

Globally Injective ReLU Networks: Supplementary Material

APPENDIX A PROOFS FROM SECTION 2

A.1 PROOFS FROM SUBSECTION 2.2

Proof of Theorem 1. Suppose that W is such that the conditions of Theorem 1 hold, and that $\text{ReLU}(Wx_1) = \text{ReLU}(Wx_2) = y$. If for $j \in [[m]]$, $y|_j > 0$ then both $\langle w_j, x_1 \rangle > 0$ and $\langle w_j, x_2 \rangle > 0$. Similarly, if $y|_j \leq 0$ then both $\langle w_j, x_1 \rangle \leq 0$ and $\langle w_j, x_2 \rangle \leq 0$. In particular, this implies that

$$\langle w_j, x_1 \rangle > 0 \iff \langle w_j, x_2 \rangle > 0 \text{ and } \langle w_j, x_1 \rangle \leq 0 \iff \langle w_j, x_2 \rangle \leq 0. \quad (11)$$

If we then consider $x_\alpha := (1 - \alpha)x_1 + \alpha x_2$ where $\alpha \in (0, 1)$, then

$$\text{ReLU}(Wx_\alpha) = y = \text{ReLU}(Wx_1) = \text{ReLU}(Wx_2). \quad (12)$$

If $\langle w_j, x_\alpha \rangle > 0$ then at least one of $\langle w_j, x_1 \rangle > 0$ or $\langle w_j, x_2 \rangle > 0$. (11) implies that both must hold, therefore $\langle w_j, x_1 \rangle = \langle w_j, x_2 \rangle > 0$. If $\langle w_j, x_\alpha \rangle = 0$ then $\langle w_j, x_1 \rangle = \langle w_j, x_2 \rangle = 0$ (otherwise (11) is violated), thus

$$\text{ReLU}(W|_{S(x_\alpha, W)}x_1) = \text{ReLU}(W|_{S(x_\alpha, W)}x_2) \implies W|_{S(x_\alpha, W)}x_1 = W|_{S(x_\alpha, W)}x_2 \quad (13)$$

and so because $W|_{S(x_\alpha, W)}$ is full rank, this implies that $x_1 = x_2$. This proves one direction.

The other direction follows from the following. Suppose that there exists a x such that $W|_{S(x, W)}$ don't span \mathbb{R}^n . If $S(x, W) = \emptyset$ non-injectivity trivially follows, so suppose w.l.o.g. that $S(x, W) \neq \emptyset$. Let $x^\perp \in \ker(W|_{S(x, W)})$ and $\alpha \in \mathbb{R}^+$ such that $\alpha < \min_{j \in S^c(x, W)} \frac{-\langle x, w_j \rangle}{|\langle x^\perp, w_j \rangle|}$ ¹. Then for $j = 1, \dots, m$ one of the following two hold

$$\text{if } j \in S(x, W) \text{ then } \langle w_j, x + \alpha x^\perp \rangle = \langle w_j, x \rangle + \alpha \langle w_j, x^\perp \rangle = \langle w_j, x \rangle \quad (14)$$

$$\text{if } j \in S^c(x, W) \text{ then } \langle w_j, x + \alpha x^\perp \rangle = \langle w_j, x \rangle + \alpha \langle w_j, x^\perp \rangle < 0. \quad (15)$$

Thus, as ReLU acts pointwise (row-wise in W), we have that

$$\text{ReLU}(W(x + \alpha x^\perp)) = \text{ReLU}(Wx) \quad (16)$$

and, hence, $\text{ReLU}(W \cdot)$ is not injective. \square

Proof of Lemma 1. First, we show that if $\text{ReLU}(W|_{b \geq 0} \cdot)$ is injective, then so is $\text{ReLU}(W \cdot + b)$. Clearly if $\text{ReLU}(Wx_1 + b) = \text{ReLU}(Wx_2 + b)$ then $\text{ReLU}(W|_{b \geq 0}x_1 + b|_{b \geq 0}) = \text{ReLU}(W|_{b \geq 0}x_2 + b|_{b \geq 0})$ as well. If we apply Lemma 8 to each component of the above equation, then we obtain that $\text{ReLU}(W|_{b \geq 0}x_1) = \text{ReLU}(W|_{b \geq 0}x_2)$ which, given the injectivity of $\text{ReLU}(W|_{b \geq 0} \cdot)$, implies that $x_1 = x_2$.

Now suppose that $\text{ReLU}(W|_{b \geq 0})$ is not injective. Let $x \in \mathbb{R}^n$ be such that $W|_{b \geq 0}$ **doesn't have** a DSS of \mathbb{R}^n w.r.t. x . Let $\beta > 0$ be small enough so that $W|_{b < 0}(\beta x) + b|_{b < 0} < 0$ component-wise, and let $x^\perp \in \mathbb{R}^n$ such that (as in (16)) $\text{ReLU}(W|_{b \geq 0}(\beta x + x^\perp)) = \text{ReLU}(W|_{b \geq 0}\beta x)$. Further let $\alpha < 1$ be small enough such that $W|_{b < 0}(\beta x + \alpha x^\perp) + b|_{b < 0} < 0$. By a component-wise analysis, we have that

$$\text{ReLU}(W(\beta x + \alpha x^\perp) + b) = \text{ReLU}(W\beta x + b)|_{b \geq 0} = \text{ReLU}(W\beta x + b); \quad (17)$$

thus, $\text{ReLU}(W \cdot + b)$ is not injective. \square

Proof of Corollary 2. If $W \in \mathbb{R}^{m \times n}$ is injective then consider a plane p in \mathbb{R}^n that none of the rows of W lie in. Apply Theorem 1 to both normals of the plane. The corresponding DSS' for each normal are disjoint, thus there must be at least $2n \geq m$, so $m < 2 \cdot n$ implies non-injectivity.

