A CONDITIONAL ORTHOGONALIZATION

Isometry after rotation. Our analysis is based on (Joudaki et al., 2023b, Corollary 3), which we restate in the following Lemma:

Lemma A.1. *For all non-degenerate matrices $X \in \mathbb{R}^{d\times d}$, we have:*

$$\mathcal{I}(\text{BN}(X)) \geq \left(1 + \frac{\textit{variance}\{\|X_{j\cdot}\|\}_{j=1}^{d}}{(\textit{mean}\{\|X_{j\cdot}\|\}_{j=1}^{d})^2}\right)\mathcal{I}(X)\,. \tag{12}$$

Lemma A.1 proves isometry bias of BN. The next lemma proves that isometry does not change under rotation

Lemma A.2 (Isometry after rotation)**.** *Let $X \in \mathbb{R}^{d\times d}$ and $W \sim \mathbb{O}_d$ be a random orthogonal matrix and $X' = WX$. Then,*

$$\mathcal{I}(\text{BN}(X')) \geq \left(1 + \frac{\textit{variance}\{\|X'_{j\cdot}\|\}_{j=1}^{d}}{(\textit{mean}\{\|X'_{j\cdot}\|\}_{j=1}^{d})^2}\right)\mathcal{I}(X)\,. \tag{13}$$

Proof. Using properties of the determinant, we have

$$\det(X'X'^{\top}) = \det(W)^2\det(XX^{\top}) = \det(XX^{\top}), \tag{14}$$

where the last equation holds since W is an orthogonal matrix. Furthermore,

$$\begin{aligned}
\text{Tr}(X'X'^{\top}) &= \text{Tr}(WXX^{\top}W^{\top}) && (15)\\
&= \text{Tr}(XX^{\top}\underbrace{W^{\top}W}_{=I}) && (16)\\
&= \text{Tr}(XX^{\top})\,. && (17)
\end{aligned}$$

Combining the last two equations with Lemma A.1 concludes the proof. □

Increasing isometry with rotations and BN. The last lemma proves the isometry bias does not decrease with rotation and BN. However, this does not prove a strict decrease in isometry with BN and rotation. The next lemma proves there exists an orthogonal matrix for which the isometry is strictly increasing.

Corollary A.3 (Increasing isometry)**.** *Let $X \in \mathbb{R}^{d\times d}$ and denote its singular value decomposition $X = U\textit{diag}\left(\{\sigma_i\}_{i=1}^{d}\right)V^{\top}$, where U and V are orthogonal matrices. Then, we have:*

$$\mathcal{I}(\text{BN}(U^{\top}H)) = 1\,. \tag{18}$$

Proof. Let $S = \text{diag}\left(\{\sigma_i\}_{i=1}^{d}\right)$ be the diagonal matrix containing the singular values of X. Then, we have:

$$\begin{aligned}
\text{BN}(U^{\top}X) &= \text{BN}(U^{\top}USV^{\top}) && (19)\\
&= \text{BN}(SV^{\top}) && (20)\\
&= \text{diag}(SV^{\top}VS)^{-\frac{1}{2}}SV^{\top} && (21)\\
&= S^{-1}SV^{\top} && (22)\\
&= V^{\top}, && (23)
\end{aligned}$$

which has maximum isometry 1 since it's an orthonormal matrix. □

Thus, there exists a rotation that increases isometry with BN for each non-orthogonal matrix. The proof of the last corollary is based on a straightforward application of Lemma 2.

Orthogonalization with randomness. The isometric is non-decreasing in Lemma 2 and provably increases for a certain rotation matrix (as stated in the last corollary). Hence, it is possible to increase isometry with random orthogonal matrices.

Theorem A.4. *Suppose $W \sim \mathbb{O}_d$ is a matrix drawn from $\mathbb{O}_d$ such that $W \overset{d}{=} WU$ for all orthogonal matrices U. Let $\{\lambda_i\}_{i=1}^d$ be the eigenvalues of $XX^\top$. Then,*

$$\mathbb{E}_W \left[\mathcal{I}(\textsc{Bn}(WX))|\, X\right] \geq \left(\frac{1}{1 - \frac{\sum_{k=1}^d (\lambda_k - 1)^2}{2d^2(d+2)}}\right) \mathcal{I}(X) \tag{24}$$

holds for all $X = \textsc{Bn}(\cdot)$, with equality for orthogonal matrices.

Remark 3. *Note that the assumption on $X = \textsc{Bn}(\cdot)$, can be viewed as an induction hypothesis, in that we can recursively apply this theorem to arrive at a quantitative rate at depth.*

Notably, $\sum_{i=1}^d \lambda_i = d$ if $X = \textsc{Bn}(\cdot)$. Hence, one can expect that $\sum_{i=1}^d \lambda_i^2 < d^2$ for all full random matrices X in form of $X = \textsc{Bn}(\cdot)$.

Proof. We need to compute the variance/mean ratio in Lemma 2. Let $X \in \mathbb{R}^{d\times d}$ have SVD decomposition $X = U\text{diag}\{\sigma_i\}_{i=1}^d V^\top$ where U and V are orthogonal matrix and $\sigma_i^2 = \lambda_i$. Since the distribution of W is invariant to transformations with orthogonal matrices, the distribution of W equates those of $X' = W\text{diag}\{\sigma_i\}V^\top$. It easy to check that

$$\|X'_{j\cdot}\| = \sqrt{\sum_{i=1}^d \sigma_i^2 W_{ji}^2} = \sqrt{\sum_{i=1}^d \lambda_i W_{ji}^2}\,. \tag{25}$$

Thus,

$$\sum_{j=1}^d \|X'_{j\cdot}\|^2 = \sum_{i=1}^d \lambda_i = d, \tag{26}$$

where the last equality holds due to the batch normalization. Thus, we have

$$\mathbb{E}\left[\frac{d\sum_{j=1}^d \|X'_{j\cdot}\|^2}{(\sum_{j=1}^d \|X'_{j\cdot}\|)^2}\right] = \mathbb{E}\left[\frac{d^2}{(\sum_j \|X'_{j\cdot}\|)^2}\right] \geq \frac{d^2}{\mathbb{E}\left[\left(\sum_{j=1}^d \|X'_{j\cdot}\|\right)^2\right]}\,. \tag{27}$$

We need to estimate

$$\mathbb{E}\left[\|X'_i\|\|X'_j\|\right] = \mathbb{E}\left[\left(\sum_{k=1}^d \lambda_k W_{ik}^2\right)^{\frac{1}{2}} \left(\sum_{k=1}^d \lambda_k W_{jk}^2\right)^{\frac{1}{2}}\right]\,. \tag{28}$$

Since square root function is concave, we have $\sqrt{x} \leq 1 + \frac{1}{2}(x-1)$. Thus

$$\mathbb{E}\left[\|X'_i\|\|X'_j\|\right] \leq \mathbb{E}\left[\left(1 + 0.5\sum_{k=1}^d (\lambda_k - 1)W_{ik}^2\right)\left(1 + 0.5\sum_{k=1}^d (\lambda_k - 1)W_{jk}^2\right)\right] \tag{29}$$

$$= 1 + \frac{1}{4}\sum_{k,q}(\lambda_k - 1)(\lambda_q - 1)\mathbb{E}\left[W_{ik}^2 W_{jq}^2\right] \tag{30}$$

$$= 1 + \frac{1}{4}\left[\sum_{k\neq q}(\lambda_k - 1)(\lambda_q - 1)\underbrace{\mathbb{E}\left[W_{ik}^2 W_{jq}^2\right]}_{E_1} + \sum_{k=q}(\lambda_k - 1)(\lambda_k - 1)\underbrace{\mathbb{E}\left[W_{ik}^2 W_{jk}^2\right]}_{E_2}\right], \tag{31}$$

where in the first equality we have used the fact that the cross terms reduce, where the expectations are applications of Weingarten calculus:

$$\mathbb{E}\left[0.5\sum_{k=1}^{d}(\lambda_k-1)W_{ik}^2+0.5\sum_{k=1}^{d}(\lambda_k-1)W_{jk}^2\right]=\mathbb{E}\left[0.5\sum_{k=1}^{d}\lambda_k W_{ik}^2+0.5\sum_{k=1}^{d}\lambda_k W_{jk}^2-1\right] \tag{32}$$

$$=0.5\sum_{k=1}^{d}\lambda_k\mathbb{E}\left[W_{ik}^2\right]+0.5\sum_{k=1}^{d}\lambda_k\mathbb{E}\left[W_{jk}^2\right]-1 \tag{33}$$

$$=\frac{0.5}{d}\sum_{k=1}^{d}\lambda_k+\frac{0.5}{d}\sum_{k=1}^{d}\lambda_k-1 \tag{34}$$

$$=0\,. \tag{35}$$

The main quantity we must compute is an expectation of polynomials taken over the Haar measure of the Orthogonal group $O(n)$. To carry out the computation, we make use of Weingarten calculus (Banica et al., 2011; Collins & Śniady, 2006; Weingarten, 1978). More specifically, we make use of the of the Weingarten formula, studied by (Collins & Śniady, 2006; Collins et al., 2022):

$$\int_{O(n)} r_{i_1j_1}r_{i_2j_2}\dots r_{i_{2d}j_{2d}}d\mu(O(n))=\sum_{\sigma\in\mathcal{M}_{2d}}\sum_{\sigma\in\mathcal{M}_{2n}}\Delta_\sigma(\boldsymbol{i})\Delta_\sigma(\boldsymbol{j})\mathrm{Wg}^O(\sigma^{-1}\tau), \tag{36}$$

where $\mu(O(n))$ is the Haar measure of Orthogonal group. For an in depth explanation of each quantity in the Weingarten formula, we refer the reader to Collins et al. (2022, Section 5.2).

The quantity we focus on is $\mathbb{E}_W\left[W_{ik}W_{ik}W_{jq}W_{jq}\right]$. We will do the computation on multiple cases, based on the equalities of k, q. Notice that $i\neq j$ in all cases. It suffices if we focus on the two distinct cases: $E_1=\mathbb{E}_W\left[W_{ik}^2W_{jq}^2\right]$ $(k\neq q)$ and $E_2=\mathbb{E}_W\left[W_{ik}^2W_{jk}^2\right]$ $(k=q)$.

We first compute E_1.

Following the procedure from Collins et al. (2022, Section 5.2), we take the index sequences to be $\boldsymbol{i}=(i,i,j,j)$ and $\boldsymbol{j}=(k,k,q,q)$. Similarly, we get $\Delta_\sigma(\boldsymbol{i})=\Delta_\tau(\boldsymbol{j})=1$ only if $\sigma=\{\{1,2\},\{3,4\}\}$ and $\tau=\{\{1,2\},\{3,4\}\}$.

Consdering σ,τ as permutations, we get:

$$\sigma=\begin{pmatrix}1&2&3&4\\1&2&3&4\end{pmatrix}, \tag{37}$$

$$\tau=\begin{pmatrix}1&2&3&4\\1&2&3&4\end{pmatrix}, \tag{38}$$

$$\sigma^{-1}\tau=\begin{pmatrix}1&2&3&4\\1&2&3&4\end{pmatrix}, \tag{39}$$

where $\sigma^{-1}\tau$ has coset-type $[1,1]$. Finally, we plug the results back into the formula and we obtain:

$$E_1=\mathbb{E}_W\left[W_{ik}^2W_{jq}^2\right]=\mathrm{Wg}^O([1,1])=\frac{d+1}{d(d+2)(d-1)},$$

where the last equality is based on the results in Section 7 of (Collins & Matsumoto, 2009).

