## 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, let $\|X\|_{\psi_2} = \inf\{t > 0 \ : \ \mathbb{E}[\exp(X^2/t^2)] \leq 2\}$.

We use the analogous definitions for vectors. In particular, let $X \in \mathbb{R}^n$ be a random vector, then $\|X\|_{\psi_2} := \sup_{\|u\|_2=1} \left\|u^\top X\right\|_{\psi_2}$ and $\|X\|_{\psi_1} := \sup_{\|u\|_2=1} \left\|u^\top X\right\|_{\psi_1}$.

We indicate with $C$ and $c$ absolute, strictly positive, numerical constants, that do not depend on the layer widths of the network $\{n_l\}_{l=0}^{L-1}$ or the number of training samples $N$. Their value may change from line to line.

## B Some Useful Estimates

**Lemma B.1.** *Under Assumption 2.5, we have that*

$$N \cdot \log^8 n_{L-1} = o(n_{L-1}n_{L-2}), \tag{22}$$

$$n_{L-2} \cdot \log^2 N \cdot \log^2 n_{L-1} = o(n_{L-1}n_{L-2}), \tag{23}$$

$$n_{L-1} \cdot \log^2 N \cdot \log^2 n_{L-1} = o(n_{L-1}n_{L-2}), \tag{24}$$

$$N \cdot \log^2 n_{L-1} \cdot \log^4 N = o(n_{L-1}n_{L-2}). \tag{25}$$

*Proof.* We start by proving (22). If $n_{L-1} = \mathcal{O}\left(N^2\right)$, then

$$N \cdot \log^8 n_{L-1} = \mathcal{O}\left(N \cdot \log^8 N\right) = o(n_{L-1}n_{L-2}), \tag{26}$$

where the last passage follows from (4). Conversely, if $n_{L-1} = \Omega(N^2)$, then

$$N \cdot \log^8 n_{L-1} = \mathcal{O}\left(\sqrt{n_{L-1}} \cdot \log^8 n_{L-1}\right) = o(n_{L-1}) = o(n_{L-1}n_{L-2}), \tag{27}$$

which concludes the proof of (22).

To obtain (23), we can exploit the second requirement of Assumption 2.5, which implies that $\log N = \mathcal{O}\left(\log n_{L-1}\right)$. This readily implies (23). Notice that (24) naturally follows since $n_{L-1} = \mathcal{O}\left(n_{L-2}\right)$ by Assumption 2.4.

Finally, to obtain (25), we write

$$N \cdot \log^2 n_{L-1} \cdot \log^4 N = \mathcal{O}\left(N \cdot \log^6 n_{L-1}\right) = o(n_{L-1}n_{L-2}), \tag{28}$$

where in the first passage we use that $\log N = \mathcal{O}\left(\log n_{L-1}\right)$ (from the second requirement of Assumption 2.5) and the last passage follows from (22). □

**Lemma B.2** (Lipschitz constant of function of the features)**.** *For all $l \in [L-1]$, and for every Lipschitz function $\varphi$, we have*

$$\|\varphi(g_l(x))\|_{\text{Lip}} = \mathcal{O}\left(1\right), \tag{29}$$

*with probability at least*

$$1 - 2l\exp\left(-n_{L-1}\right), \tag{30}$$

*over $(W_k)_{k=1}^l$. We recall that $\varphi$ is applied component-wise to $g_l(x)$, and $\varphi(g_l(x)) : \mathbb{R}^d \to \mathbb{R}^{n_l}$ is intended as a function of $x$.*

*Proof.* Note that $\varphi(g_l)$ is a composition of Lipschitz functions. Thus,

$$\begin{aligned}\|\varphi(g_l)\|_{\text{Lip}} &\leq \|\varphi\|_{\text{Lip}} \|g_l\|_{\text{Lip}} \leq \|\varphi\|_{\text{Lip}} \|W_l\|_{\text{op}} \prod_{k=1}^{l-1} \left(\|W_k\|_{\text{op}} \|\phi\|_{\text{Lip}}\right) \\ &\leq \|\varphi\|_{\text{Lip}} \|W_l\|_{\text{op}} M^{l-1} \prod_{k=1}^{l-1} \|W_k\|_{\text{op}},\end{aligned} \tag{31}$$

where the last step is justified by Assumption 2.3.

Recall that, by the assumption on the initialization of the weights, $(W_k)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_k^2/n_{k-1})$, for some constant $\beta_k$ which does not depend on the layer widths. Then, by Theorem 4.4.5 of [65], we have that, for any $k \in [l]$,

$$\|W_k\|_{\text{op}} \leq C \frac{\beta_k}{\sqrt{n_{k-1}}} (\sqrt{n_{k-1}} + 2\sqrt{n_k}), \tag{32}$$

with probability at least $1 - 2\exp(-n_k)$, $C$ being a numerical constant. By Assumption 2.4 on the topology of the network, we can rewrite this result as

$$\|W_k\|_{\text{op}} = \mathcal{O}(1), \tag{33}$$

with probability at least $1 - 2\exp(-n_{L-1})$. To conclude, using a union bound over the layers up to layer $l$, we have that

$$\|\varphi(g_l)\|_{\text{Lip}} = \mathcal{O}(1), \tag{34}$$

with probability at least $1 - 2l\exp(-n_{L-1})$ over $(W_k)_{k=1}^l$. □

**Lemma B.3.** *We have that*

$$\|D_L\|_{\text{op}} \leq \log n_{L-1}, \tag{35}$$

*with probability at least* $1 - 2\exp(-c\log^2 n_{L-1})$ *over* $W_L$*, where* $c$ *is a numerical constant.*

*Proof.* Recall that $D_L = \text{diag}(W_L)$ contains on the diagonal $n_{L-1}$ independent Gaussian random variables $(D_L)_{ii} \sim \mathcal{N}(0, \beta_L^2)$. Thus, for any $i \in [n_{L-1}]$,

$$\mathbb{P}(|(D_L)_{ii}| > \log n_{L-1}) < 2\exp(-\log^2 n_{L-1}/(2\beta_L^2)), \tag{36}$$

which gives

$$\begin{aligned}
\mathbb{P}(\|D_L\|_{\text{op}} > \log n_{L-1}) =& \mathbb{P}(\max_{i\in[n_{L-1}]} |(D_L)_{ii}| > \log n_{L-1}) \\
\leq& n_{L-1}\mathbb{P}(|(D_L)_{11}| > \log n_{L-1}) \\
<& 2\exp(\log n_{L-1} - \log^2 n_{L-1}/(2\beta_L^2)) \\
<& 2\exp(-c\log^2 n_{L-1}),
\end{aligned} \tag{37}$$

where the second step is a union bound on the entries of $D_L$. This gives the desired result. □

**Lemma B.4.** *We have that*

$$\|D_L\phi'(g_{L-1}(x))\|_{\text{Lip}} = \mathcal{O}(\log n_{L-1}), \tag{38}$$

*with probability at least* $1 - 2\exp(-c\log^2 n_{L-1}) - C\exp(-n_{L-1})$ *over* $(W_k)_{k=1}^L$*. We recall that* $\phi'$ *is applied component-wise to* $g_l(x)$*,* $D_L\phi'(g_{L-1}(x)) : \mathbb{R}^d \to \mathbb{R}^{n_{L-1}}$ *is intended as a function of* $x$*, and* $c$ *is a numerical constant.*

*Proof.* We know by composition of Lipschitz functions that

$$\|D_L\phi'(g_{L-1})\|_{\text{Lip}} \leq \|D_L\|_{\text{op}} \|\phi'(g_{L-1})\|_{\text{Lip}}. \tag{39}$$

By Assumption 2.3, $\phi'$ is Lipschitz. Hence, by combining Lemma B.2 (where we use $\varphi = \phi'$) and Lemma B.3, the result follows. □

**Lemma B.5** (Exponential tails of quadratic forms)**.** *Let* $x \sim P_X$*. Let* $u : \mathbb{R}^d \to \mathbb{R}^{d_u}$ *and* $v : \mathbb{R}^d \to \mathbb{R}^{d_v}$ *be mean-0 Lipschitz functions with respect to* $x$*, i.e.,* $\mathbb{E}_x[u(x)] = 0$*,* $\mathbb{E}_x[v(x)] = 0$*,* $\|u\|_{\text{Lip}} = c_1$ *and* $\|v\|_{\text{Lip}} = c_2$*. Let* $U$ *be a* $d_u \times d_v$ *matrix, and*

$$\Gamma(x) = u(x)^\top U v(x) - \mathbb{E}_x\left[u(x)^\top U v(x)\right]. \tag{40}$$

*Then,*

$$\|\Gamma\|_{\psi_1} < CK^2 \|U\|_F. \tag{41}$$

*where* $K = \sqrt{c_1^2 + c_2^2}$*, and* $C$ *is a numerical constant.*

*Proof.* Consider the function $z(x) : \mathbb{R}^d \to \mathbb{R}^{d_u+d_v}$ obtained by concatenating the vectors $u$ and $v$, i.e.,

$$z(x) := [u(x), v(x)]^\top. \tag{42}$$

One can readily verify that $\|z\|_{\text{Lip}} \leq \sqrt{c_1^2 + c_2^2} := K$. Let us set

$$M = \frac{1}{2}\left(\begin{array}{c|c} 0 & U \\ \hline U^\top & 0 \end{array}\right). \tag{43}$$

Then, we have that

$$\Gamma = z^\top M z - \mathbb{E}_x\left[z^\top M z\right]. \tag{44}$$

Since $x$ satisfies Assumption 2.2 and $z(x)$ is Lipschitz, in order to obtain tail bounds on $\Gamma$, we can apply the version of the Hanson-Wright inequality given by Theorem 2.3 in [1]:

$$\begin{aligned}\mathbb{P}(|\Gamma| > t) =& \mathbb{P}(|z^\top M z - \mathbb{E}_x\left[z^\top M z\right]| > t) \\ <& 2\exp\left(-\frac{1}{C_1}\min\left(\frac{t^2}{K^4\|M\|_F^2}, \frac{t}{K^2\|M\|_{\text{op}}}\right)\right) \\ \leq& 2\exp\left(-\frac{1}{C_1}\min\left(\frac{t^2}{K^4\|U\|_F^2}, \frac{t}{K^2\|U\|_{\text{op}}}\right)\right),\end{aligned} \tag{45}$$

where $C_1$ is a numerical constant, and in the last step we use that $\|M\|_{\text{op}} = \|U\|_{\text{op}}$ and $\|M\|_F^2 = \|U\|_F^2/2 \leq \|U\|_F^2$. Thus, by Lemma 5.5 of [61], we conclude that

$$\|\Gamma\|_{\psi_1} < C_2 K^2 \|U\|_F, \tag{46}$$

for some numerical constant $C_2$, which gives the desired result. □

**Lemma B.6.** *Let $u \in \mathbb{R}^{d_u}$ and $v \in \mathbb{R}^{d_v}$ be two mean-0 sub-Gaussian vectors such that $\|u\|_{\psi_2} = c_1$ and $\|v\|_{\psi_2} = c_2$. Set $A_{uv} = \mathbb{E}\left[uv^\top\right]$. Then,*

$$\|A_{uv}\|_{\text{op}} \leq C(c_1 + c_2)^2, \tag{47}$$

*where $C$ is a numerical constant.*

*Proof.* Consider the vector

$$z := [u, v]^\top. \tag{48}$$

Then, $z$ is sub-Gaussian and, by triangle inequality on the vectors $[u, 0]$ and $[0, v]$, we have that $\|z\|_{\psi_2} \leq c_1 + c_2$. Since $u$ and $v$ are mean-0, then $z$ is also mean-0 and its covariance matrix can be written as $A_z := \mathbb{E}\left[zz^\top\right]$. Furthermore, we can show that

$$\|A_z\|_{\text{op}} \leq C(c_1 + c_2)^2. \tag{49}$$

In fact, let $w$ be the unitary eigenvector associated to the maximum eigenvalue of $A_z$. Then,

$$\|A_z\|_{\text{op}} = w^\top A_z w = \mathbb{E}\left[w^\top z z^\top w\right] = \mathbb{E}\left[(w^\top z)^2\right]. \tag{50}$$

Furthermore, we have that

$$\|z\|_{\psi_2} := \sup_{w' \text{s.t.} \|w'\|_2 = 1} \left\|(w')^\top z\right\|_{\psi_2} \geq \left\|w^\top z\right\|_{\psi_2} \geq \frac{1}{C}\sqrt{\mathbb{E}\left[(w^\top z)^2\right]}, \tag{51}$$

where $C$ is a numerical constant, and the last inequality comes from Eq. (2.15) of [65]. By combining (50) and (51) with $\|z\|_{\psi_2} \leq c_1 + c_2$, (49) readily follows.

Finally, we have that

$$A_z = \left(\begin{array}{c|c} A_u & A_{uv} \\ \hline A_{uv}^\top & A_v \end{array}\right), \tag{52}$$

where $A_u := \mathbb{E}\left[uu^\top\right]$ and $A_v := \mathbb{E}\left[vv^\top\right]$. As $A_u$ and $A_v$ are PSD, we have that

$$\|A_{uv}\|_{\text{op}} \leq \|A_z\|_{\text{op}}. \tag{53}$$

Hence, the desired result follows from (49) and (53). □

**Lemma B.7.** *Let $A$ be an $N \times n$ matrix whose rows $A_i$ are i.i.d. mean-0 sub-Gaussian random vectors in $\mathbb{R}^n$. Let $K = \|A_i\|_{\psi_2}$ the sub-Gaussian norm of each row. Then, we have*

$$\left\|AA^\top\right\|_{\text{op}} = K^2 \mathcal{O}\left(N+n\right), \tag{54}$$

*with probability at least $1 - 2\exp(-cn)$, for some numerical constant $c$.*

*Proof.* Without loss of generality, we can assume $K = 1$ to simplify the proof. Let $\Sigma$ be the second moment matrix of each of the rows of $A$. Then, $\Sigma = \mathbb{E}\left[A_i A_i^\top\right]$, since the rows are mean-0. Note that, as the rows of $A$ are i.i.d., $\Sigma$ is independent of $i$. Furthermore, Lemma B.6 implies that the covariance matrix $\mathbb{E}\left[A_i A_i^\top\right]$ has operator norm bounded by a constant, since the sub-Gaussian norm of the rows is 1. Then, by using Remark 5.40 in [64], we have that

$$\left\|\frac{A^\top A}{N} - \Sigma\right\|_{\text{op}} \leq \max(\delta, \delta^2), \qquad \text{where } \delta = C\sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}}, \tag{55}$$

with probability at least $1 - 2\exp(-ct^2)$, where $c$ and $C$ are numerical constants. Setting $t = \sqrt{n}$ and using a triangular inequality gives that, with probability at least $1 - 2\exp(-ct^2)$,

$$\left\|AA^\top\right\|_{\text{op}} = \left\|A^\top A\right\|_{\text{op}} \leq N\left\|\Sigma\right\|_{\text{op}} + \max(C\sqrt{nN} + \sqrt{nN}, (C\sqrt{n} + \sqrt{n})^2) = \mathcal{O}\left(N+n\right), \tag{56}$$

which implies the desired result (after re-scaling by $K$). □

**Lemma B.8.** *Let $\tilde{F}_l = F_l - \mathbb{E}_x[F_l] \in \mathbb{R}^{N \times n_l}$ be the centered features matrix at layer $l$. Then, we have*

$$\left\|\tilde{F}_l \tilde{F}_l^\top\right\|_{\text{op}} = \mathcal{O}\left(N + n_l\right), \tag{57}$$

*with probability at least $1 - C\exp\left(-cn_{L-1}\right)$ over $(W_k)_{k=1}^l$ and $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$.*

*Proof.* From Lemma B.2, we have that

$$\left\|f_l(x)\right\|_{\text{Lip}} = \Theta(1), \tag{58}$$

with probability at least

$$1 - C'\exp\left(-n_{L-1}\right), \tag{59}$$

over $(W_k)_{k=1}^l$. We condition on this event in the rest of the proof.

Since $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$ and $P_X$ satisfies Assumption 2.2, all the rows of $\tilde{F}_l$ are mean-0 sub-Gaussian vectors, with sub-Gaussian norm bounded by a numerical constant. Here, we fix $(W_k)_{k=1}^l$ s.t. (58) holds, and the "mean-0" and the "sub-Gaussian norm" is intended w.r.t. the probability space of $(x_i)_{i=1}^N$.

An application of Lemma B.7 gives that

$$\left\|\tilde{F}_l \tilde{F}_l^\top\right\|_{\text{op}} = \mathcal{O}\left(N + n_l\right), \tag{60}$$

with probability at least $1 - 2\exp\left(-cn_l\right)$ over $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$. Taking into account the previous conditioning, we conclude that

$$\left\|\tilde{F}_l \tilde{F}_l^\top\right\|_{\text{op}} = \mathcal{O}\left(N + n_l\right), \tag{61}$$

with probability at least $1 - (C'+2)\exp\left(-cn_{L-1}\right)$ over $(W_k)_{k=1}^l$ and $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$. □

**Lemma B.9.** *Let $\tilde{B}_{L-1} = B_{L-1} - \mathbb{E}_x[B_{L-1}] \in \mathbb{R}^{N \times n_{L-1}}$ be the centered back-propagation matrix at layer $L-1$. Then, we have*

$$\left\|\tilde{B}_{L-1} \tilde{B}_{L-1}^\top\right\|_{\text{op}} = \mathcal{O}\left(\left(N + n_{L-1}\right)\log^2 n_{L-1}\right), \tag{62}$$

*with probability at least $1 - C\exp\left(-cn_{L-1}\right)$ over $(W_k)_{k=l}^L$ and $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$.*

*Proof.* From Lemma B.4, we have that

$$\|D_L\phi'(g_{L-1}(x))\|_{\mathrm{Lip}} = \mathcal{O}(\log n_{L-1}), \tag{63}$$

with probability at least

$$1 - 2\exp(-c\log^2 n_{L-1}) - C\exp(-n_{L-1}), \tag{64}$$

over $(W_k)_{k=1}^L$. We condition on this event in the rest of the proof.

Since $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$ and $P_X$ satisfies Assumption 2.2, all the rows of $\tilde{B}_{L-1}$ are mean-0 sub-Gaussian vectors, with sub-Gaussian norm $\mathcal{O}(\log n_{L-1})$. Here, we fix $(W_k)_{k=1}^L$ s.t. (63) holds, and the "mean-0" and the "sub-Gaussian norm" is intended w.r.t. the probability space of $(x_i)_{i=1}^N$.