Now we show that if W satisfies Theorem 1, then W is of the form given by (5). Suppose that there is a row vector w_i such that there are no row vectors pointing in the $-w_i$ direction. Let p be a plane

¹If $\langle x^\perp, w_j \rangle = 0$ for all $j \in S^c(x, W)$, then any $\alpha > 0$ will do.

through the origin such that $w_i \in p$, but $w_{i'} \notin p$ for $i \neq i'$. By Theorem 1, there must be at least n columns that lie on each (closed) half of p . Indeed one of the sides must have exactly n vectors on it (including w_i). By considering a small rotation of p , we can construct a plane with only $n - 1$ vectors on one side, hence there is no DSS for that rotated plane's normal. Thus for ReLU W to be injective, for every $i \in [[2n]]$ there must be a different $i' \in [[2n]]$ such that w_i and $w_{i'}$ are anti-parallel. This can only happen if for every $w \in W$ there is a corresponding $-dw \in W$, so W must have the form of (5). \square

A.2 PROOF OF THEOREM 2

A.2.1 UPPER BOUND ON MINIMAL EXPANSIVITY

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $w_i : \Omega \rightarrow \mathbb{R}^n$, $1 \leq i \leq m$ be m iid Gaussian vectors on Ω stacked in a matrix $W : \Omega \rightarrow \mathbb{R}^{m \times n}$.

We aim to find conditions for a ReLU layer with the matrix W to be injective. Injectivity fails at p the half-space $\{x : \langle x, p \rangle \geq 0\}$ contains fewer than n vectors w_i . Equivalently, injectivity fails if there is a half-space which contains more than $m - n$ vectors w_i (which implies that its opposite half-space has fewer than n).

Let $k = m - n + 1$ and denote by $A \in \mathbb{R}^{k \times n}$ some $k \times n$ submatrix of W (for example, the first k rows). Our strategy is to first bound the probability that for a fixed subset of k rows of W , there exists an x having positive inner products with all k rows (which signals non-injectivity per above discussion). Second, since there are $\binom{m}{k}$ subsets of k rows, we use the union bound to get an upper bound on the probability of non-injectivity.

For the first part we follow the proof of Bürgisser & Cucker (2013, Theorem 13.6), parts of which we reproduce for the reader's convenience. For a sign pattern $\sigma \in \{-1, 0, 1\}^k$, we denote by $R_A(\sigma)$ the set of all $x \in \mathbb{R}^n$ which produce the sign pattern σ (they belong to the σ -“wedge”, possibly empty),

$$R_A(\sigma) = \{x \in \mathbb{R}^n : \text{sign}(\langle x, a_i \rangle) = \sigma_i, i \in \{1, \dots, k\}\}. \quad (18)$$

For $\sigma \in \{-1, +1\}^k = \Sigma$, we define the event \mathcal{E}_σ as

$$\mathcal{E}_\sigma = \{\omega : R_{A(\omega)}(\sigma) \neq \emptyset\}. \quad (19)$$

We are interested in $\sigma_0 = (1, \dots, 1)$, meaning that all the inner products are positive. Note that the probability of \mathcal{E}_σ is the same for all σ due to the symmetry of the Gaussian measure. Further, note that $\sum_{\sigma \in \Sigma} \mathbf{1}_{\mathcal{E}_\sigma}(\omega) = |\{\sigma : R_{A(\omega)}(\sigma) \neq \emptyset\}|$ is the number of wedges defined by A . Then

$$\mathbb{P}(\mathcal{E}_{\sigma_0}) = \frac{1}{2^k} \sum_{\sigma \in \Sigma} \mathbb{P}(\mathcal{E}_\sigma) = \frac{1}{2^k} \sum_{\sigma \in \Sigma} \mathbf{E}(\mathbf{1}_{\mathcal{E}_\sigma}) = \frac{1}{2^k} \mathbf{E} \left(\sum_{\sigma \in \Sigma} \mathbf{1}_{\mathcal{E}_\sigma} \right) = \frac{1}{2^{k-1}} \sum_{i=0}^{n-1} \binom{k-1}{i}, \quad (20)$$

by Winder's bound (Winder, 1966). (The hyperplane arrangement is generic almost surely so we use equality.)

We now have the probability that for a subset of k vectors w_i , there exists an $x \in \mathbb{R}^n$ which has positive inner products with all k vectors. We are interested in the following event which implies non-injectivity,

$$\mathcal{E}_{\text{NI}} = \{\omega : W(\omega) \text{ has a subset of } k \text{ rows } B(\omega) \text{ such that } R_{B(\omega)}(\sigma_0) \neq \emptyset\}. \quad (21)$$

Conversely, $\omega \in \mathcal{E}_{\text{NI}}^c$ implies almost sure injectivity.