We compute E_2.

Similar to the previous expression, we take the index sequences to be to be $\boldsymbol{i}=(i,i,j,j)$ and $\boldsymbol{j}=(k,k,k,k)$. Thus, we obtain $\Delta_\sigma(\boldsymbol{i})=\Delta_\tau(\boldsymbol{j})=1$ only if $\sigma=\{\{1,2\},\{3,4\}\}$ and $\tau_1=\{\{1,2\},\{3,4\}\}$, $\tau_2=\{\{1,3\},\{2,4\}\}$, $\tau_1=\{\{1,4\},\{2,3\}\}$. Similarly, we compute $\sigma^{-1}\tau_i$ for $i\in\{1,2,3\}$. Notice that σ is the identity permutation, thus yielding $\sigma^{-1}\tau_i=\tau_i$, with the coset-types $[1,1],[2],[2]$ respectively, for each $i\in\{1,2,3\}$.

Plugging back into the original equation, we obtain:

$$E_2 = \mathbb{E}_W\left[W_{ik}^2 W_{jk}^2\right] = \mathrm{Wg}^O([1,1]) + 2\mathrm{Wg}^O([2]) \tag{40}$$

$$= \frac{d+1}{d(d+2)(d-1)} + 2\frac{-1}{d(d+2)(d-1)} \tag{41}$$

$$= \frac{d-1}{d(d+2)(d-1)}\,. \tag{42}$$

Thus, plugging back into the original inequality, we obtain:

$$\mathbb{E}\|X_i'\|\|X_j'\| \leq 1 + \frac{1}{4}\left[\sum_{k\neq q}(\lambda_k-1)(\lambda_q-1)\frac{d+1}{d(d+2)(d-1)} + \sum_{k=q}(\lambda_k-1)(\lambda_k-1)\frac{d-1}{d(d+2)(d-1)}\right] \tag{43}$$

$$= 1 + \frac{1}{4d(d+2)(d-1)}\left[\underbrace{\sum_{k\neq q}(\lambda_k-1)(\lambda_q-1)}_{S_{\neq}} - \underbrace{\sum_{k=q}(\lambda_k-1)(\lambda_k-1)}_{S_{=}}\right] \tag{44}$$

$$= 1 - \frac{\sum_k(\lambda_k-1)^2}{2d(d+2)(d-1)}, \tag{45}$$

where we have used that $S_{\neq} + S_{=} = \sum_{k,q}(\lambda_k-1)(\lambda_q-1) = (\sum_{k=1}^d(\lambda_k-1))^2 = 0$ in the last equality.

Thus, we obtain:

$$\mathbb{E}\left[\left(\sum_j \|X_{j\cdot}'\|\right)^2\right] = \mathbb{E}\left[\sum_j \|X_{j\cdot}'\|^2 + 2\sum_{i<j}\|X_{i\cdot}'\|\|X_{j\cdot}'\|\right] \tag{46}$$

$$= d + 2\sum_{i<j}\mathbb{E}\left[\|X_{i\cdot}'\|\|X_{j\cdot}'\|\right] \tag{47}$$

$$\leq d + (d^2-d)\left(1 - \frac{\sum_k(\lambda_k-1)^2}{2d(d+2)(d-1)}\right) \tag{48}$$

$$= d^2 - \frac{\sum_k(\lambda_k-1)^2}{2(d+2)}\,. \tag{49}$$

□

Corollary A.5 (Isometry gap bound)**.** *Suppose the same setup as in Theorem 3. Then, we have:*

$$\mathbb{E}_W[\phi(X')|X] \leq \phi(X) + \log\left(1 - \frac{\sum_k(\lambda_k-1)^2}{2d^2(d+2)}\right). \tag{50}$$

Remark 4. *Notice that the term* $\frac{\sum_{k=1}^d(\lambda_k-1)^2}{2d^2(d+2)} = \mathcal{O}(\frac{1}{d})$*, yielding* $\log\left[1 - \frac{\sum_{k=1}^d(\lambda_k-1)^2}{2d^2(d+2)}\right] \leq 0$.

Proof. From Lemma 2, we know that:

$$-\log \mathcal{I}(\textsc{Bn}(X')) \leq -\log \mathcal{I}(X) - \log \frac{d^2}{\left(\sum_{j=1}^{d} \|X'_{j\cdot}\|\right)^2} \tag{51}$$

$$\implies -\log \mathcal{I}(\textsc{Bn}(X')) \leq -\log \mathcal{I}(X) - \log d^2 + \log \left(\sum_{j=1}^{d} \|X'_{j\cdot}\|\right)^2 \tag{52}$$

$$\implies \mathbb{E}_W[-\log \mathcal{I}(\textsc{Bn}(X'))|X] \leq -\log \mathcal{I}(X) - \log d^2 + \mathbb{E}_W\left[\log \left(\sum_{j=1}^{d} \|X'_{j\cdot}\|\right)^2 \middle| X\right] \tag{53}$$

$$\leq -\log \mathcal{I}(X) - \log d^2 + \log \mathbb{E}_W\left[\left(\sum_{j=1}^{d} \|X'_{j\cdot}\|\right)^2 \middle| X\right] \tag{54}$$

$$\leq -\log \mathcal{I}(X) - \log d^2 + \log \left(d^2 - \frac{\sum_k (\lambda_k - 1)^2}{2(d+2)}\right) \tag{55}$$

$$\leq -\log \mathcal{I}(X) + \log \left(1 - \frac{\sum_k (\lambda_k - 1)^2}{2d^2(d+2)}\right), \tag{56}$$

where in inequality 54 we have used the fact that $\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$ and in inequality 55 we have used the bound obtained in proof Theorem 3, equation 49. Thus, we obtain:

$$\mathbb{E}_W[\phi(X')|X] \leq \phi(X) + \log \left(1 - \frac{\sum_k (\lambda_k - 1)^2}{2d^2(d+2)}\right) . \tag{57}$$

□

B ISOMETRY GAP DECAY RATE

Before we start with the main part of our analysis, let us establish a simple result on the relation between isometry gap and orthogonality:

Lemma B.1 (Isometry gap and orthogonality)**.** *If $\phi(X) \leq \frac{c}{16d}$, then eigenvalues of $X^\top X$ are within $[1-c, 1+c]$.*

Note that, in order to simplify the calculations, we use the fact that $\frac{1}{d(d+2)} \approx \frac{1}{d^2}$ in the following proofs.

Based on the conditional expectation in Corollary A.5, we have:

$$\mathbb{E}\left[\phi(X^{\ell+1})|X^\ell\right] \leq \phi(X^\ell) - \frac{\sigma_\lambda(X^\ell)}{2d^2} \,. \tag{58}$$

Now, we prove a lemma that is conditioned on the previous layer isometry gap being smaller or larger than $\frac{1}{16d}$.

Lemma B.2 (Isometry gap conditional bound)**.** *For X^ℓ being the representations of an MLP under our setting, we have:*

$$\mathbb{E}\left[\phi(X^{\ell+1})\middle| X_\ell, \phi(X^\ell) \leq \frac{1}{16d}\right] \leq \phi(X^\ell)\left(1 - \frac{1}{2d^2}\right), \tag{59}$$

$$\mathbb{E}\left[\phi(X^{\ell+1})\middle| X_\ell, \phi(X^\ell) > \frac{1}{16d}\right] \leq \phi(X^\ell) - \frac{1}{32d^3} \,. \tag{60}$$

Proof of Lemma B.2. Let $\lambda_k = 1 + \epsilon_k$, and assume without loss of generality that $\sum_{k=1}^d \epsilon_k = 0$. Then, using the numerical inequality $\log(1+x) \geq x - x^2$, when $|x| \leq \frac{1}{2}$ we have:

$$\sigma_\lambda(X) = \frac{1}{d}\sum_{k=1}^d \epsilon_k^2 \tag{61}$$

$$\max_k |\epsilon_k| \leq \frac{1}{2} \implies \phi(X) = -\frac{1}{d}\sum_{k=1}^d \log(1+\epsilon_k) \leq -\frac{1}{d}\sum_{k=1}^d (\epsilon_k - \epsilon_k^2) = \sigma_\lambda(X) \,. \tag{62}$$

Altogether, we have

$$\max_i |\epsilon_i| \leq \frac{1}{2} \implies \phi(X) \leq \sigma_\lambda(X) \,. \tag{63}$$

Now, we can restate the condition in terms of an inequality on the isometry gap. Thus, we can write:

$$d\phi(X) = -\sum_{i=1}^d \log \lambda_i = -\sum_{i=1}^d \log(1+\epsilon_i) \geq -\sum_{i=1}^d \left(\epsilon_i - \frac{3\epsilon_i^2}{6+4\epsilon_i}\right) = \sum_{i=1}^d \frac{3\epsilon_i^2}{6+4\epsilon_i}, \tag{64}$$

where we used the fact that $\sum_i \lambda_i = d$ implying $\sum_i \epsilon_i = 0$ and also used the inequality $\log(1+x) \leq x - \frac{6+x}{6+4x}$ when $x \geq -1$ for ϵ_i's. Note that because $|\epsilon_i| \leq \frac{1}{2}$, we get $6 + 4\epsilon_i \geq 4$, all terms on the right-hand side are positive, implying that each term is bounded by the upper bound:

$$\frac{3\epsilon_i^2}{6+4\epsilon_i} \leq d\phi(X) \quad \forall i \in \{1, 2, \ldots, d\} \,. \tag{65}$$

By construction, we have $\epsilon_i \geq -1$ and $\frac{3\epsilon_i^2}{6+4\epsilon_i} \leq \frac{1}{16}$. Since $6 + 4\epsilon_i \geq 2$, we can multiply both sides by $6 + 4\epsilon_i$ and conclude $3\epsilon_i^2 - \frac{6+4\epsilon_i}{16} \leq 0$. We can now solve the quadratic equation and obtain $\frac{1-\sqrt{74}}{24} \leq \epsilon_i \leq \frac{1+\sqrt{74}}{24}$ which numerically becomes $-0.35 \leq \epsilon_i \leq 0.4$, implying $|\epsilon_i| < 0.5$.