An application of Lemma B.7 gives that

$$\left\|\tilde{B}_{L-1}\tilde{B}_{L-1}^\top\right\|_{\text{op}} = \mathcal{O}\left((N+n_{L-1})\log^2 n_{L-1}\right), \tag{65}$$

with probability at least $1 - 2\exp(-cn_{L-1})$ over $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$. Taking into account the previous conditioning, we conclude that

$$\left\|\tilde{B}_{L-1}\tilde{B}_{L-1}^\top\right\|_{\text{op}} = \mathcal{O}\left((N+n_{L-1})\log^2 n_{L-1}\right), \tag{66}$$

with probability at least $1 - (C'+2)\exp(-cn_{L-1})$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$. □

**Lemma B.10.** *Let $\rho$ be a standard Gaussian random variable, then we have that*

$$\varphi_1(c) := \mathbb{E}_\rho\left[\phi(c\rho)\right], \tag{67}$$

*and*

$$\varphi_2(c) := \mathbb{E}_\rho\left[\phi^2(c\rho)\right], \tag{68}$$

*are continuous functions in c. Furthermore, $\varphi_1(c)$ is Lipschitz in c, and*

$$|\varphi_2(c_1) - \varphi_2(c_2)| \le C_1|c_1 - c_2| + C_2|c_1^2 - c_2^2|, \tag{69}$$

*where $C_1$ and $C_2$ are numerical constants (independent of $c_1, c_2$).*

*Proof.* Let $p(\rho) = \frac{1}{\sqrt{2\pi}}e^{-\rho^2/2}$. Then, we have

$$\begin{aligned}
|\varphi_1(c+\varepsilon) - \varphi_1(c)| &\le \int p(\rho)\,|\phi((c+\varepsilon)\rho) - \phi(c\rho)|\,d\rho \\
&\le \int p(\rho)\,|M\varepsilon\rho|\,d\rho \\
&= M\varepsilon\mathbb{E}_\rho\left[|\rho|\right] \\
&= C\varepsilon,
\end{aligned} \tag{70}$$

where in the second line we use that $\phi$ is $M$-Lipschitz by Assumption 2.3. Similarly, we have

$$\begin{aligned}
|\varphi_2(c+\varepsilon) - \varphi_2(c)| &\le \int p(\rho)\left|\phi^2((c+\varepsilon)\rho) - \phi^2(c\rho)\right|d\rho \\
&= \int p(\rho)\,|\phi((c+\varepsilon)\rho) - \phi(c\rho)|\,|\phi((c+\varepsilon)\rho) + \phi(c\rho)|\,d\rho \\
&\le \int p(\rho)\,|M\varepsilon\rho|\,(2\,|\phi(0)| + M(|c+\varepsilon| + |\varepsilon|)|\rho|)\,d\rho \\
&= C_1\varepsilon\mathbb{E}_\rho\left[|\rho|\right] + C_2\varepsilon|c|\mathbb{E}_\rho\left[\rho^2\right] + C_3\varepsilon^2\mathbb{E}_\rho\left[\rho^2\right] \\
&= C_4\varepsilon + C_2|c|\varepsilon + C_3\varepsilon^2 \\
&\le C_5\varepsilon + C_6\left|(c+\varepsilon)^2 - c^2\right|.
\end{aligned} \tag{71}$$

□

**Lemma B.11.** *Let $\rho$ be a standard Gaussian distribution, and $c \neq 0$ be an absolute constant. Then, we have*

$$|\mathbb{E}_\rho \left[\phi(c\rho)\right]| = \mathcal{O}\left(1\right). \tag{72}$$

*It also holds*

$$|\mathbb{E}_\rho \left[\phi'(c\rho)\right]| = \mathcal{O}\left(1\right). \tag{73}$$

*Proof.* For the first statement, we exploit the fact that $\phi$ is Lipschitz:

$$|\mathbb{E}_\rho \left[\phi(c\rho)\right]| \leq \mathbb{E}_\rho \left[|\phi(0)| + M|c\rho|\right] = |\phi(0)| + M|c|\mathbb{E}_\rho \left[|\rho|\right] = C_1. \tag{74}$$

The statement on $\phi'$ is easily derived following the same proof and using that $\|\phi'\|_{\mathrm{Lip}} \leq M'$. □

**Lemma B.12.** *Let $\rho$ be a standard Gaussian distribution, and $c \neq 0$ be an absolute constant. Then, we have*

$$\mathbb{E}_\rho \left[\phi^2(c\rho)\right] = \Theta(1), \tag{75}$$

*and*

$$\mathbb{E}_\rho \left[(\phi'(c\rho))^2\right] = \Theta(1). \tag{76}$$

*Proof.* For the upper-bound of the first statement, we exploit the fact that $\phi$ is Lipschitz:

$$\mathbb{E}_\rho \left[\phi^2(c\rho)\right] \leq \mathbb{E}_\rho \left[(|\phi(0)| + M|c\rho|)^2\right] = \phi^2(0) + 2M|c||\phi(0)|\mathbb{E}_\rho \left[|\rho|\right] + M^2c^2\mathbb{E}_\rho \left[\rho^2\right] = C_1. \tag{77}$$

For the lower bound, since $\phi$ is non-zero and continuous, we have that there exist a strictly positive constant $c' > 0$ and an interval $[c_1, c_2]$ with $c_2 > c_1$ such that $\phi^2(x) \geq c'$ for each $x \in [c_1, c_2]$. Therefore, we have

$$\mathbb{E}_\rho \left[\phi^2(c\rho)\right] \geq c'\mathbb{P}(c_1 \leq c\rho \leq c_2) = C_2. \tag{78}$$

The second statements is proved in the same way, as $\phi'$ is a non-zero Lipschitz function. □

**Lemma B.13.** *Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ a Lipschitz function, and let $x \sim P_X$. Then,*

$$\mathbb{E}_x^2 \left[\varphi(x)\right] \geq \mathbb{E}_x \left[\varphi(x)^2\right] - c\left\|\varphi\right\|_{\mathrm{Lip}}^2, \tag{79}$$

*where $c$ is a numerical constant.*

*Proof.* We have

$$\begin{aligned}
\mathbb{E}_x^2 \left[\varphi(x)\right] &= \mathbb{E}_x \left[(\varphi(x))^2\right] - \mathbb{E}_x \left[\left(\varphi(x) - \mathbb{E}_x \left[\varphi(x)\right]\right)^2\right] \\
&= \mathbb{E}_x \left[\varphi(x)^2\right] - \int_0^{+\infty} \mathbb{P}\left(\left(\varphi(x) - \mathbb{E}_x \left[\varphi(x)\right]\right)^2 > t\right) dt \\
&= \mathbb{E}_x \left[\varphi(x)^2\right] - \int_0^{+\infty} \mathbb{P}\left(|\varphi(x) - \mathbb{E}_x \left[\varphi(x)\right]| > \sqrt{t}\right) dt \\
&\geq \mathbb{E}_x \left[\varphi(x)^2\right] - \int_0^{+\infty} 2\exp\left(-Ct/\left\|\varphi\right\|_{\mathrm{Lip}}^2\right) dt \\
&= \mathbb{E}_x \left[\varphi(x)^2\right] - 2\left\|\varphi\right\|_{\mathrm{Lip}}^2/C,
\end{aligned} \tag{80}$$

where the inequality is a consequence of Assumption 2.2. □

**Lemma B.14.** *Let $x \sim P_X$, and define $\tilde{c}_l(x) = \beta_l \left\|f_l(x)\right\| / \sqrt{n_l}$.*

*Then, we have*

$$\mathbb{E}_x \left[|\tilde{c}_l(x) - \mathbb{E}_x \left[\tilde{c}_l(x)\right]|\right] \leq C\frac{\left\|f_l\right\|_{\mathrm{Lip}}}{\sqrt{n_l}}, \tag{81}$$

*and*

$$\mathbb{E}_x \left[\left(\tilde{c}_l(x) - \mathbb{E}_x \left[\tilde{c}_l(x)\right]\right)^2\right] \leq C\frac{\left\|f_l\right\|_{\mathrm{Lip}}^2}{n_l}, \tag{82}$$

*where $C$ is a numerical constant.*

*Proof.* We have that

$$\begin{aligned}\mathbb{E}_x\left[|\tilde{c}_l(x) - \mathbb{E}_x\left[\tilde{c}_l(x)\right]|\right] &= \int_0^{+\infty} \mathbb{P}\left(|\tilde{c}_l(x) - \mathbb{E}_x\left[\tilde{c}_l(x)\right]| > t\right) dt \\ &= \int_0^{+\infty} \mathbb{P}\left(|\|f_l(x)\| - \mathbb{E}_x\left[\|f_l(x)\|\right]| > \sqrt{n_l}t/\beta_l\right) dt \\ &\leq \int_0^{+\infty} 2\exp\left(-cn_l t^2/\left\|f_l\right\|_{\mathrm{Lip}}^2\right) dt \\ &= C\frac{\|f_l\|_{\mathrm{Lip}}}{\sqrt{n_l}},\end{aligned} \tag{83}$$

where $c$ and $C$ are numerical constants, and the third line is justified by Assumption 2.2.

Similarly, we have

$$\begin{aligned}\mathbb{E}_x\left[\left(\tilde{c}_l(x) - \mathbb{E}_x\left[\tilde{c}_l(x)\right]\right)^2\right] &= \int_0^{+\infty} \mathbb{P}\left(\left(\tilde{c}_l(x) - \mathbb{E}_x\left[\tilde{c}_l(x)\right]\right)^2 > t\right) dt \\ &= \int_0^{+\infty} \mathbb{P}\left(|\|f_l(x)\| - \mathbb{E}_x\left[\|f_l(x)\|\right]| > \sqrt{n_l t}/\beta_l\right) dt \\ &\leq \int_0^{+\infty} 2\exp\left(-cn_l t/\left\|f_l\right\|_{\mathrm{Lip}}^2\right) dt \\ &= C\frac{\|f_l\|_{\mathrm{Lip}}^2}{n_l},\end{aligned} \tag{84}$$

where, again, $c$ and $C$ are numerical constants and the third line is justified by Assumption 2.2. □

**Lemma B.15.** *Let $\rho_1$ and $\rho_2$ be two standard Gaussian random variables, possibly correlated. Then, we have*

$$\begin{aligned}&\left|\mathbb{E}_{\rho_1\rho_2}\left[\phi(\rho_1 x_1)\phi(\rho_2 x_2) - \phi(\rho_1 y_1)\phi(\rho_2 y_2)\right]\right| \leq \\ &\qquad \leq C_1\left|x_1 - y_1\right| + C_2\left|x_2\right|\left|x_1 - y_1\right| + C_3\left|x_2 - y_2\right| + C_4\left|y_1\right|\left|x_2 - y_2\right|,\end{aligned} \tag{85}$$

*where $C_1, C_2, C_3, C_4$ are numerical constants (which do not depend on $x_1, x_2, y_1, y_2$). Furthermore, the same result holds with $\phi'$ instead of $\phi$.*

*Proof.* We have

$$\begin{aligned}&\left|\mathbb{E}_{\rho_1\rho_2}\left[\phi(\rho_1 x_1)\phi(\rho_2 x_2) - \phi(\rho_1 y_1)\phi(\rho_2 y_2)\right]\right| \\ &\leq \left|\mathbb{E}_{\rho_1\rho_2}\left[\phi(\rho_1 x_1)\phi(\rho_2 x_2) - \phi(\rho_1 y_1)\phi(\rho_2 x_2)\right] + \mathbb{E}_{\rho_1\rho_2}\left[\phi(\rho_1 y_1)\phi(\rho_2 x_2) - \phi(\rho_1 y_1)\phi(\rho_2 y_2)\right]\right| \\ &\leq \mathbb{E}_{\rho_1\rho_2}\left[\left|\phi(\rho_1 x_1) - \phi(\rho_1 y_1)\right|\left|\phi(\rho_2 x_2)\right|\right] + \mathbb{E}_{\rho_1\rho_2}\left[\left|\phi(\rho_2 x_2) - \phi(\rho_2 y_2)\right|\left|\phi(\rho_1 y_1)\right|\right] \\ &\leq \mathbb{E}_{\rho_1\rho_2}\left[\left|M\rho_1(x_1 - y_1)\right|\left(\left|\phi(0)\right| + M\left|\rho_2 x_2\right|\right)\right] + \mathbb{E}_{\rho_1\rho_2}\left[\left|M\rho_2(x_2 - y_2)\right|\left(\left|\phi(0)\right| + M\left|\rho_1 y_1\right|\right)\right] \\ &\leq C_1\left|x_1 - y_1\right|\mathbb{E}\left[\left|\rho_1\right|\right] + C_2\left|x_2\right|\left|x_1 - y_1\right|\mathbb{E}\left[\left|\rho_1\right|\left|\rho_2\right|\right] \\ &\qquad + C_3\left|x_2 - y_2\right|\mathbb{E}\left[\left|\rho_2\right|\right] + C_4\left|y_1\right|\left|x_2 - y_2\right|\mathbb{E}\left[\left|\rho_1\right|\left|\rho_2\right|\right] \\ &\leq C_1\left|x_1 - y_1\right| + C_2\left|x_2\right|\left|x_1 - y_1\right| + C_3\left|x_2 - y_2\right| + C_4\left|y_1\right|\left|x_2 - y_2\right|,\end{aligned} \tag{86}$$

where in third inequality we use that $\phi$ is $M$-Lipschitz, and in the last inequality we use that the quantities $\mathbb{E}\left[|\rho_1|\right]$, $\mathbb{E}\left[|\rho_2|\right]$ and $\mathbb{E}\left[|\rho_1||\rho_2|\right]$ are all smaller than 1 (regardless of the correlation between $\rho_1$ and $\rho_2$). Since we only used the fact that $\phi$ is $M$-Lipschitz, the same result holds with $\phi'$ in place of $\phi$.

□

# C Concentration of $\ell_2$ Norms

In this appendix, we state and prove a number of high-probability estimates on the $\ell_2$ norms of feature and backpropagation vectors. More specifically, our results can be summarized as follows:

- Lemma C.1 gives tight bounds on $\|f_l(x)\|_2$, *i.e.* the $\ell_2$ norm of the feature vector at layer $l$. The statement holds with high probability over $x$ and $(W_k)_{k=1}^l$.
- Lemmas C.2 and C.3 give tight bounds on $\mathbb{E}_x\left[\|f_l(x)\|_2^2\right]$ and $\mathbb{E}_x\left[\|f_l(x)\|_2\right]$, respectively. These quantities represent the *expectation* with respect to $x$ of the (squared) $\ell_2$ norm of the feature vector at layer $l$. The statements hold with high probability over $(W_k)_{k=1}^l$.
- Lemma C.4 focuses on the *centered* feature vector $f_l(x) - \mathbb{E}_x\left[f_l(x)\right]$, and it gives tight bounds on *(i)* its expected (w.r.t. $x$) squared $\ell_2$ norm $\mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2^2\right]$, *(ii)* its expected (w.r.t. $x$) $\ell_2$ norm $\mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2\right]$, and *(iii)* its $\ell_2$ norm $\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2$. The first two statements hold with high probability over $(W_k)_{k=1}^l$, and the probability in the last statement is also over $x$.
- Lemma C.5 focuses on the *centered* backpropagation vector at layer $L-1$, and it gives tight bounds on its $\ell_2$ norm $\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2$. This statement holds with high probability over $x$ and $(W_k)_{k=1}^l$.

Throughout this appendix, we always assume that $P_X$ satisfies Assumptions 2.1 and 2.2, and that the layer widths satisfy Assumption 2.4. Furthermore, we use that the activation $\phi$ and its derivative $\phi'$ are Lipschitz (see Assumption 2.3).

**Lemma C.1** ($\ell_2$ norm of features)**.** *Let $x \sim P_X$. Then, for every $0 \le l \le L-1$,*

$$\|f_l(x))\|_2 = \Theta(\sqrt{n_l}), \tag{87}$$

*with probability at least $1 - C\exp(-cn_{L-1})$ over $x$ and $(W_k)_{k=1}^l$. As usual, $\phi$ is applied component-wise on $g_l(x)$, and $c$ and $C$ are numerical constants.*

*Proof.* We prove this by induction over $l$, and we start with the base case ($l = 0$). Recall that we have defined $f_0(x) := x$. As the $\ell_2$ norm is a 1-Lipschitz function, by Assumption 2.2, we have that

$$\mathbb{P}\left(\left|\|x\|_2 - \mathbb{E}[\|x\|_2]\right| > t\right) \le 2e^{-ct^2}. \tag{88}$$

Furthermore, Assumption 2.1 implies that $\mathbb{E}[\|x\|_2] = \Theta(\sqrt{d})$, hence setting $t = \mathbb{E}[\|x\|_2]/2$ in (88) proves the desired result for the base case (recalling that $n_{L-1} = \mathcal{O}(d)$ by Assumption 2.4).

By inductive hypothesis, we have

$$\|f_{l-1}(x))\|_2 = \Theta(\sqrt{n_{l-1}}), \tag{89}$$

with probability at least $1 - C\exp(-cn_{L-1})$.

Define $\tilde{c} := \beta_l \|f_{l-1}(x)\|_2 / \sqrt{n_{l-1}}$. From now on, we condition on a realization of $x$ and $(W_k)_{k=1}^{l-1}$ such that $\tilde{c} = \Theta(1)$. By (89), this happens with probability at least $1 - C\exp(-cn_{L-1})$.

To ease the notation, we use the shorthands $f := f_{l-1}(x)$ and $W := W_l$. Then, we can write

$$\|f_l(x)\|_2 = \left\|\phi(W^\top f)\right\|_2 = \sqrt{n_l}\sqrt{\frac{1}{n_l}\sum_{i=1}^{n_l}\phi^2((W^\top)_{i:}f)}. \tag{90}$$

Recall that $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/n_{l-1})$ and that the Gaussian distribution is rotationally invariant. Thus, the RHS of (90) has the same distribution as

$$\sqrt{n_l}\sqrt{\frac{1}{n_l}\sum_{i=1}^{n_l}\phi^2\left(\tilde{c}\rho_i\right)} = \sqrt{n_l}\sqrt{\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}\rho_1\right)\right] + \frac{1}{n_l}\sum_{i=1}^{n_l}Z_i}, \tag{91}$$

where $(\rho_i)_{i=1}^{n_l} \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$ and also independent of $f$, and we have defined the independent, mean-0 random variables

$$Z_i = \phi^2\left(\tilde{c}\rho_i\right) - \mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}\rho_1\right)\right]. \tag{92}$$

Note that, in the definition of $Z_i$, the randomness comes only from $\rho_i$, since we are conditioning on $\tilde{c}$.

We have that

$$\|\phi\left(\tilde{c}\rho_i\right)\|_{\psi_2} \leq \|\phi\left(\tilde{c}\rho_i\right) - \mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}\rho_i\right)\right]\|_{\psi_2} + \|\mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}\rho_i\right)\right]\|_{\psi_2} \leq C_1 + C_2 = C_3, \tag{93}$$

where the first term is bounded by a constant by Theorem 5.2.2 in [65], and the bound on the second term follows from Lemma B.11. As a consequence, we have

$$\begin{aligned}\|Z_i\|_{\psi_1} =& \left\|\phi^2\left(\tilde{c}\rho_i\right) - \mathbb{E}_{\rho_i}\left[\phi^2\left(\tilde{c}\rho_i\right)\right]\right\|_{\psi_1} \\ \leq& C_4\left\|\phi^2\left(\tilde{c}\rho_i\right)\right\|_{\psi_1} \\ =& C_4\left\|\phi\left(\tilde{c}\rho_i\right)\right\|_{\psi_2}^2 \\ \leq& C_5,\end{aligned} \tag{94}$$

where the inequality in the second line follows from Exercise 2.7.10 of [65], the equality in the third line follows from Lemma 2.7.6 of [65], and the inequality in the last line follows from (93). Hence, the $Z_i$-s are i.i.d. sub-exponential random variables, with sub-exponential norm bounded by a numerical constant. An application of Bernstein inequality (cf. Corollary 2.8.3. in [65]) gives that

$$\mathbb{P}\left(\left|\frac{1}{n_l}\sum_{i=1}^{n_l} Z_i\right| > t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{C_6^2}, \frac{t}{C_6}\right)n_l\right), \tag{95}$$

where $c, C_6$ are numerical constants. Furthermore, by Lemma B.12, we have

$$\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}\rho_1\right)\right] = \Theta(1). \tag{96}$$

By setting $t = \mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}\rho_1\right)\right]/2$ into (95) and using (91) and (96), we conclude that

$$\left\|\phi(W^\top f)\right\|_2 = \Theta(\sqrt{n_l}), \tag{97}$$

with probability at least $1 - C\exp(-cn_{L-1}) - 2\exp(-cn_l) \geq 1 - C_1\exp(-cn_{L-1})$, for some numerical constant $c$ and $C_1$, which concludes the proof. □

**Lemma C.2** (Expected squared $\ell_2$ norm of features)**.** *Let $x \sim P_X$. Then, for every $0 \leq l \leq L-1$,*

$$\mathbb{E}_x\left[\|f_l(x)\|_2^2\right] = \Theta(n_l), \tag{98}$$

*with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$. As usual, $c$ and $C$ are numerical constants.*

*Proof.* The argument is by induction over $l$. The base case is a direct consequence of Assumption 2.1, since $f_0(x) = x$.

By inductive hypothesis, we have

$$\mathbb{E}_x\left[\|f_{l-1}(x))\|_2^2\right] = \Theta(n_{l-1}), \tag{99}$$

with probability at least $1 - C\exp(-cn_{L-1})$. Define $\tilde{c}(x) := \beta_l\|f_{l-1}(x)\|_2/\sqrt{n_{l-1}}$. From now on, we condition on a realization of $(W_k)_{k=1}^{l-1}$ such that $\mathbb{E}_x\left[\tilde{c}^2(x)\right] = \Theta(1)$. By (99), this happens with probability at least $1 - C\exp(-cn_{L-1})$.