Since there are $\binom{m}{k} = \binom{m}{m-n+1} = \binom{m}{n-1}$ different subsets of k rows, we can bound the probability of \mathcal{E}_{NI} as

$$\mathbb{P}(\mathcal{E}_{\text{NI}}) \leq \binom{m}{n-1} \mathbb{P}(\mathcal{E}_{\sigma_0}) \leq \left(\frac{me}{n-1} \right)^n 2^{-(m-n)} 2^{(m-n)H(\frac{n-1}{m-n})} \quad (22)$$

$$\lesssim (ce)^n 2^{n(c-1)[H((c-1)^{-1})-1]} \quad (23)$$

$$= 2^{-n[-\log_2(ce) - (c-1)(H((c-1)^{-1})-1)]}, \quad (24)$$

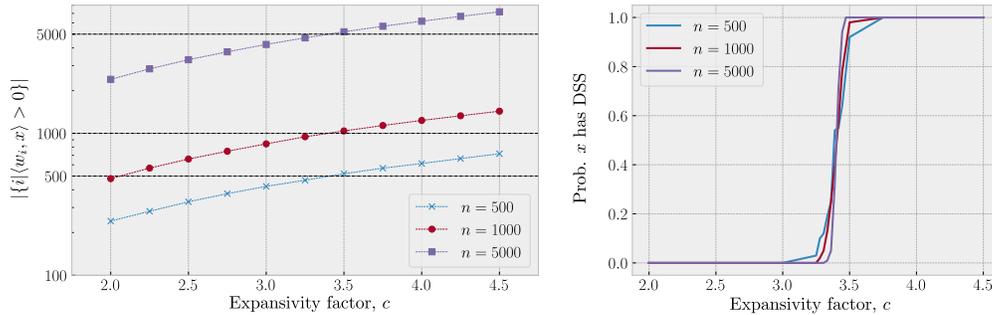


Figure 6: We empirically show that for random Gaussian matrices $\mathbf{W} \in \mathbb{R}^{m \times n}$ $m = cn$, a critical oversampling factor of at least $c > 3.5$ is required for injectivity of ReLU. We choose $x = -\frac{1}{m} \sum_j w_j / \|w_j\|$. *Left*: a plot of the number of elements of W that point in the direction of x as a function of expansivity for several different choices of n . If this quantity is less than n , then W cannot contain a DSS of \mathbb{R}^n w.r.t. x . *Right*: a plot of the empirical probability that a Gaussian matrix W contains a DSS of \mathbb{R}^n w.r.t. x as a function of expansivity for several different values of n .

where $H(\epsilon) = -\epsilon \log_2(\epsilon) - (1 - \epsilon) \log_2(1 - \epsilon)$ is the binary entropy function and we used a related bound on the sums of binomial coefficients.² To get the bound note that the bracket in the exponent is positive for $c \geq 10.43$.

A.2.2 LOWER BOUND ON MINIMAL EXPANSIVITY OF LAYERS WITH GAUSSIAN WEIGHTS

In this appendix we prove that large random Gaussian weight matrices $W \in \mathbb{R}^{m \times n}$ yield non-injective ReLU layers with high probability if $m < c^*n$, where $c^* \approx 3.4$. Note that injectivity fails if there exists a half-space in \mathbb{R}^n which contains less than n rows of W . Equivalently, injectivity fails if there is a half-space with more than $m - n$ rows of W , since then the opposite halfspace has less than n . The core idea of the proof is to make an educated guess for a half-space which has many vectors, and compute the probability that it has more than $m - n$. A good such guess is to take the halfspace defined by the average direction of a row of W . Equivalently, we study of the size of $S(\bar{w}, W)$ where $\bar{w} = \frac{1}{m} \sum_{k=1}^m w_k$, the row average of a matrix W . If $|S(\bar{w}, W)| > m - n$, then $|S(-\bar{w}, W)| < n$, and W cannot have a DSS w.r.t. $-\bar{w}$. As the Figure 6 shows, when $m < c^*n$, W does not contain a DSS w.r.t. $-\bar{w}$ with high probability.

Let $W \in \mathbb{R}^{m \times n}$ be a matrix such that the rows of W are i.i.d. random vectors distributed as $\mathcal{N}(0, I_n)$. Define the event E_i as

$$E_i := \left\{ \omega : \left\langle w_i(\omega), \sum_{k=1}^m w_k(\omega) \right\rangle \geq 0 \right\}. \quad (25)$$

Lemma A. *Let $m, n \rightarrow \infty$ so that $\frac{m}{n} \rightarrow c$. Then $\mathbb{P}(E_i) \rightarrow \frac{1}{2} \operatorname{erfc}\left(-\frac{1}{\sqrt{2c}}\right)$.*

Proof. We have

$$\mathbb{P}(E_i) = \mathbb{P}\left(\|w_i\|^2 + \left\langle w_i, \sum_{i \neq j} w_j \right\rangle \geq 0\right) = \mathbb{P}\left(\|w_i\|^2 + \|w_i\| Y \geq 0\right)$$

where $Y \sim \mathcal{N}(0, m - 1)$. We now use the fact that $\|w_i\|^2$ concentrates around n . Define the event

$$D_i = \left\{ \omega : \|w_i(\omega)\|^2 \in [(1 - \epsilon)n, n/(1 - \epsilon)] \right\}. \quad (26)$$

²https://en.wikipedia.org/wiki/Binomial_coefficient#Sums_of_binomial_coefficients

One can show using standard concentration arguments that $\mathbb{P}(D_i) \geq 1 - 2 \exp(-\epsilon^2 n/4)$, and so by the law of total probability,

$$\mathbb{P} \left\{ \|w_i\|^2 + \|w_i\| Y \geq 0 \right\} = \mathbb{P} \{ \|w_i\| + Y \geq 0 \mid D_i \} \mathbb{P} \{ D_i \} + \mathbb{P} \{ \|w_i\| + Y \geq 0 \mid D_i^c \} \mathbb{P} \{ D_i^c \}. \quad (27)$$