By solving the quadratic equation above we can guarantee that

$$\phi(X) \leq \frac{1}{16d} \implies \max_i |\epsilon_i| \leq \frac{1}{2} \implies \phi(X) \leq \sigma_\lambda(X) \,. \tag{66}$$

Furthermore, we can restate the condition on maximum using:

$$\max_k |\epsilon_k| = \sqrt{\max_k \epsilon_k^2} \leq \sqrt{\sum_k \epsilon_k^2} = \sqrt{d\sigma_\lambda(X)} \tag{67}$$

and conclude that

$$\sigma_\lambda(X) \leq \frac{1}{4d} \implies \max_i |\epsilon_i| \leq \frac{1}{2} \implies \phi(X) \leq \sigma_\lambda(X)\,. \tag{68}$$

Using this statement, we have

$$\sigma_\lambda(X) \leq \frac{1}{16d} \implies \phi(X) \leq \sigma_\lambda(X) \implies \phi(X) \leq \frac{1}{16d}\,. \tag{69}$$

If we negate and flip the two sides we arrive at

$$\phi(X) > \frac{1}{16d} \implies \sigma_\lambda(X) > \frac{1}{16d}\,. \tag{70}$$

Thus, we can simplify the recurrence

$$\mathbb{E}[\phi(X^{\ell+1})|X^\ell] \leq \phi(X^\ell) - \frac{\sigma_\lambda(X^\ell)}{2d^2} \tag{71}$$

as follows

$$\mathbb{E}\left[\phi(X^{\ell+1})\middle| X^\ell, \phi(X^\ell) \leq \frac{1}{16d}\right] \leq \phi(X^\ell)\left(1 - \frac{1}{2d^2}\right), \tag{72}$$

$$\mathbb{E}\left[\phi(X^{\ell+1})\middle| X^\ell, \phi(X^\ell) > \frac{1}{16d}\right] \leq \phi(X^\ell) - \frac{1}{32d^3}, \tag{73}$$

where we used equation 66 in the first one and equation 70 in the second one. □

Proof of Theorem 1. From Lemma A.1, we know that $\phi(X^0) \geq \phi(X^1) \geq \cdots \geq \phi(X^L) \geq 0$, for any layer $0 \leq \ell \leq L$. Thus, we get using equation 60:

$$\mathbb{E}\left[\phi(X^{\ell+1})\middle| X^\ell, \phi(X^\ell) > \frac{1}{16d}\right] \leq \phi(X^\ell) - \frac{1}{32d^3} \tag{74}$$

$$= \left(1 - \frac{1}{32d^3\phi(X^\ell)}\right)\phi(X^\ell) \tag{75}$$

$$\leq \left(1 - \frac{1}{32d^3\phi(X^0)}\right)\phi(X^\ell), \tag{76}$$

where in the last step we have used the fact that $\phi(X^\ell) \leq \phi(X^0)$.

Thus, we can combine equation 59 and equation 60 and obtain:

$$\mathbb{E}[\phi(X^{\ell+1})|X^{\ell}] = \mathbb{E}\left[\phi(X^{\ell+1})\middle| X^{\ell}, \phi(X^{\ell}) \leq \frac{1}{16d}\right] \mathbf{1}_{\phi(X^{\ell})\leq \frac{1}{16d}} \tag{77}$$

$$+ \mathbb{E}\left[\phi(X^{\ell+1})\middle| X^{\ell}, \phi(X^{\ell}) > \frac{1}{16d}\right] \mathbf{1}_{\phi(X^{\ell})> \frac{1}{16d}} \tag{78}$$

$$\leq \max\left(\mathbb{E}\left[\phi(X^{\ell+1})\middle| X^{\ell}, \phi(X^{\ell}) \leq \frac{1}{16d}\right], \mathbb{E}\left[\phi(X^{\ell+1})\middle| X^{\ell}, \phi(X^{\ell}) > \frac{1}{16d}\right]\right) \tag{79}$$

$$\leq \max\left(1 - \frac{1}{2d^2}, 1 - \frac{1}{32d^3\phi(X^0)}\right)\phi(X^{\ell}) \tag{80}$$

$$= \left(1 - \min\left(\frac{1}{2d^2}, \frac{1}{32d^3\phi(X^0)}\right)\right)\phi(X^{\ell}) \tag{81}$$

$$= \left(1 - \frac{1}{\max(2d^2, 32d^3\phi(X^0))}\right)\phi(X^{\ell}) \tag{82}$$

$$\leq \exp\left[-\underbrace{\frac{1}{\max(2d^2, 32d^3\phi(X^0))}}_{k}\right]\phi(X^{\ell}) \tag{83}$$

$$= \exp\left(-\frac{1}{k}\right)\phi(X^{\ell})\,. \tag{84}$$

By iterated expectations over X^{ℓ} we get:

$$\mathbb{E}[\phi(X^{\ell+1})] \leq \exp\left(-\frac{1}{k}\right)\mathbb{E}[\phi(X^{\ell})] \leq \exp\left(-\frac{\ell}{k}\right)\phi(X^0)\,. \tag{85}$$

Note that since $\max(2d^2, 32d^3\phi(X^0)) \leq 2d^2(1 + 16d\phi(X^0))$, we can conclude the proof. □

C GRADIENT NORM BOUND

In the following section, we denote by $H^\ell = W^\ell X^\ell$ the pre-normalization values. Moreover, we define as $\mathcal{F}_L : \mathbb{R}^{d\times d} \to \mathbb{R}^{C\times d}$, where C is the number of output classes, as the functional composition of an L layers MLP, following the update rule defined in equation 3, i.e.:

$$\mathcal{F}_L(X_L) = \textsc{Bn}(W^L \mathcal{F}_{L-1}(X_{L-1})) \,. \tag{86}$$

Let us restate the theorem for completeness of the appendix:

Theorem C.1 (Restated Thm. 5)**.** *For any $\mathcal{O}(1)$-Lipschitz loss function L and non-degenerate input X_0, we have:*

$$\log \left\| \frac{\partial \mathcal{L}}{\partial W^\ell} \right\| \lesssim d^5 \left(\phi(X_0)^3 + 1 - e^{-\frac{L}{32d^4}} \right) \tag{87}$$

holds for all $\ell \leq L$, where possibly $L \to \infty$.

In particular, the following lemma guarantees that the Lipschitz conditions are met for practical loss functions:

Lemma C.2. *In a classification setting, cross entropy and mean squared error losses are $\mathcal{O}(1)$-Lipschitz.*

The main idea for the feasibility of this theorem is the presence of perfectly isometric weight matrices that are orthonormal, and the linear activation that does not lead to vanishing or exploding gradients. The only remaining layers to be analyzed are the batch normalization layers. Thus, our main goal is to show that the sum of log-norm gradient of BN layers remains bounded even if the network has infinite depth. To do so, we shall relate the norm of the gradient of those layers to the isometry gap of representations, and use the bounds from the previous section to establish that the log-norm sum is bounded.

Proof of Thm. C.1. Now, considering an L layer deep model, where L can possibly be $L \to \infty$, we can finalize the proof of Theorem 5. Consider an MLP model as defined in equation 3. Let $H^L = W^L X^L$ be the logits of the model, where $H^L \in \mathbb{R}^{C\times d}$, $W^L \in \mathbb{R}^{C\times d}$, W^L is an orthogonal matrix and C is the number of output classes. Denote as $\mathcal{L}(H^L, y)$ the loss of model for an input matrix, with ground truth y. Then, applying the chain rule, we have:

$$\frac{\partial \mathcal{L}}{\partial W^\ell} = \frac{\partial \mathcal{L}}{\partial H^L} \frac{\partial H^L}{\partial X^L} \frac{\partial X^L}{\partial X^{L-1}} \cdots \frac{\partial X^{\ell+2}}{\partial X^{\ell+1}} \frac{\partial X^{\ell+1}}{\partial H^\ell} \frac{\partial H^\ell}{\partial W^\ell} \,. \tag{88}$$

By taking the logarithm of the norm of each factor and applying Lemma C.12, we get:

$$\log \left\| \frac{\partial \mathcal{L}}{\partial W^\ell} \right\| \leq \log \left\| \frac{\partial \mathcal{L}}{\partial H^L} \right\| + \log \underbrace{\left\| \frac{\partial H^L}{\partial X^L} \right\|}_{\|W^L\|} + \sum_{k=\ell+1}^{L} \log \left\| \frac{\partial X^{k+1}}{\partial X^k} \right\| + \log \underbrace{\left\| \frac{\partial X^{\ell+1}}{\partial H^\ell} \right\|}_{\|J_{\textsc{Bn}}(H^\ell)\|} + \log \left\| \frac{\partial H^\ell}{\partial W^\ell} \right\| \tag{89}$$

$$\leq \log \left\| \frac{\partial \mathcal{L}}{\partial H^L} \right\| + \sum_{k=\ell}^{L} \log \left\| J_{\textsc{Bn}}(H^k) \right\| + \log \left\| \frac{\partial H^\ell}{\partial W^\ell} \right\| \,. \tag{90}$$

where $\log \underbrace{\left\| \frac{\partial H^L}{\partial X^L} \right\|}_{\|W^L\|} = \log \left\| W^L \right\| = 0$, since the orthogonal matrix W^L has operator norm 1.

Since $\frac{\partial H^\ell}{\partial W^\ell} = X^\ell$ and $X^\ell = \textsc{Bn}(H^{\ell-1})$ is batch normalized, this means that $\left\| \frac{\partial H^\ell}{\partial W^\ell} \right\| \leq d$. Thus, the main part is to bound the Jacobian log-norms, which is provided by the following lemma:

Lemma C.3. *We have*

$$\sum_{k=1}^{L} \log \left\| J_{\textsc{Bn}}(H^k) \right\| \lesssim d^5 (\phi(X_0)^3 + 1 - e^{-\frac{L}{32d^4}}) \,. \tag{91}$$

Finally, we can plug the bound from Lemma C.3 in equation 90 and obtain the conclusion:

$$\log \left\| \frac{\partial \mathcal{L}}{\partial W^\ell} \right\| \lesssim \log \left\| \frac{\partial \mathcal{L}}{\partial H^L} \right\| + \log d + d^5 \left(\phi(X_0)^3 + 1 - e^{-\frac{L}{32d^4}} \right) . \tag{92}$$

Note that, for $L \to \infty$, we get:

$$\log \left\| \frac{\partial \mathcal{L}}{\partial W^\ell} \right\| \lesssim \log \left\| \frac{\partial \mathcal{L}}{\partial H^L} \right\| + \log d + d^5 (\phi(X_0)^3 + 1) . \tag{93}$$

In order to conclude the bound, it suffices to show that the norm of the gradient of the loss with respect to the logits is bounded, which is the objective of Lemma C.2. □

Proof of Lemma C.3. The proof of this lemma is chiefly relying on the following bound on the Jacobian of batch normalization layers, which we will state and prove beforehand.

Lemma C.4 (Log-norm bound)**.** *If $X \in \mathbb{R}^{d \times d}$ is the input to a* BN *layer, its Jacobian operator norm is bounded by*

$$\log \|J_{\text{BN}}(X)\|_{op} \leq d\phi(X) + 1 . \tag{94}$$

Furthermore, if $\phi(X) \leq \frac{1}{16d}$, then we have

$$\log \|J_{\text{BN}}(X)\|_{op} \leq 2\sqrt{d\phi(X)} . \tag{95}$$

Based on the lemma above, we shall define S as the hitting time, corresponding to the first layer in our case, that the isometry gap drops below the critical value of $\frac{1}{16d}$:

$$S = \min \left\{ \ell : \phi(X^\ell) \leq \frac{1}{16d} \right\} . \tag{96}$$

So, we first bound the total log-grad norm for layers 1 up to S, and subsequently $S+1$ up to L:

$$\log \|J_{\mathcal{F}_L}(X)\|_{op} \leq \sum_{\ell=1}^{L} \log \|J_{\text{BN}}(X^\ell)\|_{op} \tag{97}$$

$$\leq \sum_{\ell=1}^{S} (d\phi(X^\ell) + 1) + 2 \sum_{\ell=S+1}^{L} \sqrt{d\phi(X^\ell)} \tag{98}$$

$$\leq \sum_{\ell=1}^{S} (d\phi(X^0) + 1) + 2 \sum_{\ell=S+1}^{L} \sqrt{d\phi(X^\ell)} \tag{99}$$

$$= S(d\phi(X^0) + 1) + 2 \sum_{\ell=S+1}^{L} \sqrt{d\phi(X^\ell)}, \tag{100}$$

where we have used that $\phi(X^0)$ as an upper bound on $\phi(X^\ell)$.