To ease the notation, we use the shorthands $f := f_{l-1}(x)$, $W := W_l$ and $w_i = W_{:i}$. Then, we can write

$$\begin{aligned}\mathbb{E}_x\left[\|f_l(x))\|_2^2\right] &= \mathbb{E}_x\left[\left\|\phi(W^\top f)\right\|_2^2\right] \\ &= n_l\left(\frac{1}{n_l}\sum_{i=1}^{n_l}\mathbb{E}_x\left[\phi^2((W^\top)_{i:}f)\right]\right) \\ &= n_l\left(\mathbb{E}_{w_1}\mathbb{E}_x\left[\phi^2(w_1^\top f)\right] + \frac{1}{n_l}\sum_{i=1}^{n_l} Z_i\right),\end{aligned} \tag{100}$$

where we use that the $w_i$-s are equally distributed and we have defined the independent, mean-0 random variables

$$Z_i = \mathbb{E}_x\left[\phi^2\left(w_i^\top f(x)\right)\right] - \mathbb{E}_{w_i}\mathbb{E}_x\left[\phi^2\left(w_i^\top f(x)\right)\right]. \tag{101}$$

Note that, in the definition of $Z_i$, the randomness comes only from $w_i$, since we are conditioning on $(W_k)_{k=1}^{l-1}$.

We have that

$$\begin{aligned}\|Z_i\|_{\psi_1} \leq & \mathbb{E}_x \left[ \left\| \phi^2 \left( w_i^\top f(x) \right) - \mathbb{E}_{w_i} \left[ \phi^2 \left( w_i^\top f(x) \right) \right] \right\|_{\psi_1} \right] \\ \leq & \mathbb{E}_x \left[ C_1 \left\| \phi^2 \left( w_i^\top f(x) \right) \right\|_{\psi_1} \right] \\ = & C_1 \mathbb{E}_x \left[ \left\| \phi \left( w_i^\top f(x) \right) \right\|_{\psi_2}^2 \right],\end{aligned} \tag{102}$$

where the first line follows from Jensen's inequality as $\|\cdot\|_{\psi_1}$ is convex, the inequality in the second line follows from Exercise 2.7.10 of [65], and the equality in the third line follows from Lemma 2.7.6 of [65].

Recall that $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/n_{l-1})$ and that the Gaussian distribution is rotationally invariant. Thus, $\phi\left(w_i^\top f(x)\right)$ has the same distribution as $\phi\left(\tilde{c}(x)\rho_i\right)$, where $(\rho_i)_{i=1}^{n_l} \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$ and also independent of $\tilde{c}(x)$. We now condition on a realization of $x$ and $(W_k)_{k=1}^{l-1}$ and provide an upper bound on the sub-Gaussian norm $\left\|\phi\left(w_i^\top f(x)\right)\right\|_{\psi_2}$, where the only randomness comes again from $w_i$ (and, hence, from $\rho_i$). We have that

$$\begin{aligned}\left\|\phi\left(w_i^\top f(x)\right)\right\|_{\psi_2} &= \left\|\phi\left(\tilde{c}(x)\rho_i\right)\right\|_{\psi_2} \\ &\leq \left\|\phi\left(\tilde{c}(x)\rho_i\right) - \mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}(x)\rho_i\right)\right]\right\|_{\psi_2} + \left\|\mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}(x)\rho_i\right)\right]\right\|_{\psi_2} \\ &\leq C_1\tilde{c}(x) + C_2\tilde{c}(x) + C_3 = C_4(\tilde{c}(x)+1).\end{aligned} \tag{103}$$

where the first term in the RHS in the second line is bounded by $C_1\tilde{c}(x)$ by Theorem 5.2.2 in [65], and the second term is bounded by $C_2\tilde{c}(x) + C_3$ by following the same proof of Lemma B.11 as $\phi$ is Lipschitz. By combining (102) and (103), we get

$$\|Z_i\|_{\psi_1} \leq C_4^2 \mathbb{E}_x\left[(\tilde{c}(x)+1)^2\right] \leq C_5, \tag{104}$$

where we use that $\mathbb{E}_x\left[\tilde{c}^2(x)\right] = \Theta(1)$.

Hence, the $Z_i$-s are i.i.d. sub-exponential random variables, with sub-exponential norm bounded by a numerical constant. An application of Bernstein inequality (cf. Corollary 2.8.3. in [65]) gives that

$$\mathbb{P}\left(\left|\frac{1}{n_l}\sum_{i=1}^{n_l} Z_i\right| > t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{C_5^2}, \frac{t}{C_5}\right) n_l\right), \tag{105}$$

where $c, C_5$ are numerical constants.

Let us consider the first term in (100):

$$\mathbb{E}_x \mathbb{E}_{w_1}\left[\phi^2(w_1^\top f)\right] = \mathbb{E}_x \mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}(x)\rho_1\right)\right], \tag{106}$$

where the equality comes again from the rotational invariance of the Gaussian distribution of $w_1$. We will show that

$$\mathbb{E}_x \mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}(x)\rho_1\right)\right] = \Theta(1), \tag{107}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{l-1}$.

The upper bound in (107) follows from the same passages in (77), as $\mathbb{E}_x[\tilde{c}^2(x)] = \Theta(1)$. We now prove the lower bound. By Lemma C.1, we have that there exist numerical constants $c_2 > c_1 > 0$ such that $\tilde{c}(x) \in [c_1, c_2]$ with probability at least $1 - C\exp(-cn_{L-1})$ over $x$ and $(W_k)_{k=1}^{l-1}$. Hence, with probability at least $1 - 2C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{l-1}$, we have that

$$\mathbb{P}_x(\tilde{c}(x) \in [c_1, c_2]) \geq 1/2, \tag{108}$$

where we use the symbol $\mathbb{P}_x$ to highlight that this last probability is taken over $x$. Let us condition on a realization of $(W_k)_{k=1}^{l-1}$ s.t. (108) holds. Then, we have

$$\mathbb{E}_x \mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}(x)\rho_1\right)\right] \geq \frac{1}{2}\inf_{c\in[c_1,c_2]} \mathbb{E}_{\rho_1}\left[\phi^2\left(c\rho_1\right)\right]. \tag{109}$$

By Lemma B.10, we have that $\varphi(c) = \mathbb{E}_{\rho_1}\left[\phi^2\left(c\rho_1\right)\right]$ is continuous in $c$. Therefore, by Weierstrass theorem, there exists a strictly positive $c^* \in [c_1, c_2]$ such that $\inf_{c\in[c_1,c_2]} \mathbb{E}_{\rho_1}\left[\phi^2\left(c\rho_1\right)\right] = \mathbb{E}_{\rho_1}\left[\phi^2\left(c^*\rho_1\right)\right]$. Thus,

$$\mathbb{E}_x\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}(x)\rho_1\right)\right] \geq \frac{1}{2}\mathbb{E}_{\rho_1}\left[\phi^2\left(c^*\rho_1\right)\right] = \Theta(1), \tag{110}$$

where the last equality is a consequence of Lemma B.12. This concludes the proof of the lower bound in (107).

By setting $t = \mathbb{E}_{\rho_1}\left[\phi^2\left(c^*\rho_1\right)\right]/4$ into (105) and using (107) and (100), we conclude that

$$\mathbb{E}_x\left[\|f_l(x))\|_2^2\right] = \Theta(n_l), \tag{111}$$

with probability at least $1 - C\exp(-cn_{L-1})$, for some numerical constants $C$ and $c$, which concludes the proof. □

**Lemma C.3** (Expected $\ell_2$ norm of features)**.** *Let $x \sim P_X$. Then, for every $0 \leq l \leq L-1$,*

$$\mathbb{E}_x\left[\|f_l(x)\|_2\right] = \Theta(\sqrt{n_l}), \tag{112}$$

*with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$. As usual, $c$ and $C$ are numerical constants.*

*Proof.* We condition on the events

$$\mathbb{E}_x\left[\|f_l(x)\|_2^2\right] = \Theta(n_l), \tag{113}$$

and

$$\left\|\|f_l(x)\|_2\right\|_{\mathrm{Lip}} = \mathcal{O}\left(1\right), \tag{114}$$

which happen with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$ by Lemma C.2 and B.2.

The upper bound is a direct consequence of Jensen's inequality:

$$\mathbb{E}_x\left[\|f_l(x)\|_2\right] \leq \sqrt{\mathbb{E}_x\left[\|f_l(x)\|_2^2\right]} = \Theta(\sqrt{n_l}). \tag{115}$$

For the lower bound, we use Lemma B.13, and we obtain

$$\mathbb{E}_x\left[\|f_l(x)\|_2\right] \geq \sqrt{\mathbb{E}_x\left[\|f_l(x)\|_2^2\right] - c\left\|\|f_l(x)\|_2\right\|_{\mathrm{Lip}}^2} = \Theta(\sqrt{n_l}). \tag{116}$$

□

**Lemma C.4** ($\ell_2$ norms of centered features)**.** *Let $x \sim P_X$. Then, for every $0 \leq l \leq L-1$, the following results hold.*

1. *With probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$, we have that*

$$\mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2^2\right] = \Theta(n_l). \tag{117}$$

2. *With probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$, we have that*

$$\mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2\right] = \Theta(\sqrt{n_l}). \tag{118}$$

3. *With probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$ and $x$, we have that*

$$\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2 = \Theta(\sqrt{n_l}). \tag{119}$$

*Proof.* The argument is by induction over $l$. The base case for (117) follows directly from Assumption 2.1, since $f_0(x) = x$. Since the $\ell_2$ norm is a 1-Lipschitz function, from Jensen inequality and Lemma B.13 we readily obtain the base case for (118). Note that $\left\|\|x - \mathbb{E}_x[x]\|_2\right\|_{\text{Lip}} \leq 1$. Then, the base case for (119) is a direct consequence of Assumption 2.2 on $x$ and of the base case of (118), and it holds with probability at least $1 - C' \exp(-cd) \geq 1 - C' \exp(-cn_{L-1})$ over $x$.

By inductive hypothesis, we assume the three statements to be true for layer $l-1$, for $l \in [L-1]$. We will now prove (117) for layer $l$.

To ease the notation, we use the shorthands $f(x) := f_{l-1}(x)$, $f = \mathbb{E}_x[f(x)]$, $\tilde{f}(x) = f(x) - f$, $W := W_l$ and $w_i = W_{:i}$. We also define $\tilde{c}(x) = \beta_l \left\|f(x)\right\| / \sqrt{n_{l-1}}$ and $\tilde{c} = \mathbb{E}_x[\tilde{c}(x)]$.

We condition on the following events in the probability space of $(W_k)_{k=1}^{l-1}$:

(a) $\left\|f(x)\right\|_{\text{Lip}} = \mathcal{O}(1)$, which happens with probability at least $1 - C' \exp(-cn_{L-1})$ by Lemma B.2.

(b) $\tilde{c} = \Theta(1)$, which happens with probability at least $1 - C' \exp(-cn_{L-1})$ by Lemma C.3. Notice that by Jensen inequality this also implies $\left\|f\right\|_2^2 = \mathcal{O}(n_{l-1})$.

(c) By inductive hypothesis, we have that, with probability at least $1 - C' \exp(-cn_{n-1})$ over $x$ and $(W_k)_{k=1}^{l-1}$,

$$\left\|\tilde{f}(x)\right\|_2^2 = \Theta(n_{l-1}). \tag{120}$$

Hence, with probability at least $1 - 2C' \exp(-cn_{n-1})$ over $(W_k)_{k=1}^{l-1}$, we have that

$$\mathbb{P}_x\left(c_1 n_{l-1} \leq \left\|\tilde{f}(x)\right\|_2^2 \leq c_2 n_{l-1}\right) \geq 1/2, \tag{121}$$

for some numerical constants $c_2 > c_1 > 0$. In (121), we use the symbol $\mathbb{P}_x$ to highlight that this last probability is taken over $x$. For the rest of the argument, we condition on a realization of $(W_k)_{k=1}^{l-1}$ s.t. (121) holds.

By taking a union bound, the events (a)-(c) happen with probability at least $1 - 4C' \exp(-cn_{n-1})$ over $(W_k)_{k=1}^{l-1}$.

Now, we can write

$$\begin{aligned}
\mathbb{E}_x\left[\left\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\right\|_2^2\right] &= \mathbb{E}_x\left[\left\|\phi(W^\top f(x)) - \mathbb{E}_x\left[\phi(W^\top f(x))\right]\right\|_2^2\right] \\
&= n_l\left(\frac{1}{n_l}\sum_{i=1}^{n_l}\mathbb{E}_x\left[\left(\phi((W^\top)_{i:}f(x)) - \mathbb{E}_x\left[\phi((W^\top)_{i:}f(x))\right]\right)^2\right]\right) \\
&= n_l\left(\mathbb{E}_{xw_1}\left[\left(\phi(w_1^\top f(x)) - \mathbb{E}_x\left[\phi(w_1^\top f(x))\right]\right)^2\right] + \frac{1}{n_l}\sum_{i=1}^{n_l} Z_i\right),
\end{aligned} \tag{122}$$

where we use that the $w_i$-s are identically distributed and we have defined the independent, mean-0 random variables

$$Z_i = \mathbb{E}_x\left[\left(\phi(w_i^\top f(x)) - \mathbb{E}_x\left[\phi(w_i^\top f(x))\right]\right)^2\right] - \mathbb{E}_{w_1 x}\left[\left(\phi(w_1^\top f(x)) - \mathbb{E}_x\left[\phi(w_1^\top f(x))\right]\right)^2\right]. \tag{123}$$

As in the proof of Lemma C.2, in the definition of $Z_i$, the randomness comes only from $w_i$, since we are conditioning on $(W_k)_{k=1}^{l-1}$.

We have that

$$\begin{aligned}
\|Z_i\|_{\psi_1} &\leq C_0 \left\| \mathbb{E}_x \left[ \left( \phi(w_i^\top f(x)) - \mathbb{E}_x \left[ \phi(w_i^\top f(x)) \right] \right)^2 \right] \right\|_{\psi_1} \\
&\leq C_0 \mathbb{E}_x \left[ \left\| \left( \phi(w_i^\top f(x)) - \mathbb{E}_x \left[ \phi(w_i^\top f(x)) \right] \right)^2 \right\|_{\psi_1} \right] \\
&= C_0 \mathbb{E}_x \left[ \left\| \phi(w_i^\top f(x)) - \mathbb{E}_x \left[ \phi(w_i^\top f(x)) \right] \right\|_{\psi_2}^2 \right] \\
&\leq C_0 \mathbb{E}_x \left[ \left( \left\| \phi(w_i^\top f(x)) \right\|_{\psi_2} + \left\| \mathbb{E}_x \left[ \phi(w_i^\top f(x)) \right] \right\|_{\psi_2} \right)^2 \right] \\
&\leq C_0 \mathbb{E}_x \left[ \left( \left\| \phi(w_i^\top f(x)) \right\|_{\psi_2} + \mathbb{E}_x \left[ \left\| \phi(w_i^\top f(x)) \right\|_{\psi_2} \right] \right)^2 \right],
\end{aligned} \tag{124}$$

where $C_0$ is a numerical constant, the first inequality follows from Exercise 2.7.10 of [65], the second line follows from Jensen's inequality as $\|\cdot\|_{\psi_1}$ is convex, the equality follows from Lemma 2.7.6 of [65], and the last line follows from Jensen's inequality as $\|\cdot\|_{\psi_2}$ is convex.

Recall that $(W)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/n_{L-1})$ and that the Gaussian distribution is rotationally invariant. Thus, $\phi'\left(w_i^\top f(x)\right)$ has the same distribution as $\phi'\left(\tilde{c}(x)\rho_i\right)$, where $(\rho_i)_{i=1}^{n_{L-1}} \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$ and also independent of $\tilde{c}(x)$. Therefore,

$$\begin{aligned}
\left\| \phi\left(w_i^\top f(x)\right) \right\|_{\psi_2} &= \left\| \phi\left(\tilde{c}(x)\rho_i\right) \right\|_{\psi_2} \\
&\leq \left\| \phi\left(\tilde{c}(x)\rho_i\right) - \mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}(x)\rho_i\right)\right] \right\|_{\psi_2} + \left\| \mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}(x)\rho_i\right)\right] \right\|_{\psi_2} \\
&\leq C_1\tilde{c}(x) + C_2\tilde{c}(x) + C_3 \leq C_4(\tilde{c}(x)+1),
\end{aligned} \tag{125}$$

where the first term in the RHS in the second line is bounded by $C_1\tilde{c}(x)$ for Theorem 5.2.2 in [65], and the second term is bounded by $C_2\tilde{c}(x) + C_3$ by following the same proof of Lemma B.11.

Merging together (125) and (124) we get

$$\begin{aligned}
\|Z_i\|_{\psi_1} \leq & C_0 \mathbb{E}_x \left[ \left( C_4\tilde{c}(x) + C_4\tilde{c} + C_5 \right)^2 \right] \\
= & C_6 \mathbb{E}_x[(\tilde{c}(x) - \tilde{c})^2] + C_4\tilde{c}^2 + C_7\tilde{c} + C_8 \\
= & C_6 \mathcal{O}\left(n_l^{-1}\right) + C_4\tilde{c}^2 + C_7\tilde{c} + C_8 \\
\leq & C_9,
\end{aligned} \tag{126}$$

where in the third line we use Lemma B.14.

Hence, the $Z_i$-s are i.i.d. sub-exponential random variables, with sub-exponential norm bounded by a numerical constant. An application of Bernstein inequality (cf. Corollary 2.8.3. in [65]) gives that

$$\mathbb{P}\left( \left| \frac{1}{n_l} \sum_{i=1}^{n_l} Z_i \right| > t \right) \leq 2\exp\left( -c \min\left( \frac{t^2}{C^2}, \frac{t}{C} \right) n_l \right), \tag{127}$$

where $c, C$ are numerical constants. We recall that this probability is intended over $W_l$.

Let's now focus on the first term in the last line of (122), using the shorthand $w = w_1$, to ease the notation. We can rewrite this term as

$$\begin{aligned}
&\mathbb{E}_w \left[ \mathbb{E}_x \left[ \phi^2(w^\top f(x)) \right] - \mathbb{E}_{xy} \left[ \phi(w^\top f(x)) \phi(w^\top f(y)) \right] \right] \\
&\quad = \mathbb{E}_x \mathbb{E}_w \left[ \phi^2(w^\top f(x)) \right] - \mathbb{E}_{xy} \mathbb{E}_w \left[ \phi(w^\top f(x)) \phi(w^\top f(y)) \right].
\end{aligned} \tag{128}$$

The aim of this part of the proof is to show that the quantity in (128) is $\Theta(1)$.

Recall that $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/n_{l-1})$ and that the Gaussian distribution is rotationally invariant. Thus, $\phi\left(w^\top f(x)\right)$ has the same distribution as $\phi\left(\tilde{c}(x)\rho\right)$, where $\rho \sim \mathcal{N}(0,1)$ is independent of $\tilde{c}(x)$.

