\left(\frac{\prod_{l=0}^{k}n_l}{\min_{l\in[0,k]}n_l}\prod_{l=1}^{k-1}\log(n_l)\prod_{l=1}^{k}\beta_l^2\right). \tag{67}$$

Now, pick $t=\|\mu\|_2^2N^{-1/(r/2-0.1)}$ in (65). Then, taking the union bound over all the samples and over the events in (66) and (67), we conclude that

$$\left\|\frac{\Lambda}{\|\mu\|_2}\right\|_{\mathrm{op}}^2=\mathcal{O}\left(N^{-2/(r/2-0.1)}d\prod_{l=1}^{k}n_l\beta_l^2\right) \tag{68}$$

w.p. at least

$$1-N\exp\left(-\Omega\left(\frac{\min_{l\in[0,k]}n_l}{N^{2/(r/2-0.1)}\prod_{l=1}^{k-1}\log(n_l)}\right)\right)-\sum_{l=1}^{k}\exp\left(-\Omega\left(n_l\right)\right).$$

By combining (61) and (68), we have that

$$S_1\leq\mathcal{O}\left(N^{-2/(r/2-0.1)}\left(d\prod_{l=1}^{k}n_l\beta_l^2\right)^r\right)=o\left(\left(d\prod_{l=1}^{k}n_l\beta_l^2\right)^r\right) \tag{69}$$

w.p. lower bounded by (62). Furthermore, by combining (63) and (68), we have that

$$S_2\leq\mathcal{O}\left(N^{1-r/(r/2-0.1)}\left(d\prod_{l=1}^{k}n_l\beta_l^2\right)^r\right)=o\left(\left(d\prod_{l=1}^{k}n_l\beta_l^2\right)^r\right) \tag{70}$$

w.p. lower bounded by (62). By combining (50), (60), (64), (69) and (70), the desired result (31) follows.

F. Missing Proofs from Section 6

Definition F.1 *A subset $A \subseteq \mathbb{R}^n$ is called a polyhedron if it is the intersection of a finite family of (closed) half-spaces. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called piecewise linear if there exist a finite family of polyhedra $\{P_i\}_{i=1}^r$ such that $\mathbb{R}^n = \cup_{i=1}^r P_i$ and f coincides with a linear function on each P_i.*

The following lemma establishes a formal connection between ReLU networks and PWL functions. Its proof is contained in Appendix F.3.

Lemma F.2 *For every $k \in [L]$, $f_k, g_k : \mathbb{R}^d \to \mathbb{R}^{n_k}$ as defined in (1) are piecewise linear functions.*

An equivalent way of defining piecewise linear maps is the following, see e.g. (Gorokhovik, 2011).

Lemma F.3 *A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is piecewise linear if and only if there exist a finite family of polyhedra $\{P_i\}_{i=1}^T$ and matrices $\{A_i\}_{i=1}^T \in \mathbb{R}^{m\times n}$ such that:*

1. $\mathbb{R}^n = \bigcup_{i=1}^T P_i$,
2. $\operatorname{int}(P_i) \neq \emptyset, \quad \forall\, i \in [T]$,
3. $\operatorname{int}(P_i) \cap \operatorname{int}(P_j) = \emptyset \quad \forall\, i \neq j$,
4. $f(x) = A_i x$ *for every* $x \in P_i$.

F.1. Proof of Theorem 6.2

Let $h_{p\to q} : \mathbb{R}^{n_p} \to \mathbb{R}^{n_q}$ be defined as

$$h_{p\to q} = A_q \circ \hat{\sigma}_{q-1} \circ A_{q-1} \circ \ldots \circ \hat{\sigma}_{p+1} \circ A_{p+1},$$

where the mapping $A_l : \mathbb{R}^{n_{l-1}} \to \mathbb{R}^{n_l}$ is given by $A_l(x) = W_l^T x$, and the mapping $\hat{\sigma}_l : \mathbb{R}^{n_l} \to \mathbb{R}^{n_l}$ is given by $\hat{\sigma}(x) = [\sigma(x_1), \ldots, \sigma(x_{n_l})]^T$ for every $x \in \mathbb{R}^{n_l}$. By definition, it holds $g_k(x) = h_{0\to k}(x)$. In the following, we prove that for every $0 \leq p < q \leq L$, it holds w.p. $\geq 1 - \sum_{l=p-1}^{q} \exp\left(-\Omega\left(n_l\right)\right)$ that

$$\|h_{p\to q}\|_{\mathrm{Lip}} = \mathcal{O}\left(\frac{\prod_{l=p}^{q} n_l}{\min_{l\in[p,q]} n_l} \prod_{l=p+1}^{q-1} \log(n_l) \prod_{l=p+1}^{q} \beta_l^2\right). \tag{71}$$