Thus, taking expectation we get:

$$\mathbb{E} \log \|J_{\mathcal{F}_L}(X)\|_{op} \leq (d\phi(X^0) + 1)\mathbb{E}[S] + 2 \sum_{\ell=S+1}^{L} \mathbb{E}\sqrt{d\phi(X^\ell)} . \tag{101}$$

Note that S is a random variable, which is why the expectation over the number of layers appears at the last line. Thus, we can bound the log-norm by bounding $\mathbb{E}[S]$ and the summation separately.

Lemma C.5 (stopping time bound)**.** *We have $\mathbb{E}[S] \lesssim 512d^4\phi(X^0)^2$ if $\phi(X_0) > \frac{1}{16d}$, and $\mathbb{E}[S] = 0$ if $\phi(X_0) \leq \frac{1}{16d}$.*

Lemma C.6 (second phase bound)**.** *We have*

$$\sum_{\ell=S+1}^{L} \mathbb{E}\sqrt{d\phi(X^\ell)} \leq 32d^{4.5}\phi(X_0)^{0.5} \left(1 - e^{-\frac{L}{32d^4}} \right) .$$

Thus, we have the following 2 cases, based on whether $\phi(X_0)$ is below or over the $\frac{1}{16d}$ threshold. If we plug the bounds in equation 101 we get the following.

If $\phi(X_0) \leq \frac{1}{16d}$, then:

$$\mathbb{E} \log \|J_{\mathcal{F}_L}(X)\|_{op} \leq 32d^{4.5}\phi(X_0)^{0.5}\left(1 - e^{-\frac{L}{32d^4}}\right) \tag{102}$$

$$\lesssim d^{4.5}\phi(X_0)^{0.5}\left(1 - e^{-\frac{L}{32d^4}}\right), \tag{103}$$

and if $\phi(X_0) > \frac{1}{16d}$ then:

$$\mathbb{E} \log \|J_{\mathcal{F}_L}(X)\|_{op} \leq 512d^4\phi(X_0)^2(1 + d\phi(X_0)) + 32d^4\left(1 - e^{-\frac{L}{32d^4}}\right) \tag{104}$$

$$\lesssim d^5\phi(X_0)^3 + d^4\left(1 - e^{-\frac{L}{32d^4}}\right) \tag{105}$$

$$\lesssim d^4\left(d\phi(X_0)^3 + 1 - e^{-\frac{L}{32d^4}}\right). \tag{106}$$

In fact, the maximum of the two bounds is

$$\mathbb{E} \log \|J_{\mathcal{F}_L}(X)\|_{op} \lesssim d^5\left(\phi(X_0)^3 + 1 - e^{-\frac{L}{32d^4}}\right). \tag{107}$$

□

Proof of Lemma C.6. By the bound in Lemma B.2, we have

$$\mathbb{E}\left[\phi(X^{\ell+1})\middle|\phi(X^\ell) \leq \frac{1}{16d}\right] \leq \phi(X^\ell)\left(1 - \frac{1}{2d^2}\right). \tag{108}$$

Since we assumed $\ell \geq S$, the conditional inequality always holds and thus we have the Markov bound

$$\ell \geq S \implies q := \Pr\left\{\phi(X^{\ell+1}) \geq \left(1 - \frac{1}{4d^2}\right)\phi(X^\ell)\right\} \leq \frac{1 - \frac{1}{2d^2}}{1 - \frac{1}{4d^2}} \leq 1 - \frac{1}{4d^2}. \tag{109}$$

We define as failure the event $\bar{A} = \{\phi(X^{\ell+1}) \geq (1 - \frac{1}{4d^2})\phi(X^\ell)\}$ with probability q, and conversely as success the event A with probability $1 - q$. In other words, the probability that $\phi(X^{\ell+1})$ does not decrease by at least a factor of $1 - \frac{1}{4d^2}$ is bounded by the failure probability $1 - \frac{1}{4d^2}$.

Since $\phi(X^{\ell+1}) \leq \phi(X^\ell)$, then under the assumption that $\ell \geq S$ we can upper bound $\sqrt{\phi(X^{\ell+1})}$ with $\sqrt{\phi(X^\ell)}$ in case of failure with probability q, and with $\sqrt{(1 - \frac{1}{4d^2})\phi(X^\ell)}$ in case of success

with probability $1-q$:

$$\mathbb{E}\left[\sqrt{\phi(X^{\ell+1})}|\phi(X^{\ell})\right] \tag{110}$$

$$= \mathbb{E}\left[\sqrt{\phi(X^{\ell+1})}|\phi(X^{\ell}), \bar{A}\right] q + \mathbb{E}\left[\sqrt{\phi(X^{\ell+1})}|\phi(X^{\ell}), A\right](1-q) \tag{111}$$

$$\leq \sqrt{\phi(X^{\ell})}q + \sqrt{\phi(X^{\ell})\left(1-\frac{1}{4d^2}\right)}(1-q) \tag{112}$$

$$= \sqrt{\phi(X^{\ell})}\left(q + \sqrt{1-\frac{1}{4d^2}}(1-q)\right) \tag{113}$$

$$= \sqrt{\phi(X^{\ell})}\left(\sqrt{1-\frac{1}{4d^2}} + \left(1-\sqrt{1-\frac{1}{4d^2}}\right)q\right) \qquad \text{monotonic in } q \tag{114}$$

$$\leq \sqrt{\phi(X^{\ell})}\left(\sqrt{1-\frac{1}{4d^2}} + \left(1-\sqrt{1-\frac{1}{4d^2}}\right)\left(1-\frac{1}{4d^2}\right)\right) \qquad \text{plug } q \leq 1-\frac{1}{4d^2} \tag{115}$$

$$= \sqrt{\phi(X^{\ell})}\left(1-\frac{1}{4d^2} + \sqrt{1-\frac{1}{4d^2}}\frac{1}{4d^2}\right) \qquad \text{rearranging terms} \tag{116}$$

$$\leq \sqrt{\phi(X^{\ell})}\left(1-\frac{1}{4d^2} + \left(1-\frac{1}{8d^2}\right)\frac{1}{4d^2}\right) \qquad \sqrt{1-x} \leq 1-\frac{x}{2} \text{ for } x \geq 0 \tag{117}$$

$$= \sqrt{\phi(X^{\ell})}\left(1-\frac{1}{32d^4}\right) . \tag{118}$$

Thus, for $\ell \geq S$, we have

$$\mathbb{E}\sqrt{\phi(X^{\ell+1})} = \mathbb{E}_{X^{\ell}}\mathbb{E}\left[\sqrt{\phi(X^{\ell+1})}|\phi(X^{\ell})\right] \leq \mathbb{E}\sqrt{\phi(X^{\ell})}\left(1-\frac{1}{32d^4}\right) . \tag{119}$$

The summation starts from below $\sqrt{d\phi(X_S)}$, and will decay by rate $1-\frac{1}{32d^4}$, which is upper bounded by the geometric series:

$$\sqrt{d\phi(X_S)}\sum_{k=0}^{L}\left(1-\frac{1}{32d^4}\right)^k \leq \sqrt{d\phi(X_0)}32d^4\left(1-\left(1-\frac{1}{32d^4}\right)^{L+1}\right) \tag{120}$$

$$\leq 32d^{4.5}\phi(X_0)^{0.5}\left(1-e^{-\frac{L}{32d^4}}\right) . \tag{121}$$

□

Proof of Lemma C.5. By Lemma B.2 we have

$$\Pr\left\{\phi(X^{\ell}) \geq \frac{1}{16d}\right\} \leq \exp\left(-\frac{\ell}{\max(2d^2, 32d^3\phi(X^0))}\right)16d\phi(X^0) . \tag{122}$$

Thus, we have

$$\Pr\{S \geq \ell\} \leq \exp\left(-\frac{\ell}{\max(2d^2, 32d^3\phi(X^0))}\right)16d\phi(X^0) . \tag{123}$$

Since S is a non-negative integer valued random variable, we can thus bound $\mathbb{E}[S]$ as:

$$\mathbb{E}[S] = \sum_{\ell=1}^{\infty} P\{S \geq \ell\} \tag{124}$$

$$\leq 16d\phi(X^0) \sum_{\ell=1}^{\infty} \exp\left(\frac{-\ell}{k}\right) \tag{125}$$

$$= 16d\phi(X^0) \frac{1}{\exp(\frac{1}{k}) - 1} \tag{126}$$

$$\leq 16d\phi(X^0)k \tag{127}$$

$$= 16d\phi(X^0) \max(2d^2, 32d^3\phi(X^0)) \tag{128}$$

$$\lesssim 512d^4\phi(X^0)^2 \,. \tag{129}$$

□

1 Proof of Lemma C.4: Bounding BN grad-norm with isometry gap

The proof of the Lemma relies on two main observations that are crystallized in the following lemmas that first establish a bound on Jacobian operator norm based on the inverse of smallest eigenvalue, and then establish a lower bound for the smallest eigenvalue using the isometry gap.

Lemma C.7. *Let $X \in \mathbb{R}^{d\times d}$ and let $\{\lambda_i\}_{i=1}^d$ be the eigenvalues of $XX^\top$. Then, we have that:*

$$\|J_{\text{BN}}(X)\|_{op}^2 \leq \frac{1}{\lambda_d}, \tag{130}$$

where J_{BN} is the Jacobian of the $\text{BN}(\cdot)$ operator.

Using the above lemma we have $\log \|J_{\text{BN}}(X)\|_{op} \leq -\log \lambda_d$. The following lemma upper bounds this quantity using isometry gap:

Lemma C.8. *The minimum eigenvalue of a Gram matrix that is trace-normalized is lower-bounded by the isometry gap as $-\log \lambda_d \leq d\phi(X) + 1$. Furthermore, if $\phi(X) \leq \frac{1}{16d}$, then $-\log \lambda_d \leq 2\sqrt{d\phi(X)}$.*

Plugging these two values we have the bounds

$$\log \|J_{\text{BN}}(X)\|_{op} \leq d\phi(X) + 1, \tag{131}$$

$$\phi(X) \leq \frac{1}{16d} \implies \log \|J_{\text{BN}}(X)\|_{op} \leq 2\sqrt{d\phi(X)} \,. \tag{132}$$

Now we can turn our attention to the proof of the Lemmas used in the proof. The proof of relationship between minimum eigenvalue and isometry gap is obtained by merely a few numerical inequalities:

Proof of Lemma C.8. Let $\{\lambda_i\}_{i=1}^d$ be the eigenvalues of $X^\top X$. Since the matrix is trace-normalized, we have $\sum_{k=1}^d \lambda_k = d$.