Choosing $\epsilon = \epsilon(n) = n^{-1/4}$ yields

$$\mathbb{P} \left\{ \sqrt{n - n^{3/4}} + Y \geq 0 \right\} \leq \mathbb{P} \{ \|w_i\| + Y \geq 0 \mid D_i \} \leq \mathbb{P} \left\{ \sqrt{n/(1 - n^{-1/4})} + Y \geq 0 \right\}.$$

Finally, substituting $Z = \frac{Y}{\sqrt{m-1}} \sim \mathcal{N}(0, 1)$ yields

$$\mathbb{P} \left\{ Z \geq -\sqrt{\frac{n - n^{3/4}}{m-1}} \right\} \leq \mathbb{P} \{ \|w_i\| + Y \geq 0 \mid D_i \} \leq \mathbb{P} \left\{ Z \geq -\sqrt{\frac{n/(1 - n^{-1/4})}{m-1}} \right\}$$

Both sandwiching probabilities converge to $\mathbb{P} \left\{ Z \geq -\frac{1}{\sqrt{c}} \right\} = \frac{1}{2} \operatorname{erfc} \left(-\frac{1}{\sqrt{2c}} \right)$. Noting that $\mathbb{P} \{ D_i \} \rightarrow 1$ and $\mathbb{P} \{ D_i^c \} \rightarrow 0$ we finally have from (27) that

$$\mathbb{P} \{ E_i \} \rightarrow \frac{1}{2} \operatorname{erfc} \left(-\frac{1}{\sqrt{2c}} \right).$$

□

Lemma B. Under the same conditions as Lemma A when $i \neq j$,

$$\mathbb{P} \{ E_i \cap E_j \} \rightarrow \frac{1}{4} \operatorname{erfc} \left(-\frac{1}{\sqrt{2c}} \right)^2.$$

Proof. The proof is similar to that of Lemma A. In addition to concentration of norm of iid Gaussian vectors, we use the fact that $\langle w_i, w_j \rangle / m$ is of order $n^{-1/2}$ and that if $\mathbb{P} \{ D \} \rightarrow 1$ as $n \rightarrow \infty$, then for a fixed c and some event A , $\mathbb{P} \{ A \mid D \} - \mathbb{P} \{ A \} \rightarrow 0$. Here D is the event that the various quantities are close to the value they concentrate about, and A asymptotically has the same probability as $E_i \cap E_j$. □

Theorem 6. Given a Gaussian weight matrix $W \in \mathbb{R}^{m \times n}$, the layer $\operatorname{ReLU}(Wx)$ with $m/n \rightarrow c$ is not injective with probability $\rightarrow 1$ as $n \rightarrow \infty$ when $c < c^*$, where c^* is the unique positive real solution to

$$\frac{1}{2} \operatorname{erfc} \left(\frac{1}{\sqrt{2c}} \right) = \frac{1}{c}.$$

The numerical value of c^* is ≈ 3.4 .

Proof. Let $X_i = \mathbf{1}_{\langle w_i, \frac{1}{m} \sum_{j=1}^m w_j \rangle \geq 0}$, the indicator function of the event E_i . The expected number of w_i with a positive inner product with $\sum_{k=1}^m w_k$ is by the linearity of expectation equal to

$$\mathbb{E} \left(\sum_{i=1}^m X_i \right) = m \cdot \mathbb{P} \{ E_i \} =: mp.$$

By Chebyshev's inequality

$$\mathbb{P} \left(\left| \frac{1}{m} \sum_{i=1}^m X_i - p \right| \geq t \right) \leq \frac{\sigma^2}{t^2}$$

where $\sigma^2 = \mathbb{V} \left(\frac{1}{m} \sum_{i=1}^m X_i \right)$ is the variance of the sum. We compute

$$m^2 \sigma^2 = \mathbb{V} \left(\sum_{i=1}^m X_i \right) = \mathbb{E} \left(\sum_{i=1}^m X_i \right)^2 - \left(\mathbb{E} \sum_{i=1}^m X_i \right)^2 = mp + \sum_{i \neq j} \mathbb{P}(E_i \cap E_j) - (mp)^2.$$

Since $\mathbb{P}(E_i \cap E_j) \rightarrow p^2$ by Lemmas A and B, we have for any i and $j \neq i$

$$\sigma^2 = \frac{mp + m(m-1)\mathbb{P}(E_i \cap E_j) - (mp)^2}{m^2} \rightarrow 0.$$

Thus indeed for any $t > 0$

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m X_i - p\right| \geq t\right) \rightarrow 0.$$

Finally,

$$\mathbb{P}(\text{noninjectivity}) \geq \mathbb{P}\left(\sum_{i=1}^m X_i > m - n\right) \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m X_i > 1 - \frac{1}{c}\right). \quad (28)$$

Combining with Lemma A we get that

$$\frac{1}{m}\sum_{i=1}^m X_i \xrightarrow{\mathbb{P}} \frac{1}{2}\text{erfc}\left(-\frac{1}{\sqrt{2c}}\right)$$

Thus

$$\mathbb{P}(\text{noninjectivity}) \geq \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m X_i > 1 - \frac{1}{c}\right) \rightarrow \begin{cases} 0 & c > c^* \\ 1 & c < c^* \end{cases}. \quad (29)$$