We therefore have

$$\begin{aligned}
&\left|\mathbb{E}_x\mathbb{E}_w\left[\phi^2\left(w^\top f(x)\right)-\phi^2\left(w^\top f(x)\frac{\tilde{c}}{\tilde{c}(x)}\right)\right]\right| \\
&=\left|\mathbb{E}_x\mathbb{E}_\rho\left[\phi^2(\rho\tilde{c}(x))-\phi^2(\rho\tilde{c})\right]\right| \\
&\leq \mathbb{E}_x\left[C_1|\tilde{c}-\tilde{c}(x)|+C_2\left|\tilde{c}^2-\tilde{c}(x)^2\right|\right] \\
&\leq \mathbb{E}_x\left[C_1|\tilde{c}-\tilde{c}(x)|+C_2(\tilde{c}-\tilde{c}(x))^2+2C_2\tilde{c}\,|\tilde{c}-\tilde{c}(x)|\right] \\
&=\mathcal{O}\left(n_{l-1}^{-1/2}\right),
\end{aligned} \tag{129}$$

where the third line follows from Lemma B.10, and the last passage follows from Lemma B.14. Similarly, we have

$$\begin{aligned}
&\left|\mathbb{E}_{xy}\mathbb{E}_w\left[\phi(w^\top f(x))\phi(w^\top f(y))-\phi\left(w^\top f(x)\frac{\tilde{c}}{\tilde{c}(x)}\right)\phi\left(w^\top f(y)\frac{\tilde{c}}{\tilde{c}(y)}\right)\right]\right| \\
&=\left|\mathbb{E}_{xy}\mathbb{E}_{\rho_1\rho_2}\left[\phi(\rho_1\tilde{c}(x))\phi(\rho_2\tilde{c}(y))-\phi(\rho_1\tilde{c})\phi(\rho_2\tilde{c})\right]\right| \\
&\leq \mathbb{E}_{xy}\left[C_1\left|\tilde{c}(x)-\tilde{c}\right|+C_2\tilde{c}(y)\left|\tilde{c}(x)-\tilde{c}\right|+C_3\left|\tilde{c}(y)-\tilde{c}\right|+C_4\tilde{c}\left|\tilde{c}(y)-\tilde{c}\right|\right] \\
&=\mathcal{O}\left(n_{l-1}^{-1/2}\right),
\end{aligned} \tag{130}$$

where $\rho_1$ and $\rho_2$ indicate two standard Gaussian random variables with correlation $f(x)^\top f(y)/(\|f(x)\|\,\|f(y)\|)$, the third line follows from B.15, and the last passage follows from Lemma B.14. By combining (129) and (130), we have that

$$\left|\mathbb{E}_w\left[\mathbb{E}_x\left[\phi^2(w^\top f(x))\right]-\mathbb{E}_{xy}\left[\phi(w^\top f(x))\phi(w^\top f(y))\right]\right]-\xi\right|=\mathcal{O}\left(n_{l-1}^{-1/2}\right), \tag{131}$$

with

$$\begin{aligned}
\xi :=&\mathbb{E}_x\mathbb{E}_w\left[\phi^2\left(w^\top f(x)\frac{\tilde{c}}{\tilde{c}(x)}\right)\right]-\mathbb{E}_{xy}\mathbb{E}_w\left[\phi\left(w^\top f(x)\frac{\tilde{c}}{\tilde{c}(x)}\right)\phi\left(w^\top f(y)\frac{\tilde{c}}{\tilde{c}(y)}\right)\right] \\
=&\mathbb{E}_{\rho_1}\left[\tilde{\phi}^2(\rho_1)\right]-\mathbb{E}_{xy}\left[\mathbb{E}_{\rho_1\rho_2}\left[\tilde{\phi}(\rho_1)\tilde{\phi}(\rho_2)\right]\right],
\end{aligned} \tag{132}$$

where we have set $\tilde{\phi}(t)=\phi(\tilde{c}t)$. Hence, in order to obtain that the quantity in (128) is $\Theta(1)$, it suffices to prove that $\xi=\Theta(1)$.

As $\phi$ is Lipschitz and $\tilde{c}$ is $\Theta(1)$, $\tilde{\phi}$ is also Lipschitz, which readily implies that $\xi=\mathcal{O}(1)$. We now prove that $\xi=\Omega(1)$. By exploiting the Hermite expansion of $\tilde{\phi}$, we have that

$$\xi=\sum_{i=0}^{\infty}\mu_i^2\left(1-\mathbb{E}_{xy}\left[\left(\frac{f(x)^\top f(y)}{\|f(x)\|\,\|f(y)\|}\right)^i\right]\right), \tag{133}$$

where $\mu_i$ is the $i$-th Hermite coefficient of $\tilde{\phi}$. Note that, since we conditioned on $\tilde{c}=\Theta(1)$, these coefficients are numerical constants. As $\phi$ (and therefore $\tilde{\phi}$) is non constant, there exist $j>0$ such that $\mu_j\neq 0$. Furthermore, we have that the sum in (133) contains only positive terms, as $|f(x)^\top f(y)|\leq\|f(x)\|\cdot\|f(y)\|$ by Cauchy-Schwarz. Therefore, in order to show that $\xi=\Omega(1)$, it suffices to prove that, for all $j>0$,

$$\mathbb{E}_{xy}\left[\left(\frac{\left|f(x)^\top f(y)\right|}{\|f(x)\|\,\|f(y)\|}\right)^j\right]\leq C_0<1, \tag{134}$$

where $C_0$ is an absolute constant strictly smaller than 1. Furthermore, (134) is implied by the following:

$$\mathbb{P}_{xy}\left(\frac{\left|f(x)^\top f(y)\right|}{\|f(x)\|\,\|f(y)\|}\leq C_1\right)\geq c_1, \tag{135}$$

where $C_1<1$ and $c_1>0$ are numerical constants.

By writing $f(x)$ as $\tilde{f}(x)+f$ and $f(y)$ as $\tilde{f}(y)+f$, we have

$$\begin{aligned}\frac{\left|f(x)^\top f(y)\right|}{\|f(x)\|\,\|f(y)\|} &= \frac{\left|(\tilde{f}(x)+f)^\top(\tilde{f}(y)+f)\right|}{\left\|\tilde{f}(x)+f\right\|\left\|\tilde{f}(y)+f\right\|} \\ &\le \frac{\left|\tilde{f}(x)^\top(\tilde{f}(y)+f)\right|+\left|f^\top\tilde{f}(y)\right|+\|f\|^2}{\min_{z\in\{x,y\}}\left\|\tilde{f}(z)+f\right\|^2} \\ &\le \frac{\left|\tilde{f}(x)^\top f(y)\right|+\left|f^\top\tilde{f}(y)\right|+\|f\|^2}{\min_{z\in\{x,y\}}\left(\left\|\tilde{f}(z)\right\|^2-2\left|f^\top\tilde{f}(z)\right|\right)+\|f\|^2}.\end{aligned} \tag{136}$$

Let us provide bounds on the various terms appearing in (136):

(i) Part (d) of the conditioning (cf. (121)) gives that

$$\mathbb{P}_{xy}\left(\min_{z\in\{x,y\}}\left\|\tilde{f}(z)\right\|_2^2 \ge cn_{l-1}\right)\ge\frac{1}{4}, \tag{137}$$

for some numerical constant $c>0$.

(ii) Part (b) of the conditioning gives that

$$\|f\|_2^2 \le C' n_{l-1}. \tag{138}$$

(iii) Part (a) of the conditioning gives that $\|f_{l-1}(x)\|_{\mathrm{Lip}}=\mathcal{O}(1)$, and part (b) of the conditioning gives that $\mathbb{E}_y\left[\|f_{l-1}(y)\|\right]=\Theta(\sqrt{n_{l-1}})$. Hence, as $y\sim P_X$, Assumption 2.2 implies that

$$\|f_{l-1}(y)\|=\Theta(\sqrt{n_{l-1}}), \tag{139}$$

with probability at least $1-2\exp(-cn_{l-1})$ over $y$, where $c$ is a numerical constant. Furthermore, by recalling that $\mathbb{E}_x[\tilde{f}(x)]=0$ and using again Assumption 2.2, we have that, for any fixed vector $u$,

$$\mathbb{P}_x(|\tilde{f}(x)^\top u|>t)\le 2e^{-c_0t^2/\|u\|_2^2}, \tag{140}$$

where $c_0$ is another numerical constant. Since $x$ and $y$ are independent, (140) implies that

$$\mathbb{P}_x(|\tilde{f}(x)^\top f(y)|>n_{l-1}^{3/4})\le 2e^{-c_0n_{l-1}^{3/2}/\|f(y)\|_2^2}\le 2e^{-c_2n_{l-1}^{1/2}}, \tag{141}$$

where the first inequality holds for every $y$, and the second inequality holds with probability at least $1-2\exp(-cn_{l-1})$ over $y$ by (139). As a result, we have

$$\mathbb{P}_{xy}(|\tilde{f}(x)^\top f(y)|>n_{l-1}^{3/4})\le 2e^{-c_2\sqrt{n_{l-1}}}+2e^{-cn_{l-1}}\le 4e^{-c_3\sqrt{n_{l-1}}}. \tag{142}$$

(iv) By setting $t=n_{l-1}^{3/4}$ and $u=f$ into (140), we obtain

$$\mathbb{P}_y(|\tilde{f}(y)^\top f|>n_{l-1}^{3/4})\le 2e^{-c_4\sqrt{n_{l-1}}}, \tag{143}$$

where $c_4$ is a numerical constant and we have also used (138).

(v) Finally, as $x$ and $y$ are independent, (143) implies that

$$\mathbb{P}_{xy}(\max_{z\in\{x,y\}}|\tilde{f}(z)^\top f|>n_{l-1}^{3/4})\le 4e^{-c_4\sqrt{n_{l-1}}}. \tag{144}$$

By plugging into (136) the bounds (137), (142), (143) and (144), we obtain that

$$\mathbb{P}_{x,y}\left(\frac{\left|f(x)^\top f(y)\right|}{\|f(x)\|\,\|f(y)\|}\le\frac{2n_{l-1}^{3/4}+\|f\|_2^2}{c\,n_{l-1}-2n_{l-1}^{3/4}+\|f\|_2^2}\right)\ge 1/4-10e^{-c_5\sqrt{n_{l-1}}}. \tag{145}$$

By using also (138), we have that (135) readily follows from (145). From (135), we have that $\xi=\Theta(1)$. Hence, the quantity in (128) is $\Theta(1)$ and in particular it is lower bounded by a numerical constant, call it $C_0$.

By setting $t = C_0/2$ in (127), we conclude that

$$\mathbb{E}_x \left[ \| f_l(x) - \mathbb{E}_x \left[ f_l(x) \right] \|_2^2 \right] = \Theta(n_l), \tag{146}$$

with probability at least $1 - 2\exp(-cn_{l-1})$ over $W_l$, where $c$ is an absolute constant. By taking into account the conditioning made at the beginning of the proof over the space $(W_k)_{k=1}^{l-1}$, we obtain that (146) holds with probability at least $1 - 5C' \exp(-cn_{n-1}) - 2\exp(-cn_{n-1}) \geq 1 - C\exp(-cn_{n-1})$ over $(W_k)_{k=1}^{l}$, where $C$ is a numerical constant, which concludes the proof of (117).

Finally, we prove (118) and (119), again for layer $l$. By Lemma B.2, we have that $\| f_l(x) \|_{\mathrm{Lip}} = \mathcal{O}(1)$, with probability at least $1 - C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{l}$. By conditioning on this event, we also have that $\| \| f_l(x) - \mathbb{E}_x[f_l(x)] \|_2 \|_{\mathrm{Lip}} = \mathcal{O}(1)$. Furthermore, we condition on a realization of $(W_k)_{k=1}^{l}$ such that (117) holds.

To obtain (118), we apply Jensen's inequality and Lemma B.13, which give that

$$\mathbb{E}_x \left[ \| f_l(x) - \mathbb{E}_x[f_l(x)] \|_2 \right] = \Theta(\sqrt{n_l}), \tag{147}$$

with probability at least $1 - C'\exp(-cn_{L-1}) - 5C'\exp(-cn_{n-1}) - 2\exp(-cn_{n-1}) \geq 1 - C\exp(-cn_{n-1})$ over $(W_k)_{k=1}^{l}$.

To obtain (119), we condition on a realization of $(W_k)_{k=1}^{l}$ such that $\| \| f_l(x) - \mathbb{E}_x[f_l(x)] \|_2 \|_{\mathrm{Lip}} = \mathcal{O}(1)$ and (118) holds. Then, by Assumption 2.2, we have that

$$\begin{aligned} \mathbb{P}_x \left( \left| \| f_l(x) - \mathbb{E}_x[f_l(x)] \|_2 - \mathbb{E}_x \left[ \| f_l(x) - \mathbb{E}_x[f_l(x)] \|_2 \right] \right| > \mathbb{E}_x \left[ \| f_l(x) - \mathbb{E}_x[f_l(x)] \|_2 \right] / 2 \right) \\ \leq 2\exp(-c_1 n_l) \leq 2\exp(-cn_{L-1}), \end{aligned} \tag{148}$$

where $c$ is a numerical constant. This gives that

$$\| f_l(x) - \mathbb{E}_x[f_l(x)] \| = \Theta(\sqrt{n_l}), \tag{149}$$

with probability at least $1 - 6C'\exp(-cn_{n-1}) - 2\exp(-cn_{n-1}) - 2\exp(-cn_{n-1}) \geq 1 - C\exp(-cn_{n-1})$ over $x$ and $(W_k)_{k=1}^{l}$, which concludes the proof. □

**Lemma C.5** ($\ell_2$ norms of centered backpropagation)**.** *Let $x \sim P_X$. Then, we have*

$$\| D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[ D_L \phi'(g_{L-1}(x)) \right] \|_2 = \Theta(\sqrt{n_{L-1}}), \tag{150}$$

*with probability at least $1 - 10\exp(-c\log^2 n_{L-1}) - C\exp(-cn_{L-1})$ over $x$ and $(W_k)_{k=1}^{L}$ over $(W_k)_{k=1}^{l}$ and $x$.*

*Proof.* An application of Lemma C.4 for $l = L - 2$ gives that

$$\| f_{L-2}(x) - \mathbb{E}_x \left[ f_{L-2}(x) \right] \|_2 = \Theta(\sqrt{n_{L-2}}). \tag{151}$$

with probability at least $1 - C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L-2}$ and $x$.

To ease the notation, we use the shorthands $f(x) := f_{L-2}(x)$, $f = \mathbb{E}_x[f(x)]$, $\tilde{f}(x) = f(x) - f$, $W := W_{L-1}$ and $w_i = W_{:i}$. We also define $\tilde{c}(x) = \beta_l \| f(x) \| / \sqrt{n_{l-1}}$ and $\tilde{c} = \mathbb{E}_x[\tilde{c}(x)]$.

As in Lemma C.4, we condition on the 3 events (a)-(c), which jointly happen with probability at least $1 - 4C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L-2}$. Note that, to condition on the event (c), we use (151).

Now, we can write

$$\begin{aligned} &\mathbb{E}_x \left[ \| D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[ D_L \phi'(g_{L-1}(x)) \right] \|_2^2 \right] \\ &= \mathbb{E}_x \left[ \left\| D_L \phi'(W^\top f_{L-2}(x)) - \mathbb{E}_x \left[ D_L \phi'(W^\top f_{L-2}(x)) \right] \right\|_2^2 \right] \\ &= n_{L-1} \left( \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} (D_L)_{ii}^2 \mathbb{E}_x \left[ \left( \phi'(w_i^\top f(x)) - \mathbb{E}_x \left[ \phi'(w_i^\top f(x)) \right] \right)^2 \right] \right) \\ &= n_{L-1} \left( \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} (D_L)_{ii}^2 \mathbb{E}_{x w_1} \left[ \left( \phi'(w_1^\top f(x)) - \mathbb{E}_x \left[ \phi'(w_1^\top f(x)) \right] \right)^2 \right] + \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} Z_i \right), \end{aligned} \tag{152}$$

where we use that the $w_i$-s are identically distributed and we have defined the independent, mean-0 random variables

$$\begin{aligned} Z_i =&(D_L)_{ii}^2 \mathbb{E}_x \left[ \left( \phi'(w_i^\top f(x)) - \mathbb{E}_x \left[ \phi'(w_i^\top f(x)) \right] \right)^2 \right] \\ &- (D_L)_{ii}^2 \mathbb{E}_{w_1 x} \left[ \left( \phi'(w_1^\top f(x)) - \mathbb{E}_x \left[ \phi'(w_1^\top f(x)) \right] \right)^2 \right]. \end{aligned} \tag{153}$$

Note that in the definition of $Z_i$ the randomness comes only from $w_i$ and $(D_L)_{ii}$, since we are conditioning on $(W_k)_{k=1}^{L-2}$.

If we fix the $(D_L)_{ii}$-s and follow the same argument in (124)-(126) (cf. the proof of Lemma C.4), we have

$$\|Z_i\|_{\psi_1} \leq C_0 (D_L)_{ii}^2, \tag{154}$$

where $C_0$ is a numerical constant and we have used that $\phi'$ is Lipschitz. Let $\mathcal{E}_{\text{bad}}$ be the event s.t. $\max_i (D_L)_{ii}^2 > \log^2 n_{L-1}$. Then, by following the same argument as in Lemma B.3, we have that

$$\mathbb{P}(\mathcal{E}_{\text{bad}}) \leq 2 \exp(-c \log^2 n_{L-1}). \tag{155}$$

Hence, by conditioning on $\mathcal{E}_{\text{bad}}^c$, we have that

$$\max_i \|Z_i\|_{\psi_1} \leq C_0 \log^2 n_{L-1}. \tag{156}$$

By applying Bernstein inequality (cf. Corollary 2.8.3. in [65]), we get

$$\mathbb{P}\left( \left| \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} Z_i \right| > \frac{1}{\sqrt[4]{n_{L-1}}} \,\middle|\, \mathcal{E}_{\text{bad}}^c \right) \leq 2 \exp\left( -c \frac{\sqrt{n_{L-1}}}{\log^4 n_{L-1}} \right), \tag{157}$$

for some numerical constant $c$. By combining (155) and (157), we obtain that

$$\begin{aligned} \mathbb{P}\left( \left| \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} Z_i \right| > \frac{1}{\sqrt[4]{n_{L-1}}} \right) &\leq 2 \exp\left( -c \frac{\sqrt{n_{L-1}}}{\log^4 n_{L-1}} \right) + 2 \exp\left( -c \log^2 n_{L-1} \right) \\ &\leq 4 \exp\left( -c \log^2 n_{L-1} \right), \end{aligned} \tag{158}$$

where this probability is over $W_{L-1}$ and $W_L$.

Let's now focus on the first term in the last line of (152). In particular, we have that

$$\xi = \mathbb{E}_{x w_1} \left[ \left( \phi'(w_1^\top f(x)) - \mathbb{E}_x \left[ \phi'(w_1^\top f(x)) \right] \right)^2 \right] = \Theta(1). \tag{159}$$

This can be proven by following the same argument in (129)-(145) (cf. the proof of Lemma C.4), as $\phi'$ is Lipschitz and non-constant.

Next, we re-write (152) as

$$\begin{aligned} &\mathbb{E}_x \left[ \| D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[ D_L \phi'(g_{L-1}(x)) \right] \|_2^2 \right] \\ &= n_{L-1} \left( \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} (D_L)_{ii}^2 \mathbb{E}_{x w_1} \left[ \left( \phi'(w_1^\top f(x)) - \mathbb{E}_x \left[ \phi'(w_1^\top f(x)) \right] \right)^2 \right] + \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} Z_i \right) \\ &= n_{L-1} \left( \xi \mathbb{E}_{W_L} \left[ (D_L)_{11}^2 \right] + \xi \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} \tilde{Z}_i + \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} Z_i \right) \\ &= n_{L-1} \left( \xi \beta_L^2 + \xi \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} \tilde{Z}_i + \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} Z_i \right), \end{aligned} \tag{160}$$

where we have defined the independent, mean-0, sub-exponential random variables

$$\tilde{Z}_i = (D_L)_{ii}^2 - \mathbb{E}_{W_L} \left[ (D_L)_{11}^2 \right]. \tag{161}$$

Since the $(D_L)_{ii}$-s are standard Gaussian, we have

$$\left\| \tilde{Z}_i \right\|_{\psi_1} \leq \left\| (D_L)_{ii}^2 \right\|_{\psi_1} = \| (D_L)_{ii} \|_{\psi_2}^2 = C_1. \tag{162}$$

Hence, another application of Bernstein inequality (cf. Corollary 2.8.3. in [65]) allows us to conclude that

$$\left| \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} \tilde{Z}_i \right| = \mathcal{O}\left( n_{L-1}^{-1/4} \right), \tag{163}$$

with probability at least $1 - 2\exp(-c\sqrt{n_{L-1}})$ over $W_L$.

Thus, by using (158) and (163), and taking into account the initial conditioning over $(W_k)_{k=1}^{L-2}$, we conclude that

$$\mathbb{E}_x \left[ \left\| D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[ D_L \phi'(g_{L-1}(x)) \right] \right\|_2^2 \right] = \Theta(n_{L-1}), \tag{164}$$

with probability at least $1 - 6\exp(-c\log^2 n_{L-1}) - 6C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L}$.

Proceeding in a similar fashion as in Lemma C.4, we apply Lemma B.4, which gives that $\|D_L \phi'(g_{L-1}(x))\|_{\mathrm{Lip}} = \mathcal{O}(\log n_{L-1})$, with probability at least $1 - 2\exp(-c\log^2 n_{L-1}) - C'\exp(-n_{L-1})$ over $(W_k)_{k=1}^{L}$. By conditioning on this event, we also have that $\left\| \|D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[ D_L \phi'(g_{L-1}(x)) \right]\|_2 \right\|_{\mathrm{Lip}} = \mathcal{O}(\log n_{L-1})$. Furthermore, we condition on a realization of $(W_k)_{k=1}^{L}$ such that (164) holds.