The arithmetic mean of the top $d-1$ values can be written as

$$\frac{1}{d-1} \sum_{k=1}^{d-1} \lambda_k = 1 + \frac{1-\lambda_d}{d-1} \,. \tag{133}$$

Thus, we have that their geometric mean is bounded by the same value. Therefore, we have the following bound:

$$\prod_{k=1}^{d} \lambda_k \leq \lambda_d \left(1 + \frac{1-\lambda_d}{d-1}\right)^{d-1} \tag{134}$$

$$\implies d \log \mathcal{I}(X) \leq \log(\lambda_d) + (d-1) \log\left(1 + \frac{1-\lambda_d}{d-1}\right), \tag{135}$$

where in the second inequality we have taken logarithm of both sides. Now, we can apply the numerical inequality $\log(1+x) \leq x$ to conclude:

$$-d\phi(X) \leq \log \lambda_d + 1 - \lambda_d \,. \tag{136}$$

Since λ_d is non-negative, this clearly implies the first inequality: $-\log \lambda_d \leq d\phi(X) + 1$.

For the second inequality first we use the numerical inequality $\log(x)+1-x \leq -\frac{(x-1)^2}{2}, \forall x \in [0,1]$ to conclude

$$\frac{(1-\lambda_d)^2}{2} \leq d\phi(X) \tag{137}$$

$$\implies \lambda_d \geq 1 - \sqrt{2d\phi(X)} \tag{138}$$

$$\implies -\log \lambda_d \leq -\log(1 - \sqrt{2d\phi(X)}) \,. \tag{139}$$

We can now use the inequality $-\log(1-x) \leq \sqrt{2}x$ for $0 \leq x \leq \frac{1}{2}$ to conclude that

$$-\log \lambda_d \leq 2\sqrt{d\phi(X)} \tag{140}$$

when $2\sqrt{d\phi(X)} \leq \frac{1}{2}$, which is equivalent to $\phi(X) \leq \frac{1}{16d}$. □

For proving Lemma C.4, we first analyze BN operator on a row, and then invoke this bound and the special structure J_{BN} to derive the main proof.

Lemma C.9. *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ defined as $f(x) = \frac{x}{\|x\|}$ be the elementwise normalization of the x. Then:*

$$J_f(x) = \frac{1}{\|x\|} I_{d^2} - \frac{1}{\|x\|^3} x \otimes x, \tag{141}$$

where $\otimes$ is the outer product.

Proof. To begin, notice that for $x \in \mathbb{R}^d$ we have $\frac{\partial \|x\|}{\partial x} = \frac{x}{\|x\|}$. Denote by $y_i := [f(x)]_i = \frac{x_i}{\|x\|}$. Then the Jacobian entries become:

$$\frac{\partial y_i}{\partial x_i} = \frac{\|x\| - \frac{x_i}{\|x\|} x_i}{\|x\|^2} = \frac{1}{\|x\|} - \frac{1}{\|x\|^3} x_i x_i, \tag{142}$$

$$\frac{\partial y_i}{\partial x_j} = \frac{-\frac{x_j}{\|x\|} x_i}{\|x\|^2} = -\frac{1}{\|x\|^3} x_i x_j \,. \tag{143}$$

Assembling the equations into matrix form, we obtain:

$$J_f(x) = \frac{1}{\|x\|} I_{d^2} - \frac{1}{\|x\|^3} x \otimes x \,. \tag{144}$$

□

Corollary C.10. *The $J_f(x)$ has the eigenvalue $\frac{1}{\|x\|}$ with multiplicity $d^2 - 1$ and 0 with multiplicity 1.*

Lemma C.11. *Let $XX^\top = \sum_{i=1}^d \lambda_i u_i u_i^\top$, where $XX^\top = U\Lambda U^\top$ is the eigendecomposition of $XX^\top$. Then, we have that $\min_j \|X_{j\cdot}\|^2 \geq \min_k \lambda_k$.*

Proof.

$$\|X_{j\cdot}\|^2 = (XX^\top)_{jj} = \sum_{i=1}^d \lambda_i u_{ji}^2 \geq \min_k \lambda_k \sum_{i=1}^d u_{ji}^2 = \min_k \lambda_k, \tag{145}$$

where in the last equality we have used the fact that U orthogonal. Since this is true for all rows j, we get $\min_j \|X_{j\cdot}\|^2 \geq \min_k \lambda_k$. □

Having established the above properties, we are equipped to prove that the Jacobian of a BN layer is bounded by the inverse of the minimum eigenvalue of its input Gram matrix.

Proof of Lemma C.7. To begin, notice that since $\text{BN} : \mathbb{R}^{d^2} \to \mathbb{R}^{d^2}$, implies that $J_{\text{BN}}(X) \in \mathbb{R}^{d^2 \times d^2}$. Denote $X' = \text{BN}(X)$. Since the normalization happens on each row independently of the other rows, the only non-zero derivatives in the Jacobian correspond to changes in output row i with regards to the same input row i, i.e.:

$$\frac{\partial X'_{ik}}{\partial X_{jl}} = 0, \quad \forall i \neq j, \forall k, l \; . \tag{146}$$

This creates a block-diagonal structure in $J_{\text{BN}}(X)$, with d blocks on the main diagonal, where each block has size $d \times d$ and is equal to $J_f(X_{i\cdot})$, where f is as defined in Lemma C.9. Therefore, due to the block-diagonal structure, we know that:

$$\lambda \left[J_{\text{BN}}(X) \right] = \bigcup_{i=1}^{d} \lambda \left[J_f(X_{i\cdot}) \right], \tag{147}$$

where $\lambda[\cdot]$ denotes the eigenvalue spectrum. From Corollary C.10, we know that

$$\lambda \left[J_{\text{BN}}(X) \right] = \left\{ \frac{1}{\|X_{i\cdot}\|} \right\}_{i=1}^{d} \cup \{0\} \tag{148}$$

with their respective multiplicites. Finally, using Lemma C.11, this implies:

$$\|J_{\text{BN}}(X)\|_2^2 = \left[\max_i \frac{1}{\|X_{i\cdot}\|} \right]^2 = \left[\frac{1}{\min_i \|X_{i\cdot}\|} \right]^2 \leq \frac{1}{\min_k \lambda_k} \; . \tag{149}$$

□

Proof of Lemma C.2. Assuming the the logits are passed through a softmax layer, we analyze the case of Cross Entropy Loss for one sample i in a C-classes classification problem. Denoting $z_i = H^L_{i\cdot}$, we have:

$$\mathcal{L}(z_i, y_i) = -\sum_{i=1}^{C} y_i \log p_i, \tag{150}$$

where $p_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$ is the probability vector for sample i after passing through the softmax function.

Computing the partial derivatives, we obtain:

$$\frac{\partial p_i}{\partial z_i} = p_i(1 - p_k), \quad i = k \tag{151}$$

$$\frac{\partial p_i}{\partial z_k} = -p_i p_k, \quad i \neq k \; . \tag{152}$$

Finally, we can compute the gradient of the loss with respect to the logits:

$$\nabla_{z_k} \mathcal{L} = \sum_{i=1}^{C} \left(-y_i \frac{\partial \log(p_i)}{\partial z_k} \right) \tag{153}$$

$$= \sum_{i=1}^{C} \left(-y_i \frac{1}{p_i} \frac{\partial p_i}{\partial z_k} \right) \tag{154}$$

$$= \sum_{i \neq k} (y_i p_k) + (-y_k(1 - p_k)) \tag{155}$$

$$= p_k \sum_{i \neq k} y_i - y_k(1 - p_k) \tag{156}$$

$$\implies \|\nabla_{z_k} \mathcal{L}\| \leq \left\| p_k \sum_{i \neq k} y_k \right\| + \|y_k(1 - p_k)\| \leq 2 \; . \tag{157}$$

Since the gradient of the loss with regard to each sample is bounded, we can conclude that the operator norm of the Jacobian of the loss with regards to the logits matrix H^L is also bounded.

In a similar analysis, we now shift our attention towards the Mean Squared Error (MSE) loss:

$$\mathcal{L}(z_i, y_i) = \frac{1}{C}\sum_{i=1}^{C}(y_i - p_i)^2 \, .$$

We want to compute the gradient of the loss with respect to each logit z_k:

$$\|\nabla_{z_k}\mathcal{L}\| = \frac{1}{C}\sum_{i=1}^{C} 2(y_i - p_i)\frac{\partial(-p_i)}{\partial z_k} \tag{158}$$

$$\implies \|\nabla_{z_k}\mathcal{L}\| \leq \frac{2}{C}\sum_{i=1}^{C}\left|(y_i - p_i)\frac{\partial p_i}{\partial z_k}\right| \leq \frac{2}{C}\sum_{i=1}^{C}|y_i - p_i| \cdot \left|\frac{\partial p_i}{\partial z_k}\right| \leq 2 \, . \tag{159}$$

By substituting these derivatives into the gradient equation, we can derive the gradient for each logit with respect to the MSE loss.

□

Lemma C.12. *Let $X^\ell \in \mathbb{R}^{d\times d}$ be the hidden representations of layer $\ell > 0$ as defined in equation 3. Then, we have that:*

$$\log\left\|\frac{\partial X^{\ell+1}}{\partial X^\ell}\right\| \leq \log\left\|J_{\text{BN}}(H^\ell)\right\| \, . \tag{160}$$

Proof of Lemma C.12. By definition, we have that:

$$H^\ell = W^\ell X^\ell, \tag{161}$$

$$X^{\ell+1} = \text{BN}(H^\ell) \, . \tag{162}$$

Therefore, applying the chain rule, we get:

$$\frac{\partial X^{\ell+1}}{\partial X^\ell} = \frac{\partial X^{\ell+1}}{\partial H^\ell}\frac{\partial H^\ell}{\partial X^\ell} = J_{\text{BN}}(H^\ell)W^\ell \, . \tag{163}$$

Taking the logarithm of the norm of this quantity, we reach the conclusion:

$$\log\left\|\frac{\partial X^{\ell+1}}{\partial X^\ell}\right\| \leq \log\left\|J_{\text{BN}}(H^\ell)\right\| + \log\left\|W^\ell\right\| = \log\left\|J_{\text{BN}}(H^\ell)\right\|, \tag{164}$$

where we have used the fact that the spectrum of the orthogonal matrix W^ℓ contains only the singular value 1 with multiplicity d. □

D LINEAR INDEPENDENCE IN COMMON DATASETS

In this section, we empirically verify the assumption that popular datasets do not suffer from rank collapse in most practical settings.

We provide empirical evidence for CIFAR10, MNIST, FashionMNIST and CIFAR100. We test this assumption by randomly sampling 100 input batches of sizes $n = 16, 32, 64, 128, 256, 512$ from each of these datasets and then measuring the rank of the Gram matrix of these randomly sampled batches using the `matrix_rank()` function provided in PyTorch. We stop at size 512 since we approach the dimensionality of some datasets, i.e. FashionMNIST, MNIST. We show in Table D1 the average rank with the standard deviation for each n, over 100 randomly sampled batches.

Table D1: Average rank of Gram matrix of input batches of size n from different datasets. Mean and standard deviation are computed over 100 randomly selected input batches, where the samples are chosen without replacement.