□

A.3 EXPERIMENTS TESTING THE MINIMAL EXPANSIVITY THRESHOLD

Figure 6 suggests that an expansivity factor of 2.1 as suggested by Lei et al. (2019) is not enough to ensure global injectivity with Gaussian weights. This in turn leads to failure cases when inverting the network in applications to inverse problems, as we show in this section. We use a single ReLU layer, $f(x) = \text{ReLU}(Wx)$, and choose x to be an MNIST digit for ease of visualization of the “latent” space. We choose $W \in \mathbb{R}^{cn \times n}$ with iid Gaussian entries and set $c = m/n = 2.1$. Given $y = f(x)$, our goal is to reconstruct x using, for example, gradient descent. This requires at least n of the cn entries of $f(x)$ to be non-zero since $x \in \mathbb{R}^n$. As shown in Figure 6, $x_0 = -\frac{1}{m}\sum_j w_j / \|w_j\|$ violates this condition with high probability when $c = 2.1$.

We choose 2 data samples, x_1 and x_2 to be 2 “latent” codes. Note that W has a DSS with respect to x_1 . This ensures that given $f(x_1)$ one can reconstruct x_1 . Now, in order to show that injectivity around a few points is insufficient for global injectivity, we linearly interpolate between x_0 and x_1 and x_0 and x_2 in a 2D grid. We show the intermediate x s as images in Figure 7. The x s shown in red cannot be uniquely recovered given $f(x)$, i.e., there exist infinitely many samples close to the red samples that all map to the same output.

To further illustrate this, we also run a simple experiment where we have $y = f(x) + \eta$, with a 10dB signal-to-noise ratio. We invert f to estimate x for 3 different samples in the domain of f . We can easily see that non-injective points of the domain give poorer reconstructions due to the lower number of positive inner products with W . In order to avoid problems of non-convexity in optimization we report the best reconstruction out of 10 in 4 different trials with different random seeds, similar to experiments of Bora et al. (2017). While these results show reconstructions with single layer, the issues are only exacerbated with multiple layers.

A.4 PROOF OF THEOREM 3

We divide the proof into parts for ease of understanding.

Lemma 2 (Inverse Lipschitz Constant: Face Adjacent Wedges). *Let $W \in \mathbb{R}^{m \times n}$ have a DSS w.r.t. every $x \in \mathbb{R}^n$ and $x_0, x_1 \in \mathbb{R}^n$. Define*

$$\forall t \in [0, 1], \quad \ell^{x_0, x_1}(t) = (1-t)x_0 + tx_1 \quad (30)$$

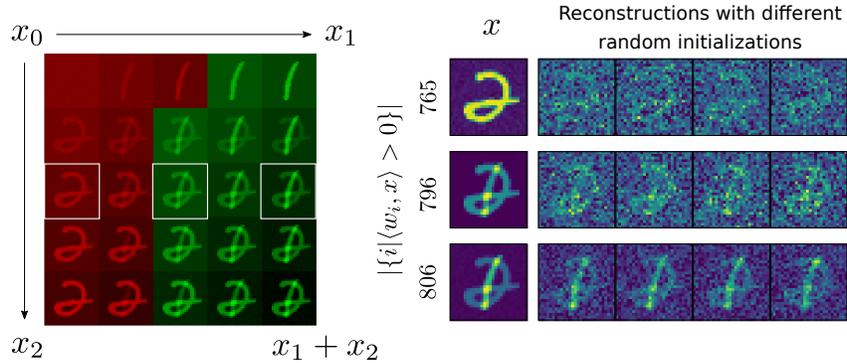


Figure 7: The mapping $\phi(x) := \text{ReLU}(Wx)$, $x \in \mathbb{R}^n$ when $W \in \mathbb{R}^{2.1n \times n}$ as suggested in [Lei et al. \(2019\)](#) is not injective as the samples shown in red do not have n positive inner products with the rows of W . This implies that around the red samples ϕ cannot be inverted. For the 3 samples delineated by white boxes on the left, we do a simple denoising test where $y = \phi(x) + \eta$ and η corresponds to 10dB noise. Each reconstruction reported is best out of 10 trials based on MSE error.

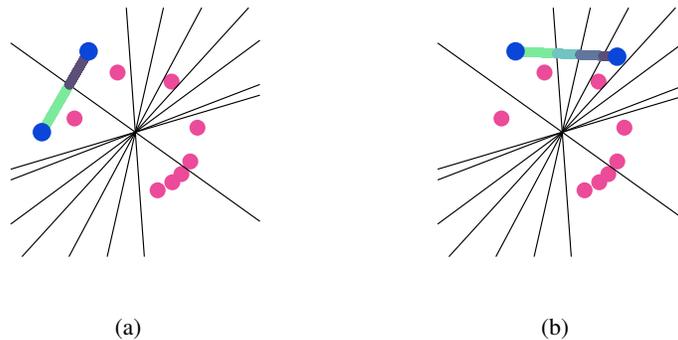


Figure 8: A Proof aid of the DSS condition when $W \in \mathbb{R}^{8 \times 2}$. The blue dots are x_0 and x_1 , the pink are the weight matrix rows, black lines denote the boundaries between adjacent wedges, and the multi-colored line is $\ell^{x_0, x_1}(t)$. This line changes color each time it crosses into a new wedge.

and suppose that x_0 and x_1 are such that there is a $t' \in (0, 1)$ such that for all $t \in [0, 1]$

$$W|_{S(\ell^{x_0, x_1}(t), W)} = \begin{cases} W|_{S(x_0, W)} & \text{if } t < t' \\ W|_{S(x_1, W)} & \text{if } t' < t \end{cases}. \quad (31)$$