The desired result follows by letting $p = 0, q = k$. The proof of (71) is by induction over the length $s = q - p$. First, (71) holds for $s = 1$. Suppose that (71) holds for all (p, q) such that $q - p \leq s - 1$, and we want to prove it for all (p, q) with $q - p = s$. It suffices to show the result for one pair and then do a union bound over all the possible pairs. Let us define

$$j = \operatorname*{arg\,min}_{l\in[p,q]} n_l, \quad t = \operatorname*{arg\,min}_{l\in[p,q]\setminus\{j\}} n_l.$$

Consider three cases below. In the first case, $j \in [p + 1, q - 1]$. By noting that

$$h_{p\to q} = h_{j\to q} \circ \hat{\sigma}_j \circ h_{p\to j}$$

and using the Lipschitz property of a composition of Lipschitz continuous functions, one obtains

$$\begin{aligned}
\|h_{p\to q}\|_{\mathrm{Lip}} &\leq \|h_{p\to j}\|_{\mathrm{Lip}} \|\hat{\sigma}_j\|_{\mathrm{Lip}} \|h_{j\to q}\|_{\mathrm{Lip}} \\
&= \mathcal{O}\left(\frac{\prod_{l=p}^{j} n_l}{\min_{l\in[p,j]} n_l} \prod_{l=p+1}^{j-1} \log(n_l) \frac{\prod_{l=j}^{q} n_l}{\min_{l\in[j,q]} n_l} \prod_{l=j+1}^{q-1} \log(n_l) \prod_{l=p+1}^{q} \beta_l^2\right) \\
&= \mathcal{O}\left(\frac{\prod_{l=p}^{q} n_l}{\min_{l\in[p,q]} n_l} \prod_{l=p+1}^{q-1} \log(n_l) \prod_{l=p+1}^{q} \beta_l^2\right),
\end{aligned}$$

where the first equality follows from induction assumption and $\|\hat{\sigma}\|_{\text{Lip}} \leq 1$, the second equality follows from definition of j. In the second case, $j = q$ and $t \in [p+1, q-1]$, then similarly,

$$\begin{aligned}
\|h_{p\to q}\|_{\text{Lip}} &\leq \|h_{p\to t}\|_{\text{Lip}} \|\hat{\sigma}_t\|_{\text{Lip}} \|h_{t\to q}\|_{\text{Lip}} \\
&= \mathcal{O}\left(\frac{\prod_{l=p}^{t} n_l}{\min_{l\in[p,t]} n_l} \prod_{l=p+1}^{t-1} \log(n_l) \frac{\prod_{l=t}^{q} n_l}{\min_{l\in[t,q]} n_l} \prod_{l=t+1}^{q-1} \log(n_l) \prod_{l=p+1}^{q} \beta_l^2 \right) \\
&= \mathcal{O}\left(\frac{n_t \prod_{l=p}^{q} n_l}{n_t n_q} \prod_{l=p+1}^{q-1} \log(n_l) \prod_{l=p+1}^{q} \beta_l^2 \right) \\
&= \mathcal{O}\left(\frac{\prod_{l=p}^{q} n_l}{\min_{l\in[p,q]} n_l} \prod_{l=p+1}^{q-1} \log(n_l) \prod_{l=p+1}^{q} \beta_l^2 \right).
\end{aligned}$$

It remains to handle the case where either $(j = p)$ or $(j = q$ and $t = p)$. By Lemma 6.1, it holds w.p. 1 over $(W_l)_{l=p+1}^{q-1}$ that there exists a set of R tuples of diagonal matrices, say $\mathcal{D} = \left\{(\Sigma_{p+1}^1, \ldots, \Sigma_{q-1}^1), \ldots, (\Sigma_{p+1}^R, \ldots, \Sigma_{q-1}^R)\right\}$, with 0-1 entries on the diagonals such that

$$\|h_{p\to q}\|_{\text{Lip}} \leq \max_{(\Sigma_{p+1},\ldots,\Sigma_{q-1})\in\mathcal{D}} \left\| \left(\prod_{l=p+1}^{q-1} W_l \Sigma_l \right) W_q \right\|_{\text{op}}. \tag{72}$$