Dataset	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$
CIFAR10	16.0 ± 0.0	32.0 ± 0.0	64.0 ± 0.0	127.99 ± 0.09	221.06 ± 2.89	203.70 ± 3.58
MNIST	16.0 ± 0.0	32.0 ± 0.0	64.0 ± 0.0	128.00 ± 0.00	250.48 ± 1.53	318.04 ± 2.82
FashionMNIST	16.0 ± 0.0	32.0 ± 0.0	64.0 ± 0.0	128.00 ± 0.00	238.19 ± 3.27	275.37 ± 4.20
CIFAR100	16.0 ± 0.0	32.0 ± 0.0	64.0 ± 0.0	127.92 ± 0.27	218.11 ± 3.69	201.51 ± 3.95

We would like to remark that these datasets are fairly simple in terms of dimensionality and semantics, which can lead to correlated samples. Furthermore, the rank degeneracy can be alleviated even in the larger batch sizes through various data augmentations techniques. Note that these datasets have a high degree of correlation between samples. Most notably, the average cosine similarity between samples in a 512 size batch is $0.81, 0.40, 0.58, 0.81$ for CIFAR10, MNIST, FashionMNIST and CIFAR100 respectively.

E ACTIVATION SHAPING

In this section, we explain the full procedure for shaping the activation, as well as expand on the heuristic we use to choose the pre-activation gain. Under the functional structure of the MLP in equation 11, let α_ℓ be the pre-activation gain.

More formally, since the gradient norm has an exponential growth in depth, as shown in Figure G5, we can compute the linear growth rate of log-norm of gradients in depth. We define the rate of explosion for a model of depth L and gain α at layer ℓ as the slope of the log norm of the gradients:

$$R(\ell, \alpha_\ell) = \frac{\log \|\nabla_{W_\ell} \mathcal{L}\| - \log \|\nabla_{W_{\ell-10}} \mathcal{L}\|}{10} \,. \tag{165}$$

Since the rate function is not perfectly linear and has noisy peaks, we measure the slope with a 10 layer gap in order to capture the true behaviour instead of the noise.

Our goal is to choose α such that the sum of the rates across the layers in depth is bounded by a constant that does not depend on the depth of the model, i.e. $R(\ell, \alpha_\ell) \leq \beta$, where β is independent of L. One choice to achieve this is to pick a gain such that the sum of the rates behaves like a decaying harmonic sum in depth.

To this end, we measure the rate of explosion at multiple layers in a 1000 layer deep model, for various gains α which are constant across the layers in Figure E1 and notice that it behaves as $R(\ell) = c_1 \alpha_2^c$. In order to have the sum of rates across layers behave like a bounded harmonic series in depth, we must choose the gain such that it decays roughly as $\alpha^{c_2} = \ell^{-k}$ where $k > 1$ results in convergence. Therefore, we can obtain a heuristic for picking a gain such that the gradients remain bounded in depth as $\alpha_\ell = \ell^{-k/c_2}$, where we refer to k/c_2 as the gain exponent.

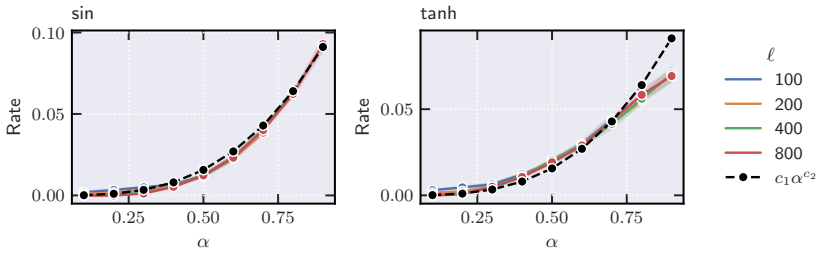

Figure E1: Explosion rate of the log norm of the gradients at initialization for an MLP model with orthogonal weights and batch normalization, for sin and tanh nonlinearities measured for a 1000 layer deep model at layers ℓ as a function of gain α. The black trace shows the fitted function $c_1 \alpha_2^c$. Traces are averaged over 10 independent runs, with the shades showing the 95% confidence interval.

This reduces the problem to picking the exponent such that the sum stays bounded. We show how the behaviour of the explosion rate at the early layers, for various models, is impacted by the exponent in Figure E2. Note that for several exponent values, we able to reduce the exponential explosion rate and obtain trainable models, which we show in Appendix G.

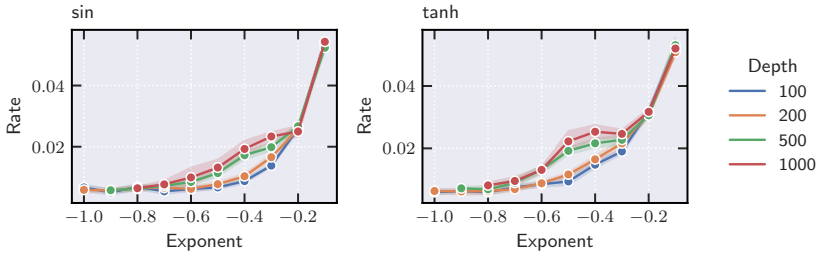

Figure E2: Explosion rate of the log norm of the gradients at initialization for an MLP model with orthogonal weights and batch normalization, for sin and tanh nonlinearities at depths $100, 200, 500, 1000$ as a function of the gain exponent. Traces are averaged over 10 independent runs, where the shade shows the 95% confidence interval. Rate is measured at $\ell = 10$ to avoid the any transient effects of the function

F IMPLICIT ORTHOGONALITY DURING TRAINING

In this section, we provide empirical evidence that our architecture during training maintains orthogonality across depths, while maintaining bounded gradients. Figure F1 shows the evolution of the isometry gap of the weight matrices W_ℓ during training, for models at different depths and different nonlinearities. In order to show that these weights are updated gradient descent, we also show the evolution of the norm of the loss gradients with regards to matrices W_ℓ in Figure F2.

These experiments are performed on an MLP with orthogonal weight matrices and batch normalization, with sin and tanh activations. The width is set to 100, batch size 100 and learning rate 0.001. The gain exponent is set to a fixed value for all experiments. The measurements are performed on a single batch of size 100 from CIFAR10, after each epoch of training on the same dataset.

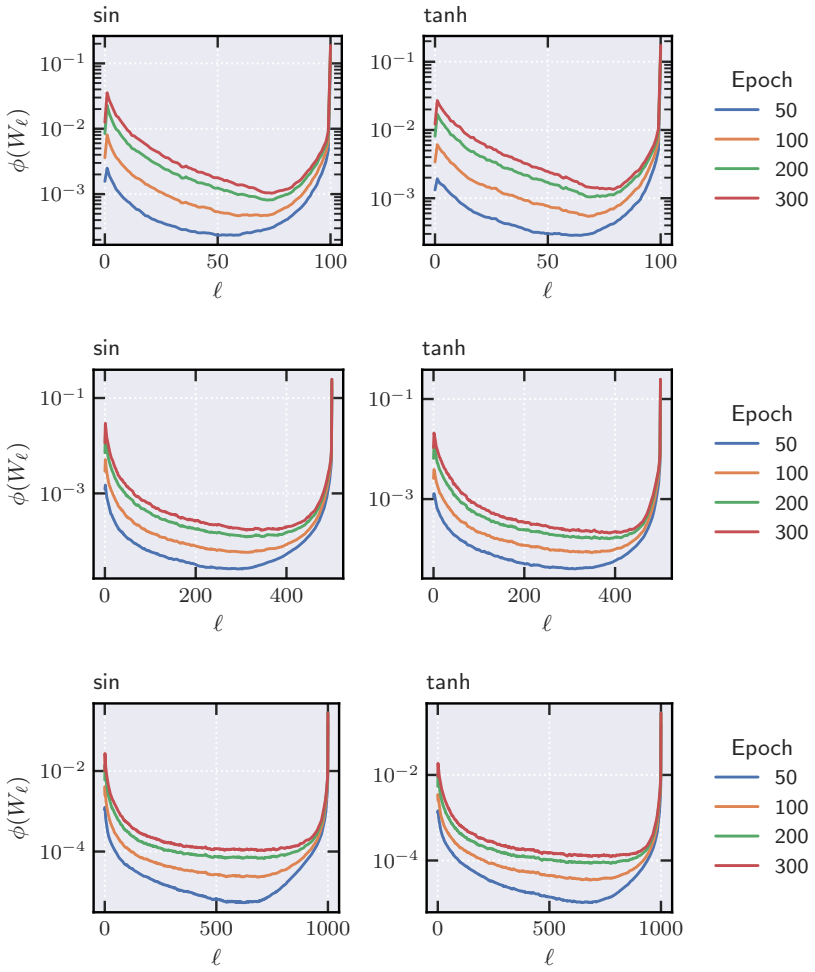

Figure F1: Contrasting the isometry gap of weight matrices during training for MLPs of depth 100 (top), 500 (middle), 1000 (bottom). The middle layers become increasingly more orthogonal with depth, while maintaining a small isometry gap. During training, the isometry gap also remains low, suggesting the matrices remain close to being orthogonal.

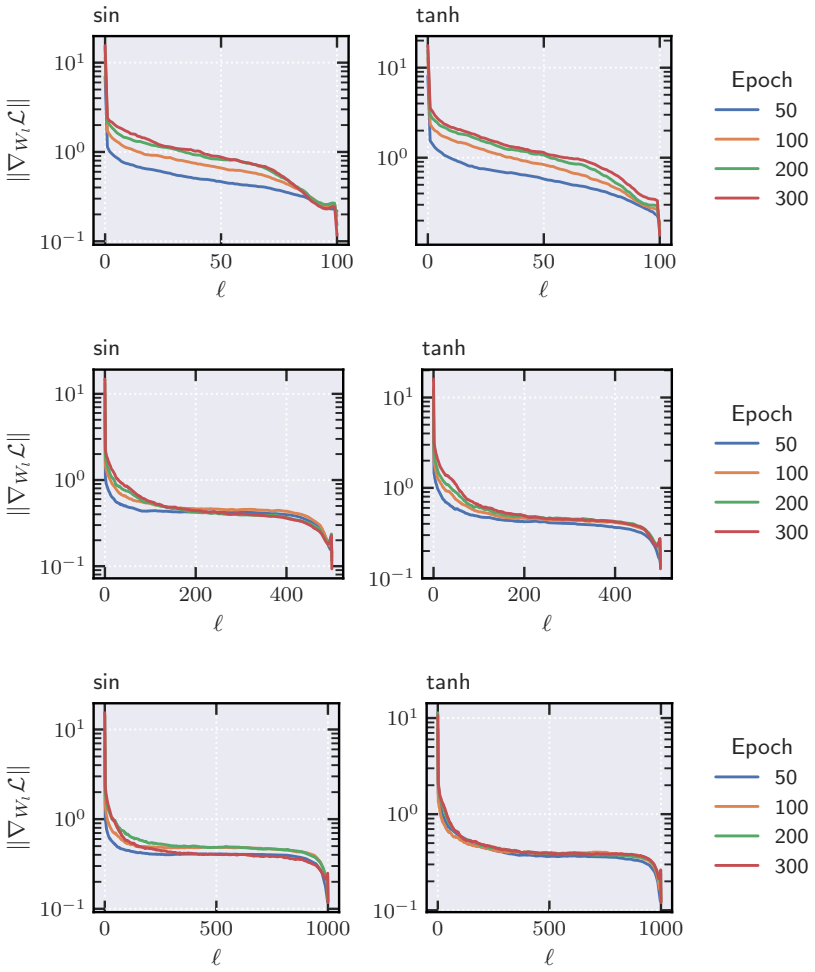

Figure F2: Contrasting the Frobenius norm of the gradients of the loss with respect to the weights during training for MLPs of depth 100 (top), 500 (middle), 1000 (bottom). The gradients do not vanish during training and across different depths for all layers, suggesting that the orthogonality evidenced in Figure F1 is not due to the weights not being updated during SGD.