Suppose further that x_0, x_1 are such that there is some $\delta > 0$ such that for all $\delta x \in \mathbb{R}^n$, if $\|\delta x\|_2 < \delta$ then there is a $t' + \delta t$ such that

$$W|_{S(\ell^{x_0, x_1 + \delta x}(t), W) \geq 0} = \begin{cases} W|_{S(x_0, W)} & \text{if } t < t' + \delta t \\ W|_{S(x_1, W)} & \text{if } t' + \delta t < t \end{cases}. \quad (32)$$

Then

$$\|\text{ReLU}(Wx_0) - \text{ReLU}(Wx_1)\|_2 \geq \frac{1}{\sqrt{2}} \min(\sigma(W|_{S(x_0, W)}), \sigma(W|_{S(x_1, W)})) \|x_0 - x_1\|_2, \quad (33)$$

where $\sigma(M)$ is the smallest singular value of the matrix M .

Remark 3. The conditions of Lemma 2 on x_0 and x_1 may look very odd, but they have a very natural geometric meaning. The DSS can be thought of as slicing \mathbb{R}^n into wedges by a series of hyperplanes that have the rows of W as normals.

The condition in (31) is interpreted as that the line segment that connects x_0 to x_1 passes from x_0 's wedge into x_1 's wedge without passing through any wedges in between (see Figure 8a, as opposed to Figure 8b). This implies that x_0 and x_1 must be in wedges that share a boundary. The condition in (32) requires that the wedges of x_0 and x_1 share a face, and not just a corner.

Proof. We denote

$$W|_{S(x_0, W)} = W_0, W|_{S(x_1, W)} = W_1, W_{0 \cap 1} = W_0 \cap W_1, W_0 = \begin{bmatrix} W_{0 \cap 1} \\ W_{0 \setminus 1} \end{bmatrix}, W_1 = \begin{bmatrix} W_{0 \cap 1} \\ W_{1 \setminus 0} \end{bmatrix}. \quad (34)$$

First we will show that if $w_i, w_{i'} \in W_{0 \setminus 1}$, then w_i and $w_{i'}$ must be parallel. From equation (31) and the continuity of $\text{ReLU}(W\ell^{x_0, x_1}(t))$ w.r.t. t , we have that there is a $t' \in (0, 1)$ such that

$$\langle w_i, (1 - t')x_0 + t'x_1 \rangle = 0 = \langle w_{i'}, (1 - t')x_0 + t'x_1 \rangle. \quad (35)$$

If w_i and $w_{i'}$ are not parallel, then let δx be some vector $0 < \|\delta x\| < \delta$ that is perpendicular to w_i but not $w_{i'}$ then

$$\langle w_i, (1 - t')x_0 - t'\delta x + t'x_1 \rangle = -t \langle w_i, \delta x \rangle = 0, \quad (36)$$

$$\langle w_{i'}, (1 - t')x_0 - t'\delta x + t'x_1 \rangle = -t \langle w_{i'}, \delta x \rangle \neq 0, \quad (37)$$

which contradicts (32). W.l.o.g. the same argument applies to $W_{1 \setminus 0}$ and also it is straight forward to see that all elements of $W_{0 \setminus 1}$ must be anti-parallel to all elements of $W_{1 \setminus 0}$. From $W_{1 \setminus 0}$ and $W_{0 \setminus 1}$ parallelism, there is a $c \geq 0$ such that for all $x \in \mathbb{R}^n$,

$$\|W_{1 \setminus 0}x\| = c^2 \|W_{0 \setminus 1}x\|. \quad (38)$$

Assume that $c \geq 1$, then

$$\|W_0x_0 - W_1x_1\|_2^2 = \|W_{0 \cap 1}x_0 - W_{0 \cap 1}x_1\|_2^2 + \|W_{0 \setminus 1}x_0\|_2^2 + \|W_{1 \setminus 0}x_1\|_2^2 \quad (39)$$

$$= \|W_{0 \cap 1}x_0 - W_{0 \cap 1}x_1\|_2^2 + \|W_{0 \setminus 1}x_0\|_2^2 + c^2 \|W_{0 \setminus 1}x_1\|_2^2 \quad (40)$$

$$\geq \|W_{0 \cap 1}x_0 - W_{0 \cap 1}x_1\|_2^2 + \|W_{0 \setminus 1}x_0\|_2^2 + \|W_{0 \setminus 1}x_1\|_2^2 \quad (41)$$

$$\geq \|W_{0 \cap 1}x_0 - W_{0 \cap 1}x_1\|_2^2 + \frac{1}{2} \|W_{0 \setminus 1}(x_0 - x_1)\|_2^2 \quad (42)$$

$$\geq \frac{1}{2} \|W_0x_0 - W_0x_1\|_2^2 \quad (43)$$

$$\geq \frac{\sigma(W_0)^2}{2} \|x_0 - x_1\|_2^2. \quad (44)$$

The antepenultimate inequality comes as from the definition of $W_{0 \setminus 1}$, we have that $W_{0 \setminus 1}x_0$ and $-W_{0 \setminus 1}x_1$ are the same sign, thus (98) applies. In the case that $c < 1$, then the rolls of $W_{0 \setminus 1}$ and