According to Lemma 6.1, R can be interpreted as the maximum number of activation patterns of a $q-p$ layer network with layer widths $(n_p, n_{p+1}, \ldots, n_q)$, where every hidden neuron has a definite sign pattern $\{-1, +1\}$. Let $n_{\max} = \max_{l\in[p+1,q-1]} n_l$, then $R = \mathcal{O}\left((n_{\max})^{n_p}\right)$ (see e.g. (Hanin & Rolnick, 2019; Serra et al., 2018)). Using the definition of operator norm and an ϵ-net argument, the inequality (72) becomes

$$\begin{aligned}
\|h_{p\to q}\|_{\text{Lip}} &\leq \max_{(\Sigma_{p+1},\ldots,\Sigma_{q-1})\in\mathcal{D}} \sup_{\|y\|_2=1} \left\| y^T \left(\prod_{l=p+1}^{q-1} W_l \Sigma_l \right) W_q \right\|_2 \\
&\leq \max_{(\Sigma_{p+1},\ldots,\Sigma_{q-1})\in\mathcal{D}} 2 \sup_{y\in\mathsf{N}_{1/2}^p} \left\| \underbrace{y^T \left(\prod_{l=p+1}^{q-1} W_l \Sigma_l \right)}_{=:z^T} W_q \right\|_2^2,
\end{aligned} \tag{73}$$

where $\mathsf{N}_{1/2}^p$ is a $\frac{1}{2}$-net of the unit sphere in $\mathbb{R}^{n_p}$ and the last inequality follows from Lemma 4.4.1 in (Vershynin, 2018). Fix $y \in \mathsf{N}_{1/2}^p$, and let z be defined as above. Note that z is independent of W_q. From the proof of Lemma 4.3, we have

$$\|z\|_2^2 \leq \left\| \prod_{l=p+1}^{q-1} W_l \Sigma_l \right\|_{\text{op}}^2 = \mathcal{O}\left(\frac{\prod_{l=p}^{q-1} n_l}{\min_{l\in[p,q-1]} n_l} \prod_{l=p+1}^{q-1} \beta_l^2 \right) \tag{74}$$

w.p. at least $1 - \sum_{l=p}^{q-1} \exp\left(-\Omega\left(n_l\right)\right)$ over $(W_l)_{l=p+1}^{q-1}$. Conditioned on the intersection of this event with the event (72) of $(W_l)_{l=p+1}^{q-1}$, let us now study a concentration bound for $\left\|z^T W_q\right\|_2^2$ where the only randomness is W_q. We have $\left\|z^T W_q\right\|_2^2 = \sum_{j=1}^{n_q} \langle z, (W_q)_{:j}\rangle^2$ and $\left\|\langle z, (W_q)_{:j}\rangle^2\right\|_{\psi_1} \leq c_1 \beta_q^2 \|z\|_2^2$. Thus by Bernstein's inequality (see Theorem 2.8.1 of (Vershynin, 2018)),

$$\mathbb{P}\left(\left| \left\|z^T W_q\right\|_2^2 - \mathbb{E}_{W_q} \left\|z^T W_q\right\|_2^2 \right| > t \right) \leq \exp\left(-c_2 \min\left(\frac{t}{c_1 \beta_q^2 \|z\|_2^2}, \frac{t^2}{n_q c_1^2 \beta_q^4 \|z\|_2^4} \right) \right),$$

for some constant c_2. Let $C = \max(c_2, 2)$. Then by substituting to the above inequality the values

$$t = \frac{Cc_1}{c_2}\max(n_q, n_p)\frac{\log(R)}{n_p}\beta_q^2 \left\|z\right\|_2^2, \quad \mathbb{E}_{W_q}\left\|z^T W_q\right\|_2^2 = n_q\beta_q^2\left\|z\right\|_2^2,$$

we have w.p. at least $1 - e^{-C\max(n_q,n_p)\log(R)/n_p}$ that

$$\left\|z^T W_q\right\|_2^2 = \mathcal{O}\left(\max(n_q, n_p)\frac{\log(R)}{n_p}\beta_q^2\left\|z\right\|_2^2\right).$$