We can now apply Jensen's inequality and Lemma B.13, to obtain that

$$\mathbb{E}_x \left[ \|D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[ D_L \phi'(g_{L-1}(x)) \right]\|_2 \right] = \Theta(\sqrt{n_l}), \tag{165}$$

with probability at least $1 - 8\exp(-c\log^2 n_{L-1}) - 7C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L}$.

Finally, we condition on a realization of $(W_k)_{k=1}^{L}$ such that $\left\| \|D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[ D_L \phi'(g_{L-1}(x)) \right]\|_2 \right\|_{\mathrm{Lip}} = \mathcal{O}(\log n_{L-1})$ and (165) hold. Then, by Assumption 2.2, we have that

$$\begin{aligned} \mathbb{P}_x \big( \big| \|D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[ D_L \phi'(g_{L-1}(x)) \right]\|_2 - \mathbb{E}_x \left[ \|D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[ D_L \phi'(g_{L-1}(x)) \right]\|_2 \right] \big| \\ > \mathbb{E}_x \left[ \|D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[ D_L \phi'(g_{L-1}(x)) \right]\|_2 \right] / 2 \big) \\ \leq 2\exp^{(}-cn_{L-1}). \end{aligned} \tag{166}$$

This gives that

$$\|D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[ D_L \phi'(g_{L-1}(x)) \right]\|_2 = \Theta(\sqrt{n_{L-1}}), \tag{167}$$

with probability at least $1 - 10\exp(-c\log^2 n_{L-1}) - 8C'\exp(-cn_{L-1})$ over $x$ and $(W_k)_{k=1}^{L}$. This concludes the proof. □

# D Proofs for Part 1: Centering

## D.1 Step (a): Centering $F_{L-2}$ and $B_{L-1}$

**Lemma D.1** (Centering $F_{L-2}$ and $B_{L-1}$)**.** *Consider the setting of Theorem 3.1, let $F_{L-2} \in \mathbb{R}^{N \times n_{L-2}}$ be the feature matrix at layer $L-2$, and let $B_{L-1}$ contain the backpropagation terms from layer $L-1$, i.e. $(B_{L-1})_{i:} = D_L \phi'(g_{L-1}(x_i))$. Let $J_{L-2} J_{L-2}^\top = F_{L-2} F_{L-2}^\top \circ B_{L-1} B_{L-1}^\top$ and $\tilde{J}_{FB} \tilde{J}_{FB}^\top = \tilde{F}_{L-2} \tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1} \tilde{B}_{L-1}^\top$, where $\tilde{F}_{L-2} = F_{L-2} - \mathbb{E}_X[F_{L-2}]$ and $\tilde{B}_{L-1} = B_{L-1} - \mathbb{E}_X[B_{L-1}]$. Then, we have that*

$$\lambda_{\min}\left( J_{L-2} J_{L-2}^\top \right) \geq \lambda_{\min}\left( \tilde{J}_{FB} \tilde{J}_{FB}^\top \right) - o(n_{L-1} n_{L-2}), \tag{168}$$

*with probability at least $1 - C\exp(-cn_{L-1}) - 4\exp(-c\log^2 N) - 8\exp(-c\log^2 n_{L-1})$ over $(W_k)_{k=1}^{L}$ and $(x_i)_{i=1}^{N} \sim_{\text{i.i.d.}} P_X$, where $c$ and $C$ are numerical constants.*

*Proof.* By Lemma B.2 and Lemma B.4, we have that $\|f_{L-2}\|_{\mathrm{Lip}} = \mathcal{O}(1)$ and $\|D_L \phi'(g_{L-1}(x))\|_{\mathrm{Lip}} = \mathcal{O}(\log n_{L-1})$ with probability $1 - C\exp(-n_{L-1}) - 2\exp(-c\log^2 n_{L-1})$ over $(W_k)_{k=1}^{L}$. We will condition on these events for the rest of the proof.

Let's define $\tilde{J}_F\tilde{J}_F^\top = \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ B_{L-1}B_{L-1}^\top$. We can now re-write the quantity $J_{L-2}J_{L-2}^\top$ as follows:

$$\begin{aligned}
J_{L-2}J_{L-2}^\top &= \tilde{J}_F\tilde{J}_F^\top + \mathbb{E}[F_{L-2}]\mathbb{E}[F_{L-2}]^\top \circ B_{L-1}B_{L-1}^\top \\
&\quad + \left((F_{L-2} - \mathbb{E}[F_{L-2}])\mathbb{E}[F_{L-2}]^\top + \mathbb{E}[F_{L-2}](F_{L-2} - \mathbb{E}[F_{L-2}])^\top\right) \circ B_{L-1}B_{L-1}^\top \\
&= \tilde{J}_F\tilde{J}_F^\top + \|\nu\|_2^2 B_{L-1}B_{L-1}^\top + \left(\Lambda\mathbf{1}\mathbf{1}^\top + \mathbf{1}\mathbf{1}^\top\Lambda\right) \circ B_{L-1}B_{L-1}^\top \\
&= \tilde{J}_F\tilde{J}_F^\top + \left(\|\nu\|_2 \mathbf{1} + \frac{\Lambda\mathbf{1}}{\|\nu\|_2}\right)\left(\|\nu\|_2 \mathbf{1} + \frac{\Lambda\mathbf{1}}{\|\nu\|_2}\right)^\top \circ B_{L-1}B_{L-1}^\top \\
&\quad - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top \\
&\succeq \tilde{J}_F\tilde{J}_F^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top,
\end{aligned} \tag{169}$$

where $\nu = \mathbb{E}_{x_i}[(F_{L-2})_{i:}] \in \mathbb{R}^{n_{L-2}}$ (independent on $i$, since the $x_i$-s are i.i.d.), $\Lambda$ is a diagonal matrix such that $\Lambda_{ii} = \nu^\top(F_{L-2})_{i:} - \|\nu\|_2^2 =: \mu(x_i)$, and $\mathbf{1} \in \mathbb{R}^N$ is a vector full of ones. The last step is justified since the Hadamard product of PSD matrices is PSD by the Schur product theorem. Notice that we are assuming $\|\nu\|_2 \neq 0$. In fact, if $\|\nu\|_2 = 0$, then we immediately have that $J = \tilde{J}_F$.

Expanding in an analogous way the term $\tilde{J}_F\tilde{J}_F^\top$, we get

$$\begin{aligned}
J_{L-2}J_{L-2}^\top &\succeq \tilde{J}_F\tilde{J}_F^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top \\
&= \tilde{J}_{FB}\tilde{J}_{FB}^\top + \left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)^\top \circ \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \\
&\quad - \frac{\Gamma\mathbf{1}\mathbf{1}^\top\Gamma}{\|\eta\|_2^2} \circ \tilde{F}_{L-2}\tilde{F}_{L-2}^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top \\
&\succeq \tilde{J}_{FB}\tilde{J}_{FB}^\top + \left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)^\top \circ \tilde{F}_{L-2}\left(\frac{\nu\nu^\top}{\|\nu\|_2^2}\right)\tilde{F}_{L-2}^\top \\
&\quad - \frac{\Gamma\mathbf{1}\mathbf{1}^\top\Gamma}{\|\eta\|_2^2} \circ \tilde{F}_{L-2}\tilde{F}_{L-2}^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top
\end{aligned} \tag{170}$$

where $\eta = \mathbb{E}_{x_i}[(B_{L-1})_{i:}] \in \mathbb{R}^{n_{L-1}}$ (independent on $i$, since the $x_i$-s are i.i.d.), $\Gamma$ is a diagonal matrix such that $\Gamma_{ii} = \eta^\top(B_{L-1})_{i:} - \|\eta\|_2^2 =: \zeta(x_i)$. The last step is justified by the fact that the following matrix is PSD

$$\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)^\top \circ \tilde{F}_{L-2}\left(I - \frac{\nu\nu^\top}{\|\nu\|_2^2}\right)\tilde{F}_{L-2}^\top,$$

since it is the Hadamard product of two PSD matrices. Notice that we are assuming $\|\eta\|_2 \neq 0$. In fact, if $\|\eta\|_2 = 0$, then we immediately have that $\tilde{J}_F = \tilde{J}_{FB}$.

Taking into account the following relations

$$\Lambda\mathbf{1} = \tilde{F}_{L-2}\nu, \qquad \Gamma\mathbf{1} = \tilde{B}_{L-1}\eta, \qquad \mathbb{E}_X[B_{L-1}] = \mathbf{1}\eta^\top, \tag{171}$$

we can simplify the second and the fourth term of the RHS of equation (170) as follows

$$\begin{aligned}
&\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)^\top \circ \tilde{F}_{L-2}\left(\frac{\nu\nu^\top}{\|\nu\|_2^2}\right)\tilde{F}_{L-2}^\top - B_{L-1}B_{L-1}^\top \circ \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \\
&= \left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)^\top \circ \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \\
&\quad - \left(\mathbf{1}\eta^\top + \tilde{B}_{L-1}\right)\left(\mathbf{1}\eta^\top + \tilde{B}_{L-1}\right)^\top \circ \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \\
&= \left(\frac{\Gamma\mathbf{1}\mathbf{1}^\top\Gamma}{\|\eta\|_2^2} - \tilde{B}_{L-1}\tilde{B}_{L-1}^\top\right) \circ \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \succeq -\tilde{B}_{L-1}\tilde{B}_{L-1}^\top \circ \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2}.
\end{aligned} \tag{172}$$

Merging this last relation with (170) we get

$$J_{L-2}J_{L-2}^\top \succeq \tilde{J}_{FB}\tilde{J}_{FB}^\top - \frac{\Gamma \mathbf{1}\mathbf{1}^\top \Gamma}{\|\eta\|_2^2} \circ \tilde{F}_{L-2}\tilde{F}_{L-2}^\top - \frac{\Lambda \mathbf{1}\mathbf{1}^\top \Lambda}{\|\nu\|_2^2} \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top$$
$$= \tilde{J}_{FB}\tilde{J}_{FB}^\top - \left(\frac{\Gamma}{\|\eta\|_2}\tilde{F}_{L-2}\right)\left(\frac{\Gamma}{\|\eta\|_2}\tilde{F}_{L-2}\right)^\top - \left(\frac{\Lambda}{\|\nu\|_2}\tilde{B}_{L-1}\right)\left(\frac{\Lambda}{\|\nu\|_2}\tilde{B}_{L-1}\right)^\top. \tag{173}$$

Note that $\|\mu\|_{\mathrm{Lip}} \leq \|f_{L-2}\|_{\mathrm{Lip}} \|\nu\|_2$, and that $\mathbb{E}_{x_i}[\mu(x_i)] = 0$ for all $i \in [N]$. Thus, by using Assumption 2.2 on $x_i$ and exploiting the initial conditioning on the weights, we have that

$$\mathbb{P}(|\mu(x_i)|/\|\nu\|_2 > t) < 2\exp(-ct^2), \tag{174}$$

where the probability is intended over $x_i \sim P_X$. Thus, following the same argument of Lemma B.3, the last relation implies that

$$\|\Lambda/\|\nu\|_2\|_{\mathrm{op}} = \mathcal{O}(\log N), \tag{175}$$

with probability at least $1 - 2\exp(-c\log^2 N)$, where $c$ is a numerical constant, and the probability is intended over $\{x_i\}_{i=1}^N$. This implies that

$$\left\|\left(\frac{\Lambda}{\|\nu\|_2}\tilde{B}_{L-1}\right)\left(\frac{\Lambda}{\|\nu\|_2}\tilde{B}_{L-1}\right)^\top\right\|_{\mathrm{op}}$$
$$\leq \left\|\frac{\Lambda}{\|\nu\|_2}\right\|_{\mathrm{op}}^2 \left\|\tilde{B}_{L-1}\tilde{B}_{L-1}^\top\right\|_{\mathrm{op}} = \mathcal{O}\left((N + n_{L-1}) \cdot \log^2 N \cdot \log^2 n_{L-1}\right) = o(n_{L-2}n_{L-1}), \tag{176}$$

where the second equality is justified by Lemma B.9, and the last by Lemma B.1. This result holds with probability $1 - C\exp(-cn_{L-1}) - 2\exp(-c\log^2 N) - 4\exp(-c\log^2 n_{L-1})$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^N$.

Note that $\|\zeta\|_{\mathrm{Lip}} \leq \|D_L\phi'(g_{L-1}(x))\|_{\mathrm{Lip}} \|\eta\|_2$, and that $\mathbb{E}_{x_i}[\zeta(x_i)] = 0$ for all $i \in [N]$. Thus, by using Assumption 2.2 on $x_i$ and exploiting the initial conditioning on the weights, we have that

$$\mathbb{P}(|\zeta(x_i)|/\|\eta\|_2 > t \cdot \log n_{L-1}) < 2\exp(-ct^2), \tag{177}$$

where the probability is intended over $x_i \sim P_X$. Thus, following the same argument of Lemma B.3, the last relation implies that

$$\|\Gamma/\|\eta\|_2\|_{\mathrm{op}} = \mathcal{O}(\log N \cdot \log n_{L-1}), \tag{178}$$

with probability at least $1 - 2\exp(-c\log^2 N)$, where $c$ is a numerical constant, and the probability is intended over $\{x_i\}_{i=1}^N$. This implies that

$$\left\|\left(\frac{\Gamma}{\|\eta\|_2}\tilde{F}_{L-2}\right)\left(\frac{\Gamma}{\|\eta\|_2}\tilde{F}_{L-2}\right)^\top\right\|_{\mathrm{op}}$$
$$\leq \left\|\frac{\Gamma}{\|\eta\|_2}\right\|_{\mathrm{op}}^2 \left\|\tilde{F}_{L-2}\tilde{F}_{L-2}^\top\right\|_{\mathrm{op}} = \mathcal{O}\left((N + n_{L-2}) \cdot \log^2 N \cdot \log^2 n_{L-1}\right) = o(n_{L-2}n_{L-1}), \tag{179}$$

where the second equality is justified by Lemma B.8, and the last by Lemma B.1. This result holds with probability $1 - C\exp(-cn_{L-1}) - 2\exp(-c\log^2 N) - 4\exp(-c\log^2 n_{L-1})$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^N$.

By merging (176) and (179) with (173), we readily obtain the desired result. □

### D.2 Step (b): Centering everything

**Lemma D.2** (Centering everything)**.** *Consider the setting of Theorem 3.1, let $F_{L-2} \in \mathbb{R}^{N \times n_{L-2}}$ be the feature matrix at layer $L-2$, and let $B_{L-1}$ contain backpropagation terms from layer $L-1$, i.e. $(B_{L-1})_{i:} = D_L\phi'(g_{L-1}(x_i))$. Let $\tilde{J}_{FB}\tilde{J}_{FB}^\top = \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top$ and $\tilde{J}_{L-2}\tilde{J}_{L-2}^\top =$*

$\tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top - \mathbb{E}_X[\tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top]$, *where* $\tilde{F}_{L-2} = F_{L-2} - \mathbb{E}_X[F_{L-2}]$ *and* $\tilde{B}_{L-1} = B_{L-1} - \mathbb{E}_X[B_{L-1}]$. *Then, we have that*

$$\lambda_{\min}\left(\tilde{J}_{FB}\tilde{J}_{FB}^\top\right) \geq \lambda_{\min}\left(\tilde{J}_{L-2}\tilde{J}_{L-2}^\top\right) - o(n_{L-1}n_{L-2}), \tag{180}$$

*with probability at least* $1 - C\exp(-n_{L-1}) - 2\exp(-c\log^2 n_{L-1}) - 2\exp(-c\log^2 N)$ *over* $(W_k)_{k=1}^L$ *and* $(x_i)_{i=1}^N$.

*Proof.* Note that the $i$-th row of $\tilde{J}_{FB}$ is now in the form

$$(\tilde{J}_{FB})_{i:} = \tilde{f}_{L-2}(x_i) \otimes (D_L\tilde{\phi}'(g_{L-1}(x_i))), \tag{181}$$

where we recall that $\tilde{f}_{L-2}(x_i) = f_{L-2}(x_i) - \mathbb{E}_{x_i}[f_{L-2}(x_i)]$ and $\tilde{\phi}'(g_{L-1}(x_i)) = \phi'(g_{L-1}(x_i)) - \mathbb{E}[\phi'(g_{L-1}(x_i))]$. Furthermore, $(\tilde{J}_{L-2})_{i:} = (\tilde{J}_{FB})_{i:} - \mathbb{E}_{x_i}[(\tilde{J}_{FB})_{i:}]$. Then, by following similar passages as in (169), we have

$$\begin{aligned}
\tilde{J}_{FB}\tilde{J}_{FB}^\top &= \tilde{J}_{L-2}\tilde{J}_{L-2}^\top + \mathbb{E}[\tilde{J}_{FB}]\mathbb{E}[\tilde{J}_{FB}]^\top \\
&\quad + (\tilde{J}_{FB} - \mathbb{E}[\tilde{J}_{FB}])\mathbb{E}[\tilde{J}_{FB}]^\top + \mathbb{E}[\tilde{J}_{FB}](\tilde{J}_{FB} - \mathbb{E}[\tilde{J}_{FB}])^\top \\
&= \tilde{J}_{L-2}\tilde{J}_{L-2}^\top + \|A\|_F^2\,\mathbf{1}\mathbf{1}^\top + \Lambda\mathbf{1}\mathbf{1}^\top + \mathbf{1}\mathbf{1}^\top\Lambda \\
&= \tilde{J}_{L-2}\tilde{J}_{L-2}^\top + \left(\|A\|_F\,\mathbf{1} + \frac{\Lambda\mathbf{1}}{\|A\|_F}\right)\left(\|A\|_F\,\mathbf{1} + \frac{\Lambda\mathbf{1}}{\|A\|_F}\right)^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|A\|_F^2} \\
&\succeq \tilde{J}_{L-2}\tilde{J}_{L-2}^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|A\|_F^2},
\end{aligned} \tag{182}$$

where we have defined

$$A = \mathbb{E}_{x_i}\left[\tilde{f}_{L-2}(x_i)(D_L\tilde{\phi}'(g_{L-1}(x_i)))^\top\right], \tag{183}$$

which is independent on $i$ (as the $x_i$-s are i.i.d.), and $\Lambda$ is a diagonal matrix that contains in the $i$-th position

$$\Lambda_{ii} = \tilde{f}_{L-2}(x_i)^\top A(D_L\tilde{\phi}'(g_{L-1}(x_i))) - \mathbb{E}_{x_i}\left[\tilde{f}_{L-2}(x_i)^\top A(D_L\tilde{\phi}'(g_{L-1}(x_i)))\right]. \tag{184}$$

The last passage of (182) is true since we are subtracting a PSD matrix.

An application of Lemma B.2 gives that $\|f_{L-2}(x_i)\|_{\mathrm{Lip}}$ and $\|\phi'(g_{L-1}(x_i))\|_{\mathrm{Lip}}$ are upper bounded by a numerical constant both with probability at least $1 - C\exp(-n_{L-1})$ over $(W_k)_{k=1}^{L-1}$. Let us condition on this event on the probability space of $(W_k)_{k=1}^{L-1}$. Then, we can apply Lemma B.5 with $u(x) := \tilde{f}_{L-2}(x)$ and $v(x) := \tilde{\phi}'(g_{L-1}(x))$, which implies that

$$\|\Lambda_{ii}\|_{\psi_1} < C\,\|AD_L\|_F \leq C\,\|A\|_F\,\|D_L\|_{\mathrm{op}} = \mathcal{O}\left(\log n_{L-1}\right)\|A\|_F. \tag{185}$$

In (185), $C$ is a numerical constant and the last equality holds with probability at least $1 - 2\exp(-c\log^2 n_{L-1})$ over $W_L$ by Lemma B.3. Thus,

$$\mathbb{P}(|\Lambda_{ii}|/\|A\|_F > t \cdot \log n_{L-1}) < 2\exp(-ct), \tag{186}$$

where the probability is intended over $x_i \sim P_X$ and $c$ is a numerical constant. Thus, following the same argument of Lemma B.3, the last relation implies that

$$\|\Lambda/\|A\|_F\|_{\mathrm{op}} = \mathcal{O}\left(\log^2 N \cdot \log n_{L-1}\right), \tag{187}$$

with probability at least $1 - 2\exp(-c\log^2 N)$, where $c$ is a numerical constant, and the probability is intended over $\{x_i\}_{i=1}^N$.