$W_{1 \setminus 0}$ can be switched, and the same result (with $\sigma(W_1)$ in place of $\sigma(W_0)$) is obtained. In either case,

$$\|W_0 x_0 - W_1 x_1\|_2^2 \geq \frac{1}{2} \min \left(\sigma(W|_{S(x_0, W)})^2, \sigma(W|_{S(x_1, W)})^2 \right) \|x_0 - x_1\|_2^2. \quad (45)$$

□

Lemma 3 (Inverse Lipschitz Constant: Connected through Faces). *Let $W \in \mathbb{R}^{m \times n}$ have a DSS w.r.t. every $x \in \mathbb{R}^n$. Let x_0, x_1 be such that the line connecting passes through n_t wedges, and through their faces (in the sense of Lemma 3). Then*

$$\|\text{ReLU}(W x_0) - \text{ReLU}(W x_1)\|_2 \geq \frac{1}{\sqrt{2n_t}} \min_{t \in [0, 1]} \sigma(W|_{S(\ell^{x_0, x_1}(t), W)}) \|x_0 - x_1\|_2. \quad (46)$$

Proof. Let $t_1 = 0$, $t_{n_t} = 1$, and let $T = \{t_k\}_{k=1}^{n_t}$ such that $t_k < t_{k+1}$ for $k < n_t$ and $x_k := \ell^{x_0, x_1}(t_k)$ are each in different wedges. Let $c = \frac{1}{\sqrt{2}} \min_{t \in [0, 1]} \sigma(W|_{S(\ell^{x_0, x_1}(t), W)})$, then

$$c \|x_0 - x_1\|_2 \leq \sum_{k=1}^{n_t-1} c \|x_{t_k} - x_{t_{k+1}}\|_2 \quad (47)$$

then by Lemma 2,

$$\sum_{k=1}^{n_t-1} c \|x_{t_k} - x_{t_{k+1}}\|_2 \leq \sum_{k=1}^{n_t-1} \|\text{ReLU}(W x_{t_k}) - \text{ReLU}(W x_{t_{k+1}})\|_2 \quad (48)$$

and by Lemma 10 we have

$$\sum_{k=1}^{n_t-1} \|\text{ReLU}(W x_{t_k}) - \text{ReLU}(W x_{t_{k+1}})\|_2 \leq \sqrt{n_t} \|\text{ReLU}(W x_0) - \text{ReLU}(W x_1)\|_2. \quad (49)$$

Combining (47) - (49) yields

$$\|\text{ReLU}(W x_0) - \text{ReLU}(W x_1)\|_2 \geq \frac{1}{\sqrt{2n_t}} \min_{t \in [0, 1]} \sigma(W|_{S(\ell^{x_0, x_1}(t), W)}) \|x_0 - x_1\|_2. \quad (50)$$

□

Proof of Theorem 3. Let x_0 and x_1 be given. If the line connecting x_0 and x_1 does not pass through any wedge corners, then we can apply Lemma 3 directly, and get that

$$\|\text{ReLU}(W x_0) - \text{ReLU}(W x_1)\|_2 \geq \frac{1}{\sqrt{2n_t}} \min_{t \in [0, 1]} \sigma(W|_{S(\ell^{x_0, x_1}(t), W)}) \|x_0 - x_1\|_2. \quad (51)$$

We now argue that the number of wedges that $\ell^{x_0, x_1}(t)$ passes through (and so n_t) is at most m . For each $j \in [m]$,

$$\text{ReLU}(W \ell^{x_0, x_1}(t))|_j = \langle w_j, \ell^{x_0, x_1}(t) \rangle \quad (52)$$

is monotone increasing or decreasing. This implies that each $w_j \in W$ can enter or exit the DSS of W w.r.t. $\ell^{x_0, x_1}(t)$ at most once, therefore the total number of unique DSS w.r.t. $\ell^{x_0, x_1}(t)$ (i.e. wedges $\ell^{x_0, x_1}(t)$ pass through) is at most m . Hence, $n_t \leq m$. Clearly

$$\min_{t \in [0, 1]} \sigma(W|_{S(\ell^{x_0, x_1}(t), W)}) \geq \min_{x \in \mathbb{R}^n} \sigma(W|_{S(x, W)}), \quad (53)$$

hence we have

$$\|\text{ReLU}(W x_0) - \text{ReLU}(W x_1)\|_2 \geq \frac{C}{\sqrt{m}} \|x_0 - x_1\|_2. \quad (54)$$

Now we show that if Lemma 3 does not apply to two points (namely, the line $\ell^{x_0, x_1}(t)$ passes through a corner in the sense of Remark 3), then the two points can be perturbed an arbitrarily small amount,

so that the perturbed points do satisfy Lemma 3. The corners (again, as in Remark 3) describe the points which are orthogonal to at least two $w_{j_1}, w_{j_2} \in W$. w_{j_1} and w_{j_2} must not be parallel to each other (otherwise Lemma 2 would apply), thus the set of points orthogonal to both w_{j_1} and w_{j_2} constitute a $n - 2$ dimensional linear space in \mathbb{R}^n .

Let x_0 and x_1 be such that $\ell^{x_0, x_1}(t)$ intersects one of these corners. By considering $\tilde{x}_0 = \delta x + x_0, \tilde{x}_1 = \delta x + x_1$ where δx is perpendicular to $x_1 - x_0$, we can obtain a line $\ell^{\tilde{x}_0, \tilde{x}_1}(t)$ so that $\ell^{\tilde{x}_0, \tilde{x}_1}(t)$ and $\ell^{x_0, x_1}(t)$ do not intersect. The choice of δx is $n - 1$ dimensional, thus for every $\delta > 0$ and w_{j_1} and w_{j_2} (that are non-perpendicular) there is a δx so that $\|\delta x\|_2 < \delta$ and $\ell^{\tilde{x}_0, \tilde{x}_1}(t)$ does not intersect the corner of w_{j_1} and w_{j_2} .