G OTHER EXPERIMENTS

In this section we provide the train and test accuracies of deep MLPs on 4 popular image datasets, namely MNIST, FashionMNIST, CIFAR10, CIFAR100. Hyperparameters and measurements procedure are described in Section 4.

1 SUPPLEMENTARY TRAIN AND TEST RESULTS ON MNIST, FASHIONMNIST, CIFAR10, CIFAR100

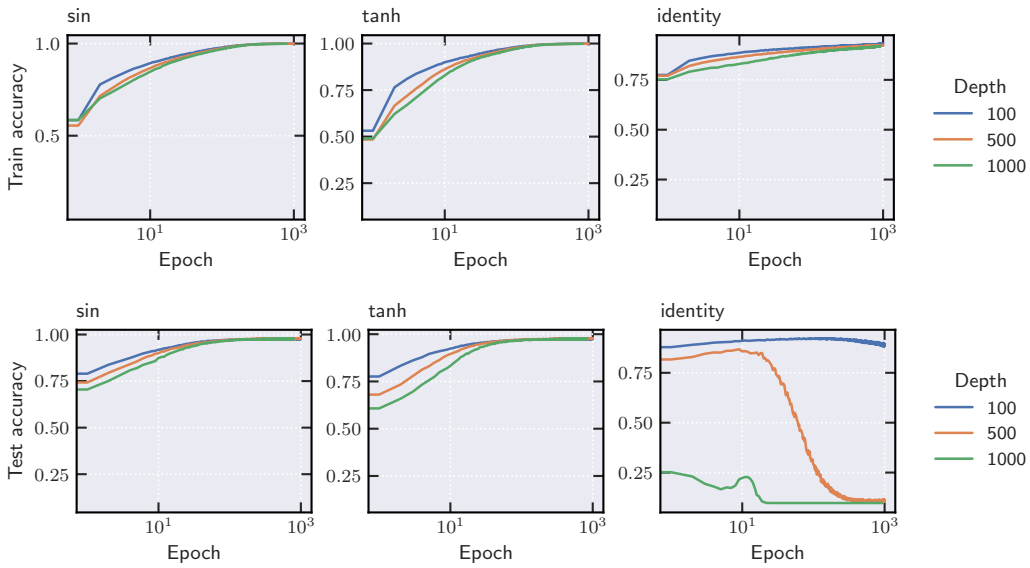

Figure G1: Contrasting the train and test accuracy of MLPs with gained sin, tanh and identity activations on MNIST. The identity activation performs much worse than the nonlinearities, indicating the fact that the sin and tanh networks are not operating in the linear regime. The network is trained with vanilla SGD and the hyperparameters are width 100, batch size 100, learning rate 0.001.

2 SUPPLEMENTAL FIGURES

We present empirical results in Figure 2 showing that degenerate input batches are a hard constraint for orthogonalization without gradient explosion. For MLPs with different depths, we show that by repeating samples in a batch of size 10 we get an exponential gradient explosion, which is unavoidable theoretically.

Furthermore, we show how non-linearities affect the gradient explosion rate in Figure G5. Using standard batch normalization and fully connected layers from PyTorch we show that non-linearities maintain a large isometry gap. This is a critical issue for our theoretical framework, since we take advantage of the fact that the identity activation achieves perfect orthogonality in order to prove that the gradients remain bounded in depth.

3 INFLUENCE OF MEAN REDUCTION ON THE GRADIENT BOUND

In this section, we compare whether adding mean reduction and the additional factor of $\frac{1}{n}$ in the denominator of the batch normalization module influences our gradient bound. As expected, we show in Figure G6 that in both cases, for the identity activation, the result remains similar, with the gradients remaining bounded in depth.

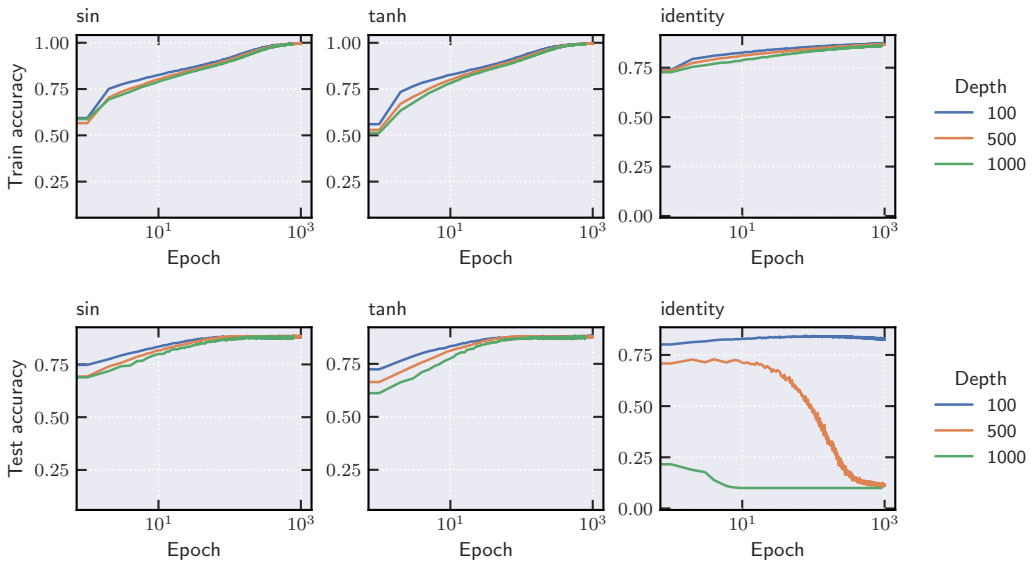

Figure G2: Contrasting the train and test accuracy of MLPs with gained sin, tanh and identity activations on FashionMNIST. The identity activation performs much worse than the nonlinearities, indicating the fact that the sin and tanh networks are not operating in the linear regime. The networks are trained with vanilla SGD and the hyperparameters are width 100, batch size 100, learning rate 0.001.

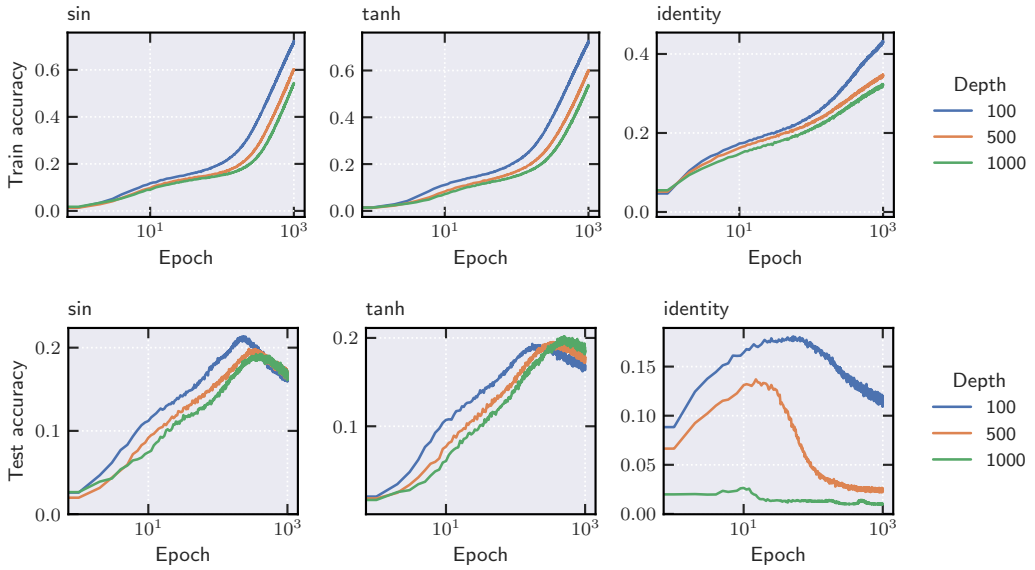

Figure G3: Contrasting the train and test accuracy of MLPs with gained sin, tanh and identity activations on CIFAR100. The identity activation performs much worse than the nonlinearities, indicating the fact that the sin and tanh networks are not operating in the linear regime. The networks are trained with vanilla SGD and the hyperparameters are width 100, batch size 100, learning rate 0.001.

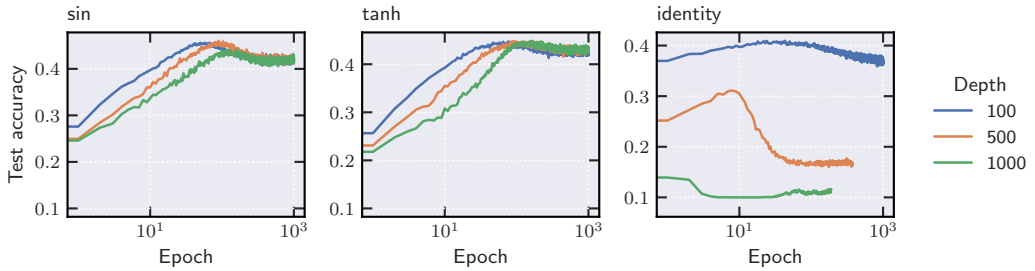

Figure G4: Contrasting the train test accuracy of MLPs with gained sin, tanh and identity activations on CIFAR10. The identity activation performs much worse than the nonlinearities, indicating the fact that the sin and tanh networks are not operating in the linear regime. The networks are trained with vanilla SGD and the hyperparameters are width 100, batch size 100, learning rate 0.001.

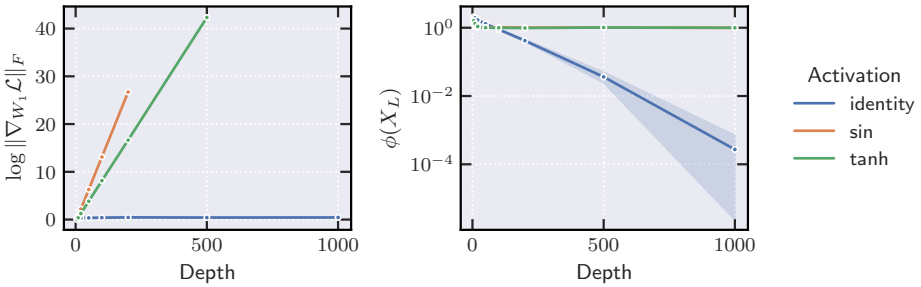

Figure G5: Left: average log-norm of gradients (log-scale y-axis) at the first layer for networks with different depths, evaluted on CIFAR10. Right: Isometry gap (log-scale y-axis) at the last layer for networks with different depths, evaluted on CIFAR10. The MLP is initialized with orthogonal weights and batch normalization, with standard modules, with $\sin, \tanh,$ identity non-linearities. After stabilizing the isometry gap, the non-linearities have an exponential gradient explosion.