Thus, with probability $1 - C\exp(-n_{L-1}) - 2\exp(-c\log^2 n_{L-1}) - 2\exp(-c\log^2 N)$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^N$, we have

$$\left\|\frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|A\|_F^2}\right\|_{\mathrm{op}} \leq \|\mathbf{1}\mathbf{1}\|_{\mathrm{op}}\left\|\frac{\Lambda}{\|A\|_F}\right\|_{\mathrm{op}}^2 = \mathcal{O}\left(N \cdot \log^4 N \cdot \log^2 n_{L-1}\right) = o(n_{L-2}n_{L-1}), \tag{188}$$

where the last equality is a consequence of Lemma B.1.

The desired result follows from merging (188) with (182). □

# E Proofs for Part 2: Bounding the Centered Jacobian

## E.1 $\ell_2$ and sub-exponential norms of centered Jacobian

We start by providing an upper bound on the quantity $\mathbb{E}_x\left[\tilde{f}_{L-2}(x)(D_L\tilde{\phi}'(g_{L-1}(x)))^\top\right]$. This preliminary result will be useful when bounding the $\ell_2$ norm of the rows of the centered Jacobian.

**Lemma E.1.** *Consider the setting of Theorem 3.1, let $x \sim P_X$, and let $A$ be defined as*

$$A = \mathbb{E}_x\left[\tilde{f}_{L-2}(x)(D_L\tilde{\phi}'(g_{L-1}(x)))^\top\right]. \tag{189}$$

*Then, we have*

$$\|A\|_F = \mathcal{O}\left(\sqrt{n_{L-1}}\log(n_{L-1})\right), \tag{190}$$

*with probability at least $1 - 2\exp(-c\log^2 n_{L-1}) - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^L$, where $c$ is an absolute constant.*

*Proof.* We condition on $\|f_{L-2}(x)\|_{\text{Lip}} = \mathcal{O}(1)$ and on $\|\phi'(g_{L-1}(x))\|_{\text{Lip}} = \mathcal{O}(1)$. By Lemma B.2, these two conditions hold with probability at least $1 - C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L-1}$. Then, as $P_X$ satisfies Assumption 2.2, we have that $\left\|\tilde{f}_{L-2}(x)\right\|_{\psi_2} = \mathcal{O}(1)$ and $\left\|\tilde{\phi}'(g_{L-1}(x))\right\|_{\psi_2} = \mathcal{O}(1)$. Hence, an application of Lemma B.6 gives that

$$\left\|\mathbb{E}_x\left[\tilde{f}_{L-2}(x)\tilde{\phi}'(g_{L-1}(x))^\top\right]\right\|_{\text{op}} \leq C_1, \tag{191}$$

where $C_1$ is a numerical constant.

The following chain of inequalities holds:

$$\begin{aligned}
\|A\|_F &\leq \left\|\mathbb{E}_x\left[\tilde{f}_{L-2}(x)\tilde{\phi}'(g_{L-1}(x))^\top\right]\right\|_F \|D_L\|_{\text{op}} \\
&\leq \sqrt{n_{L-1}}\left\|\mathbb{E}_x\left[\tilde{f}_{L-2}(x)\tilde{\phi}'(g_{L-1}(x))^\top\right]\right\|_{\text{op}} \|D_L\|_{\text{op}} \\
&\leq C_1\sqrt{n_{L-1}}\|D_L\|_{\text{op}} \\
&= \mathcal{O}\left(\sqrt{n_{L-1}}\log(n_{L-1})\right),
\end{aligned} \tag{192}$$

where the third line uses (191), and the last holds with probability $1 - 2\exp(-c\log^2 n_{L-1})$ over $W_L$, because of Lemma B.3. Taking into account the initial conditioning, we get the desired result. □

The next two results provide bounds on the $\ell_2$ norm and on the sub-exponential $\psi_1$ norm of the rows of $\tilde{J}_{L-2}$, respectively.

**Lemma E.2** ($\ell_2$ norm of rows of centered Jacobian)**.** *Consider the setting of Theorem 3.1, let $x \sim P_X$, and let $\tilde{J}_x$ be defined as*

$$\tilde{J}_x = \tilde{f}_{L-2}(x) \otimes (D_L\tilde{\phi}'(g_{L-1}(x))) - \mathbb{E}_x\left[\tilde{f}_{L-2}(x) \otimes D_L\tilde{\phi}'(g_{L-1}(x))\right]. \tag{193}$$

*Then, we have*

$$\left\|\tilde{J}_x\right\|_2 = \Theta(\sqrt{n_{L-1}n_{L-2}}), \tag{194}$$

*with probability at least $1 - C\exp(-cn_{L-1}) - 12\exp(-c\log^2 n_{L-1})$ over $x$ and $(W_k)_{k=1}^L$ and $x$.*

*Proof.* We have that

$$\begin{aligned}
\left\|\tilde{J}_x\right\|_2 &= \left\|\tilde{f}_{L-2}(x) \otimes (D_L\tilde{\phi}'(g_{L-1}(x))) - \mathbb{E}_x\left[\tilde{f}_{L-2}(x) \otimes D_L\tilde{\phi}'(g_{L-1}(x))\right]\right\|_2 \\
&= \left\|\tilde{f}_{L-2}(x)(D_L\tilde{\phi}'(g_{L-1}(x)))^\top - \mathbb{E}_x\left[\tilde{f}_{L-2}(x)(D_L\tilde{\phi}'(g_{L-1}(x)))^\top\right]\right\|_F \\
&= \left\|\tilde{f}_{L-2}(x)^\top(D_L\tilde{\phi}'(g_{L-1}(x))) - A\right\|_F,
\end{aligned} \tag{195}$$

where $A$ is defined in (189). The second equality is justified by the identity $\|u \otimes v\|_2 = \left\|uv^\top\right\|_F$ that holds for any vectors $u, v$. An application of the triangle inequality gives that

$$\eta - \|A\|_F \leq \left\|\tilde{J}_x\right\|_2 \leq \eta + \|A\|_F, \quad \text{with } \eta = \left\|\tilde{f}_{L-2}(x)\right\|_2 \left\|D_L\tilde{\phi}'(g_{L-1}(x))\right\|_2. \tag{196}$$

Lemma C.4 gives that

$$\left\|\tilde{f}_{L-2}(x)\right\|_2 = \|f_{L-2}(x) - \mathbb{E}_x\left[f_{L-2}(x)\right]\|_2 = \Theta(\sqrt{n_{L-2}}), \tag{197}$$

with probability at least $1 - C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^L$ and $x$. Furthermore, Lemma C.5 gives that

$$\left\|D_L\tilde{\phi}'(g_{L-1}(x))\right\|_2 = \|D_L\left(\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[\phi'(g_{L-1}(x))\right]\right)\|_2 = \Theta(\sqrt{n_{L-1}}), \tag{198}$$

with probability at least $1 - 10\exp(-c\log^2 n_{L-1}) - C'\exp(-cn_{L-1})$ over $x$ and $(W_k)_{k=1}^L$. By combining (196), (197), (198) and the bound on $\|A\|_F$ provided by Lemma E.1, we conclude that

$$\left\|\tilde{J}_x\right\|_2 = \Theta(\sqrt{n_{L-1}n_{L-2}}), \tag{199}$$

with probability at least

$$\begin{aligned}
&1 - C'\exp(-cn_{L-1}) - 10\exp(-c\log^2 n_{L-1}) \\
&\qquad - C'\exp(-cn_{L-1}) - 2\exp(-c\log^2 n_{L-1}) - C'\exp(-cn_{L-1}) \\
&\geq 1 - C\exp(-cn_{L-1}) - 12\exp(-c\log^2 n_{L-1}),
\end{aligned} \tag{200}$$

over $x$ and $(W_k)_{k=1}^L$, which gives the desired result. □

**Lemma E.3** (Sub-exponential norm of rows of centered Jacobian)**.** *Consider the setting of Theorem 3.1, let $x \sim P_X$, and let $\tilde{J}_x$ be defined as in* (193)*. Fix a realization of $(W_k)_{k=1}^L$. Then, with probability at least $1 - 2\exp(-c\log^2 n_{L-1}) - C\exp(-cn_{L-1})$ over this realization ($c$ being a numerical constant), we have that*

$$\left\|\tilde{J}_x\right\|_{\psi_1} = \mathcal{O}\left(\log n_{L-1}\right). \tag{201}$$

*Proof.* We condition on $\|f_{L-2}(x)\|_{\text{Lip}} = \mathcal{O}(1)$ and on $\|\phi'(g_{L-1}(x))\|_{\text{Lip}} = \mathcal{O}(1)$. By Lemma B.2, these two conditions hold with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L-1}$. Then, we have

$$\begin{aligned}
\left\|\tilde{J}_x\right\|_{\psi_1} &= \sup_{u \text{ s.t. } \|u\|_2=1} \left\|u^\top \tilde{J}_x\right\|_{\psi_1} \\
&= \sup_{U \text{ s.t. } \|U\|_F=1} \left\|\tilde{f}_{L-2}(x)U(D_L\tilde{\phi}'(g_{L-1}(x))) - \mathbb{E}_x\left[\tilde{f}_{L-2}(x)UD_L\tilde{\phi}'(g_{L-1}(x))\right]\right\|_{\psi_1} \\
&\leq C_0 \sup_{U \text{ s.t. } \|U\|_F=1} \|UD_L\|_F \\
&\leq C_0 \|D_L\|_{\text{op}} \\
&\leq C_0 \log n_{L-1},
\end{aligned} \tag{202}$$

where the third line follows from Lemma B.5 and the last inequality holds with probability at least $1 - 2\exp(-c\log^2 n_{L-1})$ over $W_L$ by Lemma B.3. Taking into account the initial conditioning, we get the desired result. □

### E.2 Proof of Proposition 3.3

*Proof of Proposition 3.3.* Following the notation in [2], we define

$$B := \sup_{z \in \mathbb{R}^N : \|z\|_2 = 1} \left| \left\| \sum_{i=1}^N z_i \tilde{J}_{i:} \right\|_2^2 - \sum_{i=1}^N z_i^2 \left\|\tilde{J}_{i:}\right\|_2^2 \right|^{\frac{1}{2}}. \tag{203}$$

Then, for any $z \in \mathbb{R}^N$ with unit norm, we have that

$$\left\|\tilde{J}z\right\|_2^2 = \left\|\sum_{i=1}^N z_i \tilde{J}_{i:}\right\|_2^2 - \sum_{i=1}^N z_i^2 \left\|\tilde{J}_{i:}\right\|_2^2 + \sum_{i=1}^N z_i^2 \left\|\tilde{J}_{i:}\right\|_2^2 \geq \min_i \left\|\tilde{J}_{i:}\right\|_2^2 - B^2, \tag{204}$$

which implies that

$$\lambda_{\min}\left(\tilde{J}\tilde{J}^\top\right) = \inf_{z\in\mathbb{R}^N:\|z\|_2=1} \left\|\tilde{J}z\right\|_2^2 \geq \min_i \left\|\tilde{J}_{i:}\right\|_2^2 - B^2. \tag{205}$$

In our case, $\tilde{J}_{i:} \in \mathbb{R}^{n_{L-2}n_{L-1}}$. Notice that this dimension is indicated with $n$ in Theorem 3.2 of [2]. In the statement of the mentioned Theorem, let's fix $r = 1$, $m = N$, and $\theta = (N/(n_{L-1}n_{L-2}))^{1/4} < 1/4$. Then, we have that the condition required to apply Theorem 3.2 is satisfied, *i.e.*

$$N \log^2\left(2\sqrt[4]{\frac{n_{L-2}n_{L-1}}{N}}\right) \leq \sqrt{N n_{L-2}n_{L-1}}, \tag{206}$$

where the inequality follows from Assumption 2.5. By combining (205) with the upper bound on $B$ given by Theorem 3.2 of [2], the desired result readily follows. □

### E.3 Proof of Theorem 3.4

*Proof of Theorem 3.4.* By Lemma E.3, we have that, with probability at least $1 - 2\exp(-c\log^2 n_{L-1}) - C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^L$, the rows of $\tilde{J}$ are sub-exponential (with respect to the randomness in $(x_i)_{i=1}^N$). In particular, we have that

$$\psi := \max_i \left\|\tilde{J}_{i:}\right\|_{\psi_1} \leq C_1 \log n_{L-1}. \tag{207}$$

Furthermore, by Lemma E.2, we have that

$$\left\|\tilde{J}_{i:}\right\|_2 = \Theta(\sqrt{n_{L-2}n_{L-1}}), \tag{208}$$

with probability at least $1 - p$ over $x_i$ and $(W_k)_{k=1}^L$, where to ease the notation we have defined $p := C'\exp(-c_0 n_{L-1}) + 12\exp(-c_0\log^2 n_{L-1})$. Hence, with probability at least $1 - \sqrt{p}$ over $(W_k)_{k=1}^L$, we have that

$$\mathbb{P}_{x_i}\left(c_1\sqrt{n_{L-2}n_{L-1}} \leq \left\|\tilde{J}_{i:}\right\|_2 \leq c_2\sqrt{n_{L-2}n_{L-1}}\right) \geq 1 - \sqrt{p}, \tag{209}$$

for some numerical constants $c_2 > c_1 > 0$. In (209), we use the symbol $\mathbb{P}_{x_i}$ to highlight that this probability is taken over $x_i$. For the rest of the argument, we condition on a realization of $(W_k)_{k=1}^L$ s.t. (207) and (209) hold. Then, by performing a union bound over the samples, we have that

$$\eta_{\text{min}} = \min_i \left\|\tilde{J}_{i:}\right\|_2 \geq c_1\sqrt{n_{L-2}n_{L-1}}, \tag{210}$$

and

$$\eta_{\text{max}} = \max_i \left\|\tilde{J}_{i:}\right\|_2 \leq c_2\sqrt{n_{L-2}n_{L-1}}, \tag{211}$$

with probability at least $1 - N\sqrt{p}$ over $(x_i)_{i=1}^N$.

Next, we apply Proposition 3.3 with $K = 1$, $K' = c_2$ and

$$\begin{aligned}\Delta &= C_1(\psi K + K')^2 N^{1/4}(n_{L-1}n_{L-2})^{3/4} \\ &\leq C_2 \log^2 n_{L-1} N^{1/4}(n_{L-1}n_{L-2})^{3/4} \\ &= o(n_{L-1}n_{L-2}).\end{aligned} \tag{212}$$

Note that (22) in Lemma B.1 gives that $N^{1/4}\cdot\log^2 n_{L-1} = o((n_{L-1}n_{L-2})^{1/4})$, which justifies the last line. Thus, (13) implies that

$$\lambda_{\min}\left(\tilde{J}\tilde{J}^\top\right) \geq \eta_{\text{min}}^2 - \Delta \geq c_1 n_{L-2}n_{L-1} - o(n_{L-1}n_{L-2}) = \Theta(n_{L-2}n_{L-1}), \tag{213}$$

with probability at least

$$\begin{aligned}1-\exp\left(-cK\sqrt{N}\log\left(\frac{2n_{L-1}n_{L-2}}{N}\right)\right)-\mathbb{P}\left(\eta_{\max}\geq K'\sqrt{n_{L-1}n_{L-2}}\right)\\ \geq 1-\exp\left(-c\sqrt{N}\right)-\mathbb{P}\left(\eta_{\max}\geq c_2\sqrt{n_{L-1}n_{L-2}}\right)\\ \geq 1-\exp\left(-c\sqrt{N}\right)-N\sqrt{p},\end{aligned} \tag{214}$$

where the last inequality follows from (211). By taking into account the conditioning over $(W_k)_{k=1}^L$ made in order to guarantee (207) and (209), we conclude that $\lambda_{\min}\left(\tilde{J}\tilde{J}^\top\right)=\Omega(n_{L-1}n_{L-2})$ with probability at least

$$\begin{aligned}&1-\exp\left(-c\sqrt{N}\right)-N\sqrt{p}-\sqrt{p}-2\exp(-c\log^2 n_{L-1})-C'\exp(-cn_{L-1})\\ &=1-\exp\left(-c\sqrt{N}\right)-(N+1)\sqrt{C'\exp(-c_0n_{L-1})+12\exp(-c_0\log^2 n_{L-1})}\\ &\qquad\qquad-2\exp(-c\log^2 n_{L-1})-C'\exp(-cn_{L-1})\\ &\geq 1-\exp\left(-c\sqrt{N}\right)-(N+1)\left(\sqrt{C'}\exp(-c_0n_{L-1}/2)+\sqrt{12}\exp(-c_0\log^2 n_{L-1}/2)\right)\\ &\qquad\qquad-2\exp(-c\log^2 n_{L-1})-C'\exp(-cn_{L-1})\\ &\geq 1-\exp\left(-c\sqrt{N}\right)-C''N\exp(-c_1n_{L-1})-C''N\exp(-c\log^2 n_{L-1}),\end{aligned} \tag{215}$$

over $(x_i)_{i=1}^N$ and $(W_k)_{k=1}^L$, which gives the desired result. □

# F Proof of the Upper Bound 7

Before giving the proof of the upper bound 7, we provide again its statement for the reader's convenience.

**Lemma F.1** (Upper bound on the smallest NTK eigenvalue)**.** *Consider the setting of Theorem 3.1, and let $K$ be the NTK Gram matrix* (3)*. Then, we have*

$$\lambda_{\min}(K)=\mathcal{O}(dn_{L-1}), \tag{216}$$

*with probability at least $1-C\exp(-cn_{L-1})$ over $(x_i)_{i=1}^N$ and $(W_k)_{k=1}^L$, where $c$ and $C$ are numerical constants.*

*Proof.* By using the expression in (8), we have that

$$\lambda_{\min}(K)=\lambda_{\min}\left(JJ^\top\right)\leq(JJ^\top)_{11}=\sum_{l=0}^{L-1}\|(F_l)_{1:}\|_2^2\,\|(B_{l+1})_{1:}\|_2^2. \tag{217}$$

An application of Lemma C.1 gives that

$$\|(F_l)_{1:}\|_2^2=\|f_l(x_1)\|_2^2=\Theta(n_l), \tag{218}$$

with probability at least $1-C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$ and $x_1$. We condition on the event such that (218) holds for all $l\in\{0,\dots,L-1\}$. This happens with probability at least $1-C''\exp(-cn_{L-1})$ over $(W_k)_{k=1}^L$ and $x_1$.