Now taking the union bound over $y \in \mathsf{N}_{1/2}^p$ and all tuples from $\mathcal{D}$, the RHS of (73) is bounded as

$$\begin{aligned}\max_{(\Sigma_{p+1},\ldots,\Sigma_{q-1})\in\mathcal{D}} \; 2\sup_{y\in\mathsf{N}_{1/2}^p}\left\|z^T W_q\right\|_2^2 &= \mathcal{O}\left(\max(n_q, n_p)\frac{\log(R)}{n_p}\beta_q^2\left\|z\right\|_2^2\right)\\ &= \mathcal{O}\left(\max(n_q, n_p)\log(n_{\max})\beta_q^2\left\|z\right\|_2^2\right)\end{aligned}$$

w.p. at least

$$1 - R\left|\mathsf{N}_{1/2}^p\right| e^{-C\max(n_q,n_p)\frac{\log(R)}{n_p}} \geq 1 - e^{-\Omega(\max(n_q,n_p))},$$

where we used $\left|\mathsf{N}_{1/2}^p\right| \leq 5^{n_p}$, $R = \mathcal{O}\left((n_{\max})^{n_p}\right)$ and $C > 1$. This combined with (73), (74) implies

$$\begin{aligned}\left\|h_{p\to q}\right\|_{\mathrm{Lip}} &= \mathcal{O}\left(\max(n_q, n_p)\log(n_{\max})\beta_q^2\frac{\prod_{l=p}^{q-1} n_l}{\min_{l\in[p,q-1]} n_l}\prod_{l=p+1}^{q-1}\beta_l^2\right)\\ &= \mathcal{O}\left(\frac{\prod_{l=p}^{q} n_l}{\min_{l\in[p,q]} n_l}\log(\max_{l\in[p+1,q-1]} n_l)\prod_{l=p+1}^{q}\beta_l^2\right)\\ &= \mathcal{O}\left(\frac{\prod_{l=p}^{q} n_l}{\min_{l\in[p,q]} n_l}\prod_{l=p+1}^{q-1}\log(n_l)\prod_{l=p+1}^{q}\beta_l^2\right),\end{aligned}$$

where the second estimate follows from the current value of (j, t). So, we have shown that (71) holds for every pair (p, q) with $q - p = s$. Taking the union bound over all these pairs finishes the proof. Note that this last step does not affect the final probability as the number of pairs is only a constant.

F.2. Proof of Lemma 6.1

Let γ_d be the Lebesgue measure in $\mathbb{R}^d$. Let us associate to $g_k : \mathbb{R}^d \to \mathbb{R}^{n_k}$ a set of polyhedra $\{P_i\}_{i=1}^T$ and matrices $\{A_i\}_{i=1}^T \in \mathbb{R}^{n_k\times n_d}$ as in Lemma F.3. First, let us show that

$$\left\|g_k\right\|_{\mathrm{Lip}} = \max_{i\in[T]}\left\|A_i\right\|_{\mathrm{op}}. \tag{75}$$

Pick any $x, y \in \mathbb{R}^d$. By intersecting the line segment $[x, y]$ with the polyhedra, there exists a finite set of points $\{u_i\}_{i=1}^r$ on $[x, y]$ such that: *(i)* $u_0 = x, u_r = y$, *(ii)* $\left\|x - y\right\|_2 = \sum_{i=0}^{r-1}\left\|u_i - u_{i+1}\right\|_2$, and *(iii)* $[u_i, u_{i+1}]$ is contained in P_{j_i} for some $j_i \in [T]$. This implies

$$\begin{aligned}\left\|g_k(x) - g_k(y)\right\|_2 \leq \sum_{i=0}^{r-1}\left\|g_k(u_i) - g_k(u_{i+1})\right\|_2 = \sum_{i=0}^{r-1}\left\|A_{j_i}(u_i - u_{i+1})\right\|_2 &\leq \sum_{i=0}^{r-1}\left\|A_{j_i}\right\|_{\mathrm{op}}\left\|u_i - u_{i+1}\right\|_2\\ &\leq \max_{i\in[T]}\left\|A_i\right\|_{\mathrm{op}}\left\|x - y\right\|_2,\end{aligned}$$

which means

$$\left\|g_k\right\|_{\mathrm{Lip}} = \sup_{x,y}\frac{\left\|g_k(x) - g_k(y)\right\|_2}{\left\|x - y\right\|_2} \leq \max_{i\in[T]}\left\|A\right\|_{\mathrm{op}}.$$