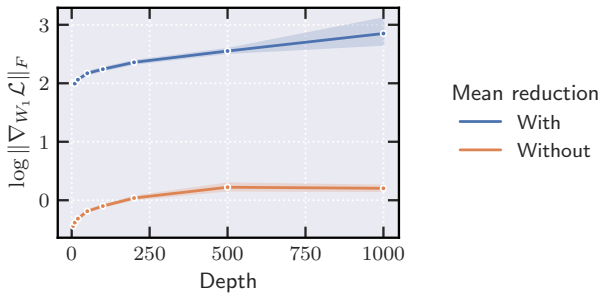

Figure G6: Comparing the gradient explosion rate in networks with standard batch normalization (blue) and networks with the simplified batch normalization operator from our theoretical framework (orange). Notice that the 2 traces are similar in terms of gradient explosion. Traces are averaged over 10 runs with the shaded regions showing the 95% confidence interval. Samples are from CIFAR10.

H HIGH PROBABILITY LOG-GRADIENT BOUND

Theorem H.1. *In the same conditions as Theorem C.1, it there is constant C such that:*

$$\Pr\left(\log \|\nabla_{W_\ell} \mathcal{L}\|_{op} \geq C \log(1/\delta) d^6 (\phi(X^0)^3 + 1)\right) \leq 2\delta$$

for any $\delta \in (0, 1)$.

For example for $\delta = 1/2$, we we have $\|\nabla_{W_\ell} \mathcal{L}\|_{op} \lesssim d^6(\phi(X^0)^3 + 1)$ with at most $1/2$ probability.

Proof. We follow the same proof as Theorem C.1 by decomposing the log gradient to sum of log gradients of each layer. Note that Lemma C.4 holds deterministically, as it is valid by construction. Thus, what remains is to derive a probabilistic version of Lemma C.3. We start from equation 100 which holds deterministically, to arrive at a probabilistic bound:

$$\sum_{k=1}^{L} \log \left\| J_{\text{BN}}(H^k) \right\| \leq S(d\phi(X^0)) + 2 \sum_{\ell=S+1}^{L} \sqrt{d\phi(X^\ell)}, \qquad S := \min \left\{ \ell : \phi(X^\ell) \leq \frac{1}{16d^2} \right\} \tag{166}$$

We derive a probabilistic bound for each term.

First term. Based on Lemma C.5 we have $E[S] \lesssim d^4 \phi(X^0)^2$. Thus, there is C_0 such that

$$\exists C_0 : E[S] \leq \frac{C_0}{2} d^4 \phi(X^0)^2.$$

Thus, by Markov inequality we have

$$P(S \geq B) \leq \frac{\mathbb{E}[S]}{B} \leq \frac{1}{2}, \qquad B := C_0 d^4 \phi(X^0)^2.$$

First, we discretize the layers into blocks of size B, where B is defined above. With a slight abuse of notation, define B_i as the end of the ith block of size B, and let E_i be the event $E_i = \{\phi(X_{B_i}) > \frac{1}{16d^2}\}$, which is the event of failure for ϕ to drop below the threshold of $\frac{1}{16d^2}$ after the last layer from block i.

By the inequality established above, we know that in each block, ϕ can either fall below the threshold, or stay above, with probability at most $1/2$. Moreover, knowing that ϕ is non-increasing, we know that $P(E_i | E_{<i}) \leq 1/2$.

Thus, by the non-increasing property of ϕ, the probability of failure after k blocks of size B is at most the probability that ϕ did not fall below the threshold in any of the k successive blocks:

$$\begin{aligned} P(E_k) &\leq P(E_1 \wedge \cdots \wedge E_{k-1}) && (167)\\ &= P(E_1) P(E_2|E_1) \dots P(E_k | E_{k-1} \dots E_1) && (168)\\ &\leq \left(\frac{1}{2}\right)^k && (169) \end{aligned}$$

Thus, we obtain the probability:

$$P(S \geq kB) = P(S \geq kC_0 d^4 \phi(X_0)^2) \leq 2^{-k} \tag{170}$$

Thus, connecting back to gradients, we obtain:

$$\Pr\Big(\sum_{\ell=0}^{S} \|J_{\text{BN}}(H^\ell)\|_{op} \geq kC_0 d^5 \phi(X^0)^3 \Big) \leq 2^{-k}. \tag{171}$$

Second term Starting from equation 109, we have

$$\ell \geq S \implies \Pr\Big\{ \phi(X^{\ell+1}) \geq (1 - 1/4d^2) \phi(X^\ell) \Big\} \leq 1 - \frac{1}{4d^2}$$

Let E_ℓ denote the event that $\phi(X^{\ell+1})$ does not decrease by $1 - 1/4d^2$ compared to its previous layer: $E_\ell = \mathbf{1}\Big\{ \phi(X^{\ell+1}) \geq (1 - 1/4d^2) \phi(X^\ell) \Big\}$. Due to the non-increasing property of ϕ, we know that $\Pr(E_\ell) = \Pr(E_\ell | E_{\ell-1})$. Repeating this for s steps, we obtain:

$$\ell \geq S \implies \Pr\Big\{ \phi(X^{\ell+s}) \geq (1 - 1/4d^2) \phi(X^\ell) \Big\} = \prod_{i=\ell}^{\ell+s-1} \Pr\{E_{i+1} | \bar{E}_i\} \leq (1 - \frac{1}{4d^2})^s \tag{172}$$

Inspired by this, consider the sequence $\ell_0, \ell_1, \dots$ defined as

$$\ell_0 = S, \qquad \ell_k = \ell_{k-1} + 4ckd^2$$

Thus, we get:

$$\Pr\Big\{\phi(X^{\ell_{k+1}}) \geq (1 - \frac{1}{4d^2})\phi(X^{\ell_k})\Big\} \leq (1 - \frac{1}{4d^2})^{4ckd^2} \leq e^{-ck}$$

Thus, we have the union bound

$$\Pr\Big\{\bigvee_{k=0}^{\infty} \phi(X^{\ell_{k+1}}) \geq (1 - \frac{1}{4d^2})\phi(X^{\ell_k})\Big\} \leq \sum_{k=1}^{\infty} e^{-ck} = \frac{e^{-c}}{1 - e^{-c}}$$

$$\implies \Pr\Big\{ \underbrace{\bigwedge_{k=0}^{\infty} \phi(X^{\ell_{k+1}}) \leq (1 - \frac{1}{4d^2})\phi(X^{\ell_k})}_{Q:=} \Big\} \geq 1 - \frac{e^{-c}}{1 - e^{-c}} = \frac{1 - 2e^{-c}}{1 - e^{-c}}$$

Note that in the event that Q holds, we can the isometry gaps in each $[\ell_{k-1}, \ell_k]$ interval as:

$$Q \implies \phi(X_\ell) \leq \frac{1}{16d}(1 - \frac{1}{4d^2})^{k-1} \text{ for all } \ell \in [\ell_{k-1}, \ell_k)$$

We can upper bound $\sum_{\ell=S+1}^{\infty} \sqrt{d\phi(X^\ell)}$ by using the numbers of items and upper bound on each block. Thus, assuming for all k we have $\phi(X^{\ell_{k+1}}) \leq (1 - \frac{1}{4d^2})\phi(X^{\ell_k})$, we can derive

$$\begin{aligned}
Q \implies \sum_{\ell=S+1}^{\infty} \sqrt{d\phi(X_\ell)} &\leq \sum_{k=1}^{\infty} (\ell_k - \ell_{k-1}) \sqrt{d \frac{1}{16d}(1 - \frac{1}{4d^2})^k} \\
&= cd^2 \sum_{k=1}^{\infty} k(1 - \frac{1}{4d^2})^{k/2} \\
&\leq cd^2 \sum_{k=1}^{\infty} k(1 - \frac{1}{8d^2})^k && \text{using } (1+x)^{1/2} \leq 1 + x/2 \\
&= cd^2 \frac{1 - 8/d^2}{(1/8d^2)^2} && \text{using } \sum_{k=1}^{\infty} k\alpha^k = \alpha/(1-\alpha)^2. \\
&\leq 64cd^6
\end{aligned}$$

Thus we have:

$$\Pr\left(\sum_{\ell=S+1}^{\infty} \sqrt{d\phi(X_\ell)} \leq 64cd^6\right) \geq \Pr\Big\{Q\Big\} \geq \frac{1 - 2e^{-c}}{1 - e^{-c}}$$

which yields

$$\Pr\left(\sum_{\ell=S+1}^{\infty} \log \|J_{BN}(H^\ell)\|_{op} \geq 64cd^6\right) \leq \Pr(\overline{Q}) = \frac{e^{-c}}{1 - e^{-c}}$$

Combining first and second term bounds for the Jacobian log-norms we have

$$\Pr\Big(\sum_{l=\ell}^{S} \log \|J_{\text{BN}}(H^l)\|_{op} > kC_0 d^5 \phi(X^0)^3\Big) \leq 2^{-k} \tag{173}$$

$$P\left(\sum_{\ell=S+1}^{\infty} \log \|J_{\text{BN}}(H^l)\|_{op} \geq 64cd^6\right) \leq \frac{e^{-c}}{1 - e^{-c}} \tag{174}$$

And thus we have

$$\Pr\Big(\sum_{l=\ell}^{L} \log \|J_{BN}(H^l)\|_{op} \geq kC_0 d^5 \phi(X^0)^3 + 64cd^6\Big) \leq 2^{-k} + \frac{e^{-c}}{1 - e^{-c}}$$

Thus, we can find C such that

$$\Pr\Big(\log \|\nabla_{W_\ell} \mathcal{L}\|_{op} \geq kCd^6(\phi(X^0)^3 + 1)\Big) \leq 2^{-k+1}$$

We can finish the proof by defining $\delta = 2^{-k}$ and change of variables $k = \log(1/\delta)$. □

I GRADIENTS IN RESIDUAL NETWORKS

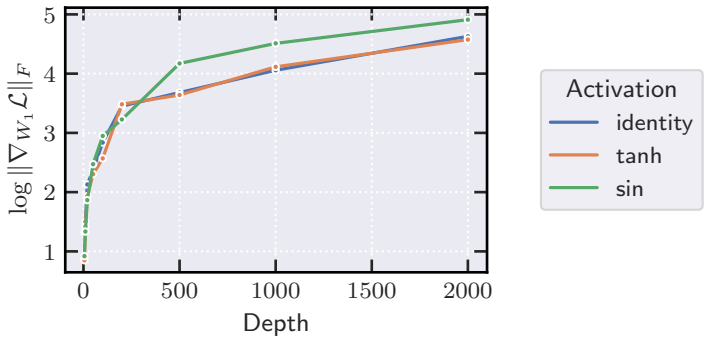

Figure I1: Logarithmic plot for the gradient norm of the first layer for residual networks with batch normalization initialized with Gaussian weights at different depths, evaluated on CIFAR10. The traces show that the gradients do not remain bounded in depth for different activations.