By definition, we have that $\|B_L\|_2=1$ and that, for $l\in[L-1]$,

$$\|(B_l)_{1:}\|_2^2=\left\|\prod_{k=l}^{L-1}\Sigma_k(x_1)W_{k+1}\right\|_2^2. \tag{219}$$

Since $\Sigma_k(x_1)=\operatorname{diag}([\phi'(g_{k,j}(x_1))]_{j=1}^{n_k})$, by Assumption 2.3, we have that

$$\|\Sigma_k(x_1)\|_{\text{op}}\leq M. \tag{220}$$

Let us now condition on the following two events: *(i)* $\|W_k\|_{\text{op}} = \Theta(1)$, for all $k \in [L-1]$ (this happens with probability at least $1 - C' \exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L-1}$, see (33) in the proof of Lemma B.2), and *(ii)* $\|W_L\|_2 = \Theta(\sqrt{n_{L-1}})$ (this happens with probability at least $1 - \exp(-cn_{L-1})$ over $W_L$, by Theorem 3.1.1 in [65]). Then, we readily get

$$\|(B_l)_{1:}\|_2^2 = \mathcal{O}\left(n_{L-1}\right). \tag{221}$$

Taking the intersection of all the events over which we conditioned, we finally obtain

$$\sum_{l=0}^{L-1} \|(F_l)_{1:}\|_2^2 \, \|(B_{l+1})_{1:}\|_2^2 = \mathcal{O}\left(n_{L-1} \sum_{l=0}^{L-1} n_l\right) = \mathcal{O}\left(dn_{L-1}\right), \tag{222}$$

with probability at least $1 - (1 + C' + C'') \exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L}$ and $x_1$, where in the last step we have used Assumption 2.4. By combining (217) and (222), the desired result follows. □

# G Proof of Corollary 4.1

*Proof of Corollary 4.1.* By Theorem 3.1, we have that the smallest eigenvalue of $JJ^\top$ is bounded away from zero with probability at least $1 - p$ over $(x_i)_{i=1}^N$ and $(W_k)_{k=1}^L$, where

$$p := C\,Ne^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}. \tag{223}$$

Hence, with probability at least $1 - p$ over $(x_i)_{i=1}^N$, there exists a set of parameters $\theta_0$ such that $J(\theta_0)$ has a right inverse. Thus, for all $Y \in \mathbb{R}^N$, there exists $\theta'$ such that

$$J(\theta_0)\theta' = \left.\frac{\partial F_L(\theta)}{\partial \theta}\right|_{\theta=\theta_0} \theta' = Y. \tag{224}$$

This can also be written, for all $i \in [N]$, as

$$y_i = \left.\frac{\partial f_L(\theta, x_i)}{\partial \theta}\right|_{\theta=\theta_0}^\top \theta' = \lim_{h\to 0} \frac{f_L(\theta_0 + h\theta', x_i) - f_L(\theta_0, x_i)}{h}. \tag{225}$$

Then, for all $\varepsilon > 0$, there exists $h^*$ such that, for all $i \in [N]$,

$$|y_i - f^*(x_i)| \le \frac{\varepsilon}{\sqrt{N}}, \tag{226}$$

where

$$f^*(x_i) := \frac{f_L(\theta_0 + h^*\theta', x_i) - f_L(\theta_0, x_i)}{h^*}. \tag{227}$$

Finally, the desired result follows by noticing that $f^*$ can be implemented by a network with the same depth and twice more neurons at every hidden layer. □

# H Proof of Theorem 4.2

**Notation for this appendix.** In this appendix, we use $J(\theta)$ to denote the Jacobian of the network output $F_L$, evaluated in $\theta$. We recall that $J(\theta)$ is a matrix with $N$ rows and $\sum_{l=0}^{L-3} n_l n_{l+1} + 2n_{L-2}n_{L-1} + 2n_{L-1}$ columns (for the optimization result, we assume that the $(L-1)$-th layer has an even number of neurons and denote its width as $2n_{L-1}$). Let $K(\theta) = J(\theta)(J(\theta))^\top$ be the associated empirical NTK Gram matrix, and let $\theta_0$ be the initialization defined in (18). We also make the dependence on $\theta$ explicit for feature vectors and backpropagation terms: the feature vector at the $l$-th layer with input $x_i$ and network parameter $\theta$ is denoted by $f_l(\theta, x_i)$, and the corresponding backpropagation term is denoted by $b_l(\theta, x_i)$, where $b_l(\theta, x_i) = (B_l(\theta))_{i:}$. Finally, we use $W_l(\theta)$ to denote the weights of the $l$-th layer evaluated at the parameter $\theta$.

A straightforward application of Theorem 2.1 in [49] gives the following proposition.

**Proposition H.1.** *Consider solving the least-squares optimization problem*

$$\min_\theta \mathcal{L}(\theta) := \frac{1}{2} \min_\theta \|F_L(\theta) - Y\|_2^2, \tag{228}$$

*by running gradient descent updates of the form $\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}(\theta_t)$, with some initialization $\tilde{\theta}_0$. Assume there exists $\alpha, \beta \in \mathbb{R}$, such that, if we define $\mathcal{D} = \mathcal{B}(\tilde{\theta}_0, R)$ as the $\ell_2$ ball centered in $\tilde{\theta}_0$ with radius $R$, with*

$$R := \frac{4 \left\| F_L(\tilde{\theta}_0) - Y \right\|_2}{\alpha}, \tag{229}$$

*the following holds*

$$\forall \theta \in \mathcal{D} : \ \alpha \leq \sigma_{\min}(J(\theta)) \leq \|J(\theta)\|_{\text{op}} \leq \beta, \tag{230}$$

$$\forall \theta_1, \theta_2 \in \mathcal{D} : \ \|J(\theta_1) - J(\theta_2)\|_{\text{op}} \leq \frac{\alpha^2}{2\beta}. \tag{231}$$

*Then, by setting $\eta \leq 1/(2\beta^2)$, we have that, for all $t \geq 1$,*

$$\mathcal{L}(\theta_t) \leq \left(1 - \frac{\eta \alpha^2}{2}\right)^t \mathcal{L}(\tilde{\theta}_0). \tag{232}$$

In order to apply this proposition with initialization $\tilde{\theta}_0 = \theta_0$, we need to prove that the necessary assumptions hold. We will do so by showing the following intermediate results:

- Lemma H.2 shows that, at the initial point $\theta_0$, the network output is 0 and the smallest NTK eigenvalue is lower bounded.
- Lemma H.3 gives a tight estimate on the operator norm of the weights inside a ball $\mathcal{D}$ centered at $\theta_0$ and with radius $R = o(1)$.
- Lemma H.4 gives an upper bound on the distance between a feature vector in $\mathcal{D}$ and the feature vector at $\theta_0$.
- Lemma H.5 gives upper bounds on the $\ell_2$ norm and the $\ell_2$ distance between feature vectors in $\mathcal{D}$.
- Lemmas H.6 and H.7 give upper bounds on the $\ell_2$ norm and the $\ell_2$ distance of backpropagation terms in $\mathcal{D}$, respectively.
- Lemma H.8 gives an upper bound on the difference in operator norm between Jacobians in $\mathcal{D}$.
- Finally, Lemma H.9 gives upper and lower bounds on the NTK spectrum in $\mathcal{D}$.

**Lemma H.2** (Network output and smallest NTK eigenvalue at initialization)**.** *Let $\theta_0$ be defined in* (18)*. Then, we have that, for all $x \in \mathbb{R}^d$,*

$$f_L(x, \theta_0) = 0. \tag{233}$$

*Furthermore, we have that*

$$\sigma_{\min}(J(\theta_0)) \geq c_1 \sqrt{\gamma n_{L-2} n_{L-1}}), \tag{234}$$

*with probability at least $1 - C\, N e^{-c \log^2 n_{L-1}} - C e^{-c \log^2 N}$ over $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$ and $\theta_0$, where $c, c_1$ and $C$ are numerical constants.*

*Proof.* By definition (18) of the initialization $\theta_0$, we have that

$$\begin{aligned} f_L(x, \theta_0) &= (W_L^{(1)}(\theta_0))^\top \phi((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x)) \\ &\quad + (W_L^{(2)}(\theta_0))^\top \phi((W_{L-1}^{(2)}(\theta_0))^\top f_{L-2}(\theta_0, x)) \\ &= (W_L^{(1)}(\theta_0))^\top \phi((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x)) \\ &\quad + (-W_L^{(1)}(\theta_0))^\top \phi((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x)) \\ &= 0, \end{aligned} \tag{235}$$

where in the second equality we use that $W_{L-1}^{(2)}(\theta_0) = W_{L-1}^{(1)}(\theta_0)$ and that $W_L^{(2)}(\theta_0) = -W_L^{(1)}(\theta_0)$. This proves (233).

Let us now compute the Jacobian at initialization $J(\theta_0)$. For $l \in [L-2]$, we have that

$$\begin{aligned}
\frac{\partial f_L(x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0} &= (W_L^{(1)}(\theta_0))^\top \Bigg( \phi'\left((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0,x)\right)\left((W_{L-1}^{(1)}(\theta_0))^\top \frac{\partial f_{L-2}(\theta,x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0}\right)\Bigg) \\
&+ (W_L^{(2)}(\theta_0))^\top \left( \phi'\left((W_{L-1}^{(2)}(\theta_0))^\top f_{L-2}(\theta_0,x)\right)\left((W_{L-1}^{(2)}(\theta_0))^\top \frac{\partial f_{L-2}(\theta,x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0}\right)\right) \\
&= (W_L^{(1)}(\theta_0))^\top \Bigg( \phi'\left((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0,x)\right)\left((W_{L-1}^{(1)}(\theta_0))^\top \frac{\partial f_{L-2}(\theta,x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0}\right)\Bigg) \\
&- (W_L^{(1)}(\theta_0))^\top \left( \phi'\left((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0,x)\right)\left((W_{L-1}^{(1)}(\theta_0))^\top \frac{\partial f_{L-2}(\theta,x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0}\right)\right) \\
&= 0,
\end{aligned} \tag{236}$$

where in the second equality we use again that $W_{L-1}^{(2)}(\theta_0) = W_{L-1}^{(1)}(\theta_0)$ and that $W_L^{(2)}(\theta_0) = -W_L^{(1)}(\theta_0)$.

Let us define $f_{L-1}^{(k)}(\theta,x) := \phi((W_{L-1}^{(k)}(\theta))^\top f_{L-2}(\theta,x))$ for $k \in \{1,2\}$. Then, for the $(L-1)$-th layer, by isolating the computation over $W_{L-1}^{(1)}$, we have that

$$\begin{aligned}
\frac{\partial f_L(\theta,x)}{\partial (W_{L-1}^{(1)})_{ij}}\bigg|_{\theta=\theta_0} &= (W_L^{(1)}(\theta_0))^\top \frac{\partial f_{L-1}^{(1)}(\theta,x)}{\partial (W_{L-1}^{(1)})_{ij}}\bigg|_{\theta=\theta_0} + (W_L^{(2)}(\theta_0))^\top \frac{\partial f_{L-1}^{(2)}(\theta,x)}{\partial (W_{L-1}^{(1)})_{ij}}\bigg|_{\theta=\theta_0} \\
&= (W_L^{(1)}(\theta_0))^\top \frac{\partial f_{L-1}^{(1)}(\theta,x)}{\partial (W_{L-1}^{(1)})_{ij}}\bigg|_{\theta=\theta_0} \\
&=: J^{(1)}(\theta_0),
\end{aligned} \tag{237}$$

where we use that $f_{L-1}^{(2)}(\theta,x)$ does not depend on the parameters $W_{L-1}^{(1)}$. Proceeding in the same way and using that $W_{L-1}^{(2)}(\theta_0) = W_{L-1}^{(1)}(\theta_0)$ and $W_L^{(2)}(\theta_0) = -W_L^{(1)}(\theta_0)$, we also obtain that

$$\frac{\partial f_L(\theta,x)}{\partial (W_{L-1}^{(2)})_{ij}}\bigg|_{\theta=\theta_0} = -J^{(1)}(\theta_0). \tag{238}$$

Finally, by observing that $f_L(\theta,x) = (W_L^{(1)})^\top f_{L-1}^{(1)}(\theta,x) + (W_L^{(2)})^\top f_{L-1}^{(2)}(\theta,x)$, we deduce

$$\frac{\partial f_L(\theta,x)}{\partial (W_L^{(k)})_i}\bigg|_{\theta=\theta_0} = \left(f_{L-1}^{(k)}(\theta_0,x)\right)_i, \quad \text{for } k \in \{1,2\}. \tag{239}$$

Hence, the NTK at initialization $K(\theta_0)$ can be expressed as

$$\begin{aligned}
K(\theta_0) = J^{(1)}(\theta_0)(J^{(1)}(\theta_0))^\top + J^{(1)}(\theta_0)(J^{(1)}(\theta_0))^\top \\
+ F_{L-1}^{(1)}(\theta_0)(F_{L-1}^{(1)}(\theta_0))^\top + F_{L-1}^{(2)}(\theta_0)(F_{L-1}^{(2)}(\theta_0))^\top.
\end{aligned} \tag{240}$$

Note that, by construction, $J^{(1)}(\theta_0)$ has the same distribution of $J_{L-2}$, whose rows are given by (10). Therefore, by combining the results from Theorem 3.2 and Theorem 3.4, we conclude that

$$\sigma_{\min}(J(\theta_0)) \geq 2\sqrt{\gamma}\sigma_{\min}(J_{L-2}) \geq c_1\sqrt{\gamma n_{L-2} n_{L-1}}, \tag{241}$$

with probability at least $1 - C\,N e^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}$ over $(x_i)_{i=1}^N$ and $\theta_0$, where $c_1, C$ are numerical constants. This proves (234), and concludes the proof of the lemma. □

**Lemma H.3** (Operator norm of weights in $\mathcal{D}$)**.** *Let $\theta_0$ be defined in* (18)*, let $\mathcal{D} = \mathcal{B}(\theta_0, R)$ and assume that $R = o(1)$. Then, for any $l \in [L-1]$,*

$$\sup_{\theta\in\mathcal{D}} \|W_l(\theta)\|_{\text{op}} = \mathcal{O}(1), \tag{242}$$

*with probability at least $1 - 2\exp(-cn_{L-1})$ over $W_l(\theta_0)$.*

*Proof.* By Weyl's theorem, we have that, for all $l \in [L-2]$,

$$\begin{aligned}\sup_{\theta\in\mathcal{D}} \|W_l(\theta)\|_{\text{op}} &\le \|W_l(\theta_0)\|_{\text{op}} + \sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_{\text{op}} \\ &\le \|W_l(\theta_0)\|_{\text{op}} + \sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_F \\ &\le \|W_l(\theta_0)\|_{\text{op}} + \sup_{\theta\in\mathcal{D}} \|\theta - \theta_0\|_2 \\ &= \|W_l(\theta_0)\|_{\text{op}} + o(1) \\ &= \mathcal{O}(1),\end{aligned} \tag{243}$$

where in the fourth line we use that $R = o(1)$, and the result of the last line holds with probability at least $1 - 2\exp(-cn_{L-1})$ over $W_l(\theta_0)$ by Theorem 4.4.5 of [65]. By following the same argument, we have that, with probability at least $1 - 2\exp(-cn_{L-1})$ over $W_{L-1}(\theta_0)$,

$$\sup_{\theta\in\mathcal{D}} \left\| W_{L-1}^{(k)}(\theta) \right\|_{\text{op}} = \mathcal{O}(1), \quad \text{for } k \in \{1, 2\}, \tag{244}$$

which readily implies that $\sup_{\theta\in\mathcal{D}} \|W_{L-1}(\theta)\|_{\text{op}} = \mathcal{O}(1)$ and concludes the proof. □

**Lemma H.4** (Distance of features in $\mathcal{D}$ from initialization)**.** *Let $\theta_0$ be defined in* (18)*, $x \sim P_X$, $\mathcal{D} = \mathcal{B}(\theta_0, R)$ and assume that $R = o(1)$. Then, for any $0 \le l \le L-1$, we have*

$$\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x) - f_l(\theta_0, x)\|_2 \le C\, R\sqrt{d}, \tag{245}$$

*with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^l$ and $x$, where $c, C$ are numerical constants.*

*Proof.* We prove the claim by induction over $l$. For the base case, we have $f_0(\theta, x) = f_0(\theta_0, x)$, hence (245) holds with probability 1.

For the induction case, let $l > 0$. Then,

$$\begin{aligned}\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x) - f_l(\theta_0, x)\|_2 &= \sup_{\theta\in\mathcal{D}} \left\|\phi\left((W_l(\theta))^\top f_{l-1}(\theta, x)\right) - \phi\left((W_l(\theta_0))^\top f_{l-1}(\theta_0, x)\right)\right\|_2 \\ &\le M \sup_{\theta\in\mathcal{D}} \left\|(W_l(\theta))^\top f_{l-1}(\theta, x) - (W_l(\theta_0))^\top f_{l-1}(\theta_0, x)\right\|_2 \\ &\le M \sup_{\theta\in\mathcal{D}} \left\|(W_l(\theta))^\top f_{l-1}(\theta, x) - (W_l(\theta))^\top f_{l-1}(\theta_0, x)\right\|_2 \\ &\quad + M \sup_{\theta\in\mathcal{D}} \left\|(W_l(\theta))^\top f_{l-1}(\theta_0, x) - (W_l(\theta_0))^\top f_{l-1}(\theta_0, x)\right\|_2 \\ &\le M \sup_{\theta\in\mathcal{D}} \|W_l(\theta)\|_{\text{op}} \sup_{\theta\in\mathcal{D}} \|f_{l-1}(\theta, x) - f_{l-1}(\theta_0, x)\|_2 \\ &\quad + M \sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_{\text{op}} \|f_{l-1}(\theta_0, x)\|_2.\end{aligned} \tag{246}$$

By Lemma H.3, we have that

$$\sup_{\theta\in\mathcal{D}} \|W_l(\theta)\|_{\text{op}} = \mathcal{O}(1), \tag{247}$$

with probability at least $1 - 2\exp(-cn_{L-1})$ over $W_l(\theta_0)$. By inductive hypothesis, we have

$$\sup_{\theta\in\mathcal{D}} \|f_{l-1}(\theta, x) - f_{l-1}(\theta_0, x)\|_2 \le C\, R\sqrt{d}, \tag{248}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l-1}$ and $x$. Clearly, we also have that

$$\sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_{\text{op}} \le \sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_F \le \sup_{\theta\in\mathcal{D}} \|\theta - \theta_0\| \le R. \tag{249}$$

Furthermore, an application of Lemma C.1 gives that

$$\|f_{l-1}(\theta_0, x)\|_2 = \Theta(\sqrt{n_{l-1}}) = \mathcal{O}\left(\sqrt{d}\right), \tag{250}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l-1}$ and $x$. By combining (246), (247), (248), (249) and (250), we obtain that

$$\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x) - f_l(\theta_0, x)\|_2 \leq C\,R\sqrt{d}, \tag{251}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and $x$, which completes the proof. □

**Lemma H.5** ($\ell_2$ norm and $\ell_2$ distance of features in $\mathcal{D}$)**.** *Let $\theta_0$ be defined in* (18)*, $x \sim P_X$, $\mathcal{D} = \mathcal{B}(\theta_0, R)$ and assume that $R = o(1)$. Then, for any $0 \leq l \leq L-1$, we have*

$$\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x)\|_2 = \mathcal{O}\left(\sqrt{d}\right), \tag{252}$$

*with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and $x$, where $c, C$ are numerical constants. Furthermore,*

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1, x) - f_l(\theta_2, x)\|_2 \leq C\,R\,\sqrt{d}, \tag{253}$$

*with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and $x$.*

*Proof.* The first statement follows from the chain of inequalities below:

$$\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x)\|_2 \leq \|f_l(\theta_0, x)\|_2 + \sup_{\theta\in\mathcal{D}} \|f_l(\theta) - f_l(\theta_0)\|_2 \leq C\sqrt{n_l} + C\,R\sqrt{d}, \tag{254}$$

where the second inequality holds with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and $x$ by combining Lemma C.1 and Lemma H.4.

We prove the second statement by induction over $l$. For the base case, we have $f_0(\theta_1, x) = f_0(\theta_2, x)$, hence (253) holds with probability 1.

For the induction case, let $l > 0$. Then,

$$\begin{aligned}
\sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1, x) - f_l(\theta_2, x)\|_2 &= \sup_{\theta_1,\theta_2\in\mathcal{D}} \left\|\phi\left((W_l(\theta_1))^\top f_{l-1}(\theta_1, x)\right) - \phi\left((W_l(\theta_2))^\top f_{l-1}(\theta_2, x)\right)\right\|_2 \\
&\leq M \sup_{\theta_1,\theta_2\in\mathcal{D}} \left\|(W_l(\theta_1))^\top f_{l-1}(\theta_1, x) - (W_l(\theta_2))^\top f_{l-1}(\theta_2, x)\right\|_2 \\
&\leq M \sup_{\theta_1,\theta_2\in\mathcal{D}} \left\|(W_l(\theta_1))^\top f_{l-1}(\theta_1, x) - (W_l(\theta_1))^\top f_{l-1}(\theta_2, x)\right\|_2 \\
&\quad + M \sup_{\theta_1,\theta_2\in\mathcal{D}} \left\|(W_l(\theta_1))^\top f_{l-1}(\theta_2, x) - (W_l(\theta_2))^\top f_{l-1}(\theta_2, x)\right\|_2 \\
&\leq M \sup_{\theta_1\in\mathcal{D}} \|W_l(\theta_1)\|_{\text{op}} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_{l-1}(\theta_1, x) - f_{l-1}(\theta_2, x)\|_2 \\
&\quad + M \sup_{\theta_1,\theta_2\in\mathcal{D}} \|W_l(\theta_1) - W_l(\theta_2)\|_{\text{op}} \sup_{\theta_2\in\mathcal{D}} \|f_{l-1}(\theta_2, x)\|_2 .
\end{aligned} \tag{255}$$

By Lemma H.3, we have that

$$\sup_{\theta_1\in\mathcal{D}} \|W_l(\theta_1)\|_{\text{op}} = \mathcal{O}(1), \tag{256}$$

with probability at least $1 - 2\exp(-cn_{L-1})$ over $W_l(\theta_0)$. By inductive hypothesis, we have

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_{l-1}(\theta_1, x) - f_{l-1}(\theta_2, x)\|_2 \leq C\,R\,\sqrt{d}, \tag{257}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l-1}$ and $x$. Clearly, we also have that

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|W_l(\theta_1) - W_l(\theta_2)\|_{\text{op}} \leq \sup_{\theta_1,\theta_2\in\mathcal{D}} \|W_l(\theta_1) - W_l(\theta_2)\|_F \leq \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\theta_1 - \theta_2\| \leq R. \tag{258}$$

Furthermore, by using (252), we have that

$$\sup_{\theta_2\in\mathcal{D}} \|f_{l-1}(\theta_2, x)\|_2 \leq C\,R\,\sqrt{d}, \tag{259}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l-1}$ and $x$. By combining (255), (256), (257), (258) and (259), we obtain that

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1, x) - f_l(\theta_2, x)\|_2 \le C\,R\sqrt{d}, \tag{260}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and $x$, which completes the proof. □

**Lemma H.6** ($\ell_2$ norm of backpropagation in $\mathcal{D}$)**.** *Let $\theta_0$ be defined in* (18)*, $x \sim P_X$, and $\mathcal{D} = \mathcal{B}(\theta_0, R)$. Assume that $R = o(1)$ and that $\gamma > 1$. Then, for any $l \in [L]$, we have*

$$\sup_{\theta\in\mathcal{D}} \|b_l(\theta, x)\|_2 \le C\sqrt{\gamma \cdot n_{L-1}}, \tag{261}$$

*with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^{L}$, where $c, C$ are numerical constants.*

*Proof.* We prove the claim by induction on $l \in \{L, L-1, \dots, 1\}$. For the base case, we have that $\|b_L(\theta, x)\|_2 = 1$, hence (261) clearly holds.