To show that the above inequality can be attained, let $i_* = \arg\max_{i\in[T]} \|A_i\|_{\mathrm{op}}$. Since $\mathrm{int}(P_{i_*}) \neq \emptyset$, it holds

$$\left\{ \frac{x-y}{\|x-y\|_2} \;\middle|\; x, y \in P_{i_*} \right\} = \mathcal{S}^{n-1},$$

where $\mathcal{S}^{n-1}$ denotes the unit sphere in $\mathbb{R}^n$, and thus

$$\sup_{x,y} \frac{\|g_k(x) - g_k(y)\|_2}{\|x-y\|_2} \geq \sup_{x,y\in P_{i_*}} \frac{\|g_k(x) - g_k(y)\|_2}{\|x-y\|_2} = \sup_{x,y\in P_{i_*}} \frac{\|A_{i_*}(x-y)\|_2}{\|x-y\|_2} = \|A_{i_*}\|_{\mathrm{op}}.$$

This proves the equation (75). Next, let us define the following sets:

$$\begin{aligned} S &= \left\{ x \in \mathbb{R}^d \;\middle|\; f_{k-1}(x) = 0 \right\}, \\ B &= \left\{ x \in \mathbb{R}^d \setminus S \;\middle|\; \exists\, l \in [k-1], i_l \in [n_l] : g_{l,i_l}(x) = 0 \right\}, \\ G &= \mathbb{R}^d \setminus (B \cup S). \end{aligned}$$

Let $\partial S = S \setminus \mathrm{int}(S)$. Then clearly, $\mathbb{R}^d = G \cup B \cup \partial S \cup \mathrm{int}(S)$. Let us show that $\gamma_d(B) = \gamma_d(\partial S) = 0$. By Lemma F.2, f_{k-1} is a PWL function, thus every level set of f_{k-1} can be written as a union of finitely many polyhedra in $\mathbb{R}^d$. This means that ∂S is a union of finitely many polyhedra with dimension at most $d-1$, thus $\gamma_d(\partial S) = 0$. Concerning the set B, note that for every $l \in [k-1], i_l \in [n_l]$,

$$g_{l,i_l}(x) = \sum_{i_0=1}^{d} \sum_{i_1=1}^{n_1} \cdots \sum_{i_{l-1}=1}^{n_{l-1}} \prod_{p=1}^{l} x_{i_0} (W_p)_{i_{p-1},i_p} \prod_{q=1}^{l-1} \mathbb{1}_{g_{q,i_q}(x)>0}.$$

By definition, any $x \in B$ satisfies $f_l(x) \neq 0$ for all $l \in [k-1]$. This implies that at each layer $q \in [k-1]$, there exists at least one active neuron, i.e. some $i_q \in [n_q]$ such that $g_{q,i_q}(x) > 0$. Let $\mathcal{I}_l$ denote the set of active neurons that an input $x \in B$ may have at layer $l \in [k-1]$. Then it holds

$$B \subseteq \bigcup_{l\in[k-1]} \bigcup_{i_l\in[n_l]} \bigcup_{\substack{\mathcal{I}_1\subseteq[n_1]\\ \mathcal{I}_1\neq\emptyset}} \cdots \bigcup_{\substack{\mathcal{I}_{l-1}\subseteq[n_{l-1}]\\ \mathcal{I}_{l-1}\neq\emptyset}} \left\{ x \in \mathbb{R}^d \;\middle|\; \sum_{i_0=1}^{d} \sum_{i_1\in\mathcal{I}_1} \cdots \sum_{i_{l-1}\in\mathcal{I}_{l-1}} \prod_{p=1}^{l} x_{i_0} (W_p)_{i_{p-1},i_p} = 0 \right\}.$$

With probability 1 over $(W_l)_{l=1}^{k-1}$, the set of zeros of each polynomial inside the bracket above has measure zero. Since there are only finitely many such polynomials, one obtains $\gamma_d(B) = 0$.