For the induction case, pick $l \in [L-1]$. Then,

$$\begin{aligned}
\sup_{\theta\in\mathcal{D}} \|b_l(\theta, x)\|_2 &= \sup_{\theta\in\mathcal{D}} \left\| \left( \prod_{k=l}^{L-2} \Sigma_k(\theta, x) W_{k+1}(\theta) \right) \Sigma_{L-1}(\theta, x) W_L(\theta) \right\|_2 \\
&\le \sup_{\theta\in\mathcal{D}} \left\| \left( \prod_{k=l}^{L-2} \Sigma_k(\theta, x) W_{k+1}(\theta) \right) \Sigma_{L-1}(\theta, x) \right\|_{\text{op}} \sup_{\theta\in\mathcal{D}} \|W_L(\theta)\|_2 \\
&\le \left( \prod_{k=l}^{L-2} \sup_{\theta\in\mathcal{D}} \|\Sigma_k(\theta, x)\|_{\text{op}} \sup_{\theta\in\mathcal{D}} \|W_{k+1}(\theta)\|_{\text{op}} \right) \sup_{\theta\in\mathcal{D}} \|\Sigma_{L-1}(\theta, x)\|_{\text{op}} \sup_{\theta\in\mathcal{D}} \|W_L(\theta)\|_2 \\
&\le M^{L-l} \left( \prod_{k=l+1}^{L-1} \sup_{\theta\in\mathcal{D}} \|W_k(\theta)\|_{\text{op}} \right) \left( \|W_L(\theta_0)\|_2 + \sup_{\theta\in\mathcal{D}} \|W_L(\theta) - W_L(\theta_0)\|_2 \right) \\
&\le C\,M^{L-l}(\|W_L(\theta_0)\|_2 + \sup_{\theta\in\mathcal{D}} \|\theta - \theta_0\|_2) \\
&\le C\,M^{L-l}(\sqrt{\gamma n_{L-1}} + \sup_{\theta\in\mathcal{D}} \|\theta - \theta_0\|_2) \\
&= C\sqrt{\gamma n_{L-1}}.
\end{aligned} \tag{262}$$

Here, the fourth line follows from Assumption 2.3, which gives $\sup_{\theta\in\mathcal{D}} \|\Sigma_k(\theta, x)\|_{\text{op}} \le M$; the fifth line holds with probability $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^{L-1}$ by Lemma H.3; the sixth line holds with probability at least $1 - \exp(-cn_{L-1})$ over $W_L(\theta_0)$ by Theorem 3.1.1 in [65]; and the last line follows from $R = o(1)$. Taking the intersection of these events gives the desired result. □

**Lemma H.7** ($\ell_2$ distance of backpropagation in $\mathcal{D}$)**.** *Let $\theta_0$ be defined in* (18)*, $x \sim P_X$, and $\mathcal{D} = \mathcal{B}(\theta_0, R)$. Assume that $R = o(1)$ and that $\gamma > 1$. Then, for any $l \in [L]$, we have*

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_l(\theta_1, x) - b_l(\theta_2, x)\|_2 \le C\,R\,\sqrt{\gamma d n_{L-1}}, \tag{263}$$

*with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^{L}$ and $x$, where $c, C$ are numerical constants.*

*Proof.* We prove the claim by induction on $l \in \{L, L-1, \ldots, 1\}$. For the base case, $b_L(\theta, x)$ does not depend on $\theta$, hence (263) clearly holds. For the induction case, pick $l \in [L-1]$. Then,

$$\begin{aligned}
&\sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_l(\theta_1,x) - b_l(\theta_2,x)\|_2 \\
&\quad= \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_1,x)W_{l+1}(\theta_1)b_{l+1}(\theta_1,x) - \Sigma_l(\theta_2,x)W_{l+1}(\theta_2)b_{l+1}(\theta_2,x)\|_2 \\
&\quad\le \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_1,x)W_{l+1}(\theta_1)b_{l+1}(\theta_1,x) - \Sigma_l(\theta_1,x)W_{l+1}(\theta_1)b_{l+1}(\theta_2,x)\|_2 \\
&\qquad+ \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_1,x)W_{l+1}(\theta_1)b_{l+1}(\theta_2,x) - \Sigma_l(\theta_2,x)W_{l+1}(\theta_2)b_{l+1}(\theta_2,x)\|_2 \\
&\quad\le \sup_{\theta_1\in\mathcal{D}} \|\Sigma_l(\theta_1,x)\|_{\text{op}} \sup_{\theta_1\in\mathcal{D}} \|W_{l+1}(\theta_1)\|_{\text{op}} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_1,x) - b_{l+1}(\theta_2,x)\|_2 \\
&\qquad+ \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_1,x)W_{l+1}(\theta_1) - \Sigma_l(\theta_2,x)W_{l+1}(\theta_2)\|_{\text{op}} \sup_{\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_2,x)\|_2 \\
&\quad\le \sup_{\theta_1\in\mathcal{D}} \|\Sigma_l(\theta_1,x)\|_{\text{op}} \sup_{\theta_1\in\mathcal{D}} \|W_{l+1}(\theta_1)\|_{\text{op}} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_1,x) - b_{l+1}(\theta_2,x)\|_2 \\
&\qquad+ \sup_{\theta_1,\theta_2\in\mathcal{D}} \|(\Sigma_l(\theta_1,x) - \Sigma_l(\theta_2,x))\, W_{l+1}(\theta_2)\|_{\text{op}} \sup_{\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_2,x)\|_2 \\
&\qquad+ \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_2,x)\,(W_{l+1}(\theta_1) - W_{l+1}(\theta_2))\|_{\text{op}} \sup_{\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_2,x)\|_2 .
\end{aligned} \tag{264}$$

Furthermore, we have that the following results hold.

(i) By Assumption 2.3 and Lemma H.3,

$$\sup_{\theta_1\in\mathcal{D}} \|\Sigma_l(\theta_1,x)\|_{\text{op}} \sup_{\theta_1\in\mathcal{D}} \|W_{l+1}(\theta_1)\|_{\text{op}} = \mathcal{O}(1),$$

with probability $1 - 2\exp(-cn_{L-1})$ over $W_{l+1}(\theta_0)$;

(ii) By inductive hypothesis,

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_1,x) - b_{l+1}(\theta_2,x)\|_2 \le C\,R\,\sqrt{\gamma d n_{L-1}},$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=l+2}^{L}$ and $x$;

(iii) By the same argument of the second statement in Lemma H.5 and again Lemma H.3,

$$\begin{aligned}
&\sup_{\theta_1,\theta_2\in\mathcal{D}} \|(\Sigma_l(\theta_1,x) - \Sigma_l(\theta_2,x))\, W_{l+1}(\theta_2)\|_{\text{op}} \\
&\qquad\le \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\phi'(g_l(\theta_1,x)) - \phi'(g_l(\theta_2,x))\|_2 \sup_{\theta_2\in\mathcal{D}} \|W_{l+1}(\theta_2)\|_{\text{op}} \\
&\qquad\le C\,R\,\sqrt{d},
\end{aligned}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l+1}$ and $x$;

(iv) By Lemma H.6,

$$\sup_{\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_2,x)\|_2 \le C\,\sqrt{\gamma n_{L-1}},$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^{L}$;

(v) By Assumption 2.3,

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_2,x)\,(W_{l+1}(\theta_1) - W_{l+1}(\theta_2))\|_{\text{op}} \le C\,R.$$

By combining (i)-(v) with (264), we conclude that

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_l(\theta_1,x) - b_l(\theta_2,x)\|_2 \le C\,R\,\sqrt{\gamma d n_{L-1}}, \tag{265}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $x$ and $(W_k(\theta_0))_{k=1}^{L}$, which concludes the proof. □

**Lemma H.8** (Difference of Jacobians in $\mathcal{D}$)**.** *Let $\theta_0$ be defined in* (18)*, $x \sim P_X$, and $\mathcal{D} = \mathcal{B}(\theta_0, R)$. Assume that $R = o(1)$ and that $\gamma > 1$. Then, we have*

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|J(\theta_1) - J(\theta_2)\|_{\text{op}} \leq C\, R\, d\sqrt{\gamma n_{L-1} N}, \tag{266}$$

*with probability at least $1 - CN\exp(-cn_{L-1})$ over $(x_i)_{i=1}^N$ and $\theta_0$, where $c, C$ are numerical constants.*

*Proof.* Pick $i \in [N]$. Then, we have

$$\begin{aligned}
&\sup_{\theta_1,\theta_2\in\mathcal{D}} \|(J(\theta_1))_{i:} - (J(\theta_2))_{i:}\|_2^2 \\
&\quad\leq \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|(F_l(\theta_1))_{:i} \otimes (B_{l+1}(\theta_1))_{:i} - (F_l(\theta_2))_{:i} \otimes (B_{l+1}(\theta_2))_{:i}\|_2^2 \\
&\quad= \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1, x_i) \otimes b_{l+1}(\theta_1, x_i) - f_l(\theta_2, x_i) \otimes b_{l+1}(\theta_2, x_i)\|_2^2 \\
&\quad\leq \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|(f_l(\theta_1, x_i) - f_l(\theta_2, x_i)) \otimes b_{l+1}(\theta_1, x_i)\|_2^2 \\
&\qquad+ \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_2, x_i) \otimes (b_{l+1}(\theta_1, x_i) - b_{l+1}(\theta_1, x_i))\|_2^2 \\
&\quad\leq \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1, x_i) - f_l(\theta_2, x_i)\|_2^2 \sup_{\theta_1\in\mathcal{D}} \|b_{l+1}(\theta_1, x_i)\|_2^2 \\
&\qquad+ \sum_{l=0}^{L-1} \sup_{\theta_2\in\mathcal{D}} \|f_l(\theta_2, x_i)\|_2^2 \sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_1, x_i) - b_{l+1}(\theta_1, x_i)\|_2^2 .
\end{aligned} \tag{267}$$

Since $x_i \sim P_X$, we can merge together the results from Lemmas H.4, H.5, H.6 and H.7 and obtain

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|(J(\theta_1))_{i:} - (J(\theta_2))_{i:}\|_2^2 \leq C\gamma R^2 d^2 n_{L-1}, \tag{268}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $x_i$ and $\theta_0$.

Therefore, we have

$$\begin{aligned}
\sup_{\theta_1,\theta_2\in\mathcal{D}} \|J(\theta_1) - J(\theta_2)\|_{\text{op}} &\leq \sup_{\theta_1,\theta_2\in\mathcal{D}} \|J(\theta_1) - J(\theta_2)\|_F \\
&\leq \sqrt{\sum_{i=1}^N \sup_{\theta_1,\theta_2\in\mathcal{D}} \|(J(\theta_1))_{i:} - (J(\theta_2))_{i:}\|_2^2} \\
&\leq C\, R\, d\sqrt{\gamma n_{L-1} N},
\end{aligned} \tag{269}$$

with probability $1 - CN\exp(-cn_{L-1})$ over $(x_i)_{i=1}^N$ and $\theta_0$. □

**Lemma H.9** (NTK spectrum in $\mathcal{D}$)**.** *Let $\theta_0$ be defined in* (18)*, $x \sim P_X$, and $\mathcal{D} = \mathcal{B}(\theta_0, R)$. Assume that $R = o(1)$ and that $\gamma > 1$. Then, we have*

$$\sup_{\theta\in\mathcal{D}} \|K(\theta)\|_{\text{op}} \leq C\,\gamma\, N\, d\, n_{L-1}, \tag{270}$$

*with probability at least $1 - C\,N\exp(-cn_{L-1})$ over $\theta_0$ and $(x_i)_{i=1}^N$, where $c, C$ are numerical constants. Furthermore,*

$$\inf_{\theta\in\mathcal{D}} \sigma_{\min}(J(\theta)) \geq c_1\sqrt{\gamma n_{L-2} n_{L-1}} - C_1\, R\, d\sqrt{\gamma n_{L-1} N}, \tag{271}$$

*with probability at least $1 - C\,Ne^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}$ over $\theta_0$ and $(x_i)_{i=1}^N$, where $c_1, C_1$ are also numerical constants.*

*Proof.* We have

$$\begin{aligned}
\sup_{\theta\in\mathcal{D}} \|K(\theta)\|_{\text{op}} &= \sup_{\theta\in\mathcal{D}} \left\| \sum_{l=0}^{L-1} F_l(\theta) F_l^\top(\theta) \circ B_{l+1}(\theta) B_{l+1}^\top(\theta) \right\|_{\text{op}} \\
&\leq \sum_{l=0}^{L-1} \sup_{\theta\in\mathcal{D}} \left\| F_l(\theta) F_l^\top(\theta) \circ B_{l+1}(\theta) B_{l+1}^\top(\theta) \right\|_{\text{op}} \\
&\leq \sum_{l=0}^{L-1} \sup_{\theta\in\mathcal{D}} \left\| F_l(\theta) F_l^\top(\theta) \right\|_{\text{op}} \sup_{\theta\in\mathcal{D}} \max_{i\in[N]} \left\| (B_{l+1}(\theta))_{i:} \right\|_2^2 \\
&\leq \sum_{l=0}^{L-1} \sup_{\theta\in\mathcal{D}} \|F_l(\theta)\|_F^2 \sup_{\theta\in\mathcal{D}} \max_{i\in[N]} \|b_{l+1}(\theta, x_i)\|_2^2 \\
&\leq \sum_{l=0}^{L-1} \left( \sum_{i=1}^{N} \sup_{\theta\in\mathcal{D}} \|f_l(\theta, x_i)\|_2^2 \right) \sup_{\theta\in\mathcal{D}} \max_i \|b_{l+1}(\theta, x_i)\|_2^2 .
\end{aligned} \tag{272}$$

By Lemma H.5, we have that $\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x_i)\|_2^2 \leq C\, d$ with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^l$ and $x_i$, for any $0 \leq l \leq L-1$ and $i \in [N]$. By Lemma H.6, we have that $\sup_{\theta\in\mathcal{D}} \|b_l(\theta, x_i)\|_2^2 \leq C\gamma n_{L-1}$ with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^L$, for any $l \in [L]$ and $i \in [N]$. Therefore, we obtain

$$\sup_{\theta\in\mathcal{D}} \|K(\theta)\|_{\text{op}} \leq C\,\gamma\, N\, d\, n_{L-1}, \tag{273}$$

with probability at least $1 - C\,N\exp(-cn_{L-1})$ over $\theta_0$ and $(x_i)_{i=1}^N$, which gives the first statement of the lemma.

By using Weyl's inequality, we get

$$\begin{aligned}
\inf_{\theta\in\mathcal{D}} \sigma_{\min}(J(\theta)) &\geq \sigma_{\min}(J(\theta_0)) - \sup_{\theta\in\mathcal{D}} \|J(\theta_1) - J(\theta_0)\|_{\text{op}} \\
&\geq c_1\sqrt{\gamma n_{L-2} n_{L-1}} - C_1\, R\, d\, \sqrt{\gamma\, n_{L-1}\, N},
\end{aligned} \tag{274}$$

where the last inequality follows from Lemma H.2 and Lemma H.8, and it holds with probability $1 - C\,Ne^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}$ over $(x_i)_{i=1}^N$ and $\theta_0$. This gives the second statement of the lemma and concludes the proof. □

Armed with Proposition H.1 and the intermediate estimates of Lemmas H.2-H.9, we are finally ready to prove Theorem 4.2.

*Proof of Theorem 4.2.* We show that there exist two absolute constants $\tilde{c}$ and $\tilde{C}$ such that

$$\alpha = \tilde{c}\sqrt{\gamma n_{L-2} n_{L-1}} \tag{275}$$

and

$$\beta = \tilde{C}\sqrt{\gamma N d n_{L-1}} \tag{276}$$

satisfy the two assumptions in Proposition H.1 with initialization $\tilde{\theta}_0 := \theta_0$, where $\theta_0$ is defined in (18). This holds with probability at least $1 - C\,Ne^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}$ over $(x_i)_{i=1}^N$ and $\theta_0$.

Recall from Proposition H.1 that $R$ is defined as $4\left\|F_L(\theta_0) - Y\right\|_2 / \alpha$, since we have set $\tilde{\theta}_0 = \theta_0$. For the moment, we assume that

$$R = \mathcal{O}\left( \sqrt{\frac{N}{\gamma n_{L-2} n_{L-1}}} \right), \tag{277}$$

and we will verify that this is the case later. Note that $\gamma = d^3 N^2 > 1$ and, hence, (277) and Assumption 2.5 imply that $R = o(1)$. Thus, we can apply Lemma H.9 and obtain

$$\inf_{\theta\in\mathcal{D}} \sigma_{\min}(J(\theta)) \geq c_1\sqrt{\gamma n_{L-2} n_{L-1}} - C_1\, R\, d\sqrt{\gamma n_{L-1} N} \geq \tilde{c}\sqrt{\gamma n_{L-2} n_{L-1}}, \tag{278}$$

with probability at least $1 - C\,N e^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}$ over $(x_i)_{i=1}^N$ and $\theta_0$, where the last inequality uses (277). This shows that the lower bound in (230) holds.

Now, by using (278), we verify that (277) holds. Recall that, by assumption of the theorem, $\|Y\|_2 = \Theta(\sqrt{N})$. Furthermore, by Lemma H.2, $F_L(\theta_0)$ is a vector of all zeros. Then,

$$R = \frac{4\,\|F_L(\theta_0) - Y\|_2}{\alpha} = \frac{4\,\|Y\|_2}{\alpha} = \mathcal{O}\left(\sqrt{\frac{N}{\gamma n_{L-2} n_{L-1}}}\right). \tag{279}$$

By Lemma H.9, we have that

$$\sup_{\theta\in\mathcal{D}} \|J(\theta)\|_{\text{op}} \le C\sqrt{\gamma N d n_{L-1}}, \tag{280}$$

with probability at least $1 - C\,N\exp(-cn_{L-1})$ over $\theta_0$. Thus, by our choice (276) of $\beta$, we obtain that the upper bound in (230) holds.

Next, we verify the second assumption of Proposition H.1. To do so, let us write

$$\frac{\alpha^2}{2\beta} = \frac{\tilde{c}^2 n_{L-2} n_{L-1}\gamma}{2\tilde{C}\sqrt{\gamma N d n_{L-1}}} = \Omega(\sqrt{n_{L-2}} N d), \tag{281}$$

where we have used Assumption 2.5. Thus,

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|J(\theta_1) - J(\theta_2)\|_{\text{op}} \le C\,R\,d\sqrt{\gamma n_{L-1} N} = \mathcal{O}\left(\frac{dN}{\sqrt{n_{L-2}}}\right) \le \frac{\alpha^2}{2\beta}, \tag{282}$$

with probability at least $1 - C\,N\exp(-cn_{L-1})$ over $(x_i)_{i=1}^N$ and $\theta_0$. Here, the first passage follows from Lemma H.8, in the second passage we use (277), and in the last one we use (281). This completes the proof of (231) and also of the theorem, since the desired claim follows from an application of Proposition H.1. □