We are now ready to prove the lemma. From $\mathrm{int}(P_i) \neq \emptyset$ and $\gamma_d(B \cup \partial S) = 0$, it follows that

$$\mathrm{int}(P_i) \cap (G \cup \mathrm{int}(S)) = \mathrm{int}(P_i) \cap (\mathbb{R}^d \setminus (B \cup \partial S)) \neq \emptyset.$$

For every $i \in [T]$, let $z_i \in \mathrm{int}(P_i) \cap (G \cup \mathrm{int}(S))$. Since $z_i \in \mathrm{int}(P_i)$, it follows from (75) that

$$\|g_k\|_{\mathrm{Lip}} = \max_{i\in[T]} \|A_i\|_{\mathrm{op}} = \max_{i\in[T]} \|J(g_k)(z_i)\|_{\mathrm{op}}.$$

Now if $z_i \in \mathrm{int}(S)$, then $J(g_k)(z_i) = 0$, as g_k is constant zero in a neighborhood of z_i. Otherwise, we must have $z_i \in G$, which implies $\mathcal{A}_{1\to k-1}(z_i) \in \{-1,+1\}^{\sum_{l=1}^{k-1} n_l}$. Combining all these facts, we get

$$\|g_k\|_{\mathrm{Lip}} = \max_{z:\ \mathcal{A}_{1\to k-1}(z)\in\{-1,+1\}^{\sum_{l=1}^{k-1} n_l}} \|J(g_k)(z)\|_{\mathrm{op}}.$$

Finally, the inequality $\|f_k\|_{\mathrm{Lip}} \leq \|g_k\|_{\mathrm{Lip}}$ follows from the 1-Lipschitz property of ReLU.

F.3. Proof of Lemma F.2

Let $T = 2^{\sum_{l=1}^{k} n_l}$, and $\{\mathcal{A}_1, \ldots, \mathcal{A}_T\} \in \{-1,+1\}^{\sum_{l=1}^{k} n_l}$ denote the set of all possible binary strings of dimension $\sum_{l=1}^{k} n_l$, where each entry takes value -1 or $+1$. Let us index the entries of each string by $\mathcal{A}_j = \{\mathcal{A}_{j,l,i_l}\}_{l\in[k],i_l\in[n_l]}$. Let

$P_j \subseteq \mathbb{R}^d$ be the set of inputs where the activation pattern of all neurons up to layer k matches perfectly with $\mathcal{A}_j$, namely

$$\begin{aligned} P_j &= \bigcap_{l\in[k]} \bigcap_{i_l\in[n_l]} \left\{x \in \mathbb{R}^d \;\middle|\; g_{l,i_l}(x)\mathcal{A}_{j,l,i_l} \geq 0\right\} \\ &= \bigcap_{l\in[k]} \bigcap_{i_l\in[n_l]} \left\{ x \in \mathbb{R}^d \;\middle|\; \sum_{i_0=1}^{d}\sum_{i_1=1}^{n_1}\cdots\sum_{i_{l-1}=1}^{n_{l-1}}\prod_{p=1}^{l} x_{i_0}(W_p)_{i_{p-1},i_p}\prod_{p=1}^{l-1}\mathbb{1}_{\mathcal{A}_{j,p,i_p}>0}\,\mathcal{A}_{j,l,i_l} \geq 0 \right\}. \end{aligned}$$

It is clear that P_j is a polyhedron. Also, every coordinate function f_{k,i_k} admits the following linear representation on P_j

$$f_{k,i_k}(x) = \sum_{i_0=1}^{d}\sum_{i_1=1}^{n_1}\cdots\sum_{i_{l-1}=1}^{n_{k-1}}\prod_{p=1}^{k} x_{i_0}(W_p)_{i_{p-1},i_p}\mathbb{1}_{\mathcal{A}_{j,p,i_p}>0}, \quad \forall\, x \in P_j.$$

This implies that f_k coincides with a linear function on P_j. As every input must take one of the T strings as an activation pattern, we also have $\mathbb{R}^d = \cup_{i=1}^{T} P_j$. Thus according to Definition F.1, f_k is a PWL function. Similarly, g_k is also piecewise linear.