Consider a sequence of $\tilde{x}_0^{(i)}, \tilde{x}_1^{(i)}, i = 1, \dots$ such that $\lim_{i \rightarrow \infty} (\tilde{x}_0^{(i)}, \tilde{x}_1^{(i)}) = (x_0, x_1)$ and $\ell^{\tilde{x}_0^{(i)}, \tilde{x}_1^{(i)}}(t)$ does not pass through a corner for any i . Given that $\|\cdot\|_2$ and $\text{ReLU}(W \cdot)$ are continuous and so by Lemma 3

$$\left\| \text{ReLU}(W \tilde{x}_0^{(i)}) - \text{ReLU}(W \tilde{x}_1^{(i)}) \right\|_2 - \frac{C}{\sqrt{2m}} \left\| \tilde{x}_0^{(i)} - \tilde{x}_1^{(i)} \right\|_2 \geq 0, \quad (55)$$

thus

$$\left\| \text{ReLU}(W x_0) - \text{ReLU}(W x_1) \right\|_2 - \frac{C}{\sqrt{2m}} \|x_0 - x_1\|_2, \quad (56)$$

$$= \lim_{i \rightarrow \infty} \left\| \text{ReLU}(W \tilde{x}_0^{(i)}) - \text{ReLU}(W \tilde{x}_1^{(i)}) \right\|_2 - \frac{C}{\sqrt{2m}} \left\| \tilde{x}_0^{(i)} - \tilde{x}_1^{(i)} \right\|_2 \geq 0. \quad (57)$$

and so

$$\left\| \text{ReLU}(W x_0) - \text{ReLU}(W x_1) \right\|_2 \geq \frac{C}{\sqrt{2m}} \|x_0 - x_1\|_2. \quad (58)$$

□

Remark 4 (Factor of $\frac{1}{\sqrt{m}}$ in Theorem 3.). Throughout this paper we define the discrete norm of $y \in \mathbb{R}^d$ as

$$\|y\|_2 = \left(\sum_{j=1}^d [y]_j^2 \right)^{\frac{1}{2}}. \quad (59)$$

This is to be contrasted with the norm that arise from the discretization of the L_2 function norm on a finite domain. For example, if we instead thought of y as a discrete sampling of a continuous function $\tilde{y} \in L_2([0, 1])$, such that $\forall j = 1, \dots, d$

$$\tilde{y} \left(\frac{j-1}{m} \right) = [y]_j, \quad (60)$$

then we could approximate the $L_2([0, 1])$ norm of \tilde{y} by

$$\|\tilde{y}\|_{L_2([0,1])} \approx \|y\|_{l_2([0,1])} := \frac{1}{\sqrt{m}} \left(\sum_{j=1}^d [y]_j^2 \right)^{\frac{1}{2}}. \quad (61)$$

If we express Theorem 3 in terms of $\|\cdot\|_{l_2([0,1])}$, then it would become

$$\left\| \text{ReLU}(W x_0) - \text{ReLU}(W x_1) \right\|_{l_2([0,1])} \geq \frac{C(W)}{m} \|x_0 - x_1\|_2. \quad (62)$$

A.5 THEOREM 4

Example 1 (Applying Theorem 4, One Channel). Consider a layer of the form $\text{Reshape}(\text{ReLU}(W x))$ where $W = [C_1^T, \dots, C_q^T]^T$ and each C_k is a convolution operator with kernel c_k . Suppose further that $W \in \mathbb{R}^{4 \times 4 \times 1024 \times 100} = \mathbb{R}^{16384 \times 100}$ (as in Radford et al. (2015)). The reshaping operator takes the 16384 single-channel output of W and transforms it into a multi-channel signal. This is necessary

for subsequent convolutions, but plays no role in injectivity. Let $q = 8$, and the 2×2 convolution kernels be given as

$$c_1 = \begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}, \quad c_2 = \begin{bmatrix} -1 & 3 \\ -1 & -1 \end{bmatrix}, \quad c_3 = \begin{bmatrix} -1 & -1 \\ 3 & -1 \end{bmatrix}, \quad c_4 = \begin{bmatrix} -1 & -1 \\ -1 & 3 \end{bmatrix} \quad (63)$$

$$c_5 = -c_1, c_6 = -c_2, c_7 = -c_3, c_8 = -c_4. \quad (64)$$

Directly proving that a 16384×100 dimension operator has a DSS w.r.t. every $x \in \mathbb{R}^{100}$ is daunting. However, since each layer of the operator is given by one of only 8 simple convolutions, we can leverage Theorem 4 to significantly simplify the problem. Choosing $P = (2, 2)$, implies that $\mathcal{Z}_{(2,2)}(c_k) = c_k$, and so $W|_{\mathcal{Z}_{(2,2)}} = \bigcup_{k=1}^8 c_k$. Further, it is easy to see, that $\{c_1, c_2, c_3, c_4\}$ is a basis for $\mathbb{R}^{2 \times 2}$, so Corollary 2 applies, $W|_{\mathcal{Z}_P}$ has a DSS for \mathbb{R}^4 , and by Theorem 4, $\text{ReLU}(Wx)$ is injective.

An example when a layer is injective but P in Theorem 4 must be greater than O is when W is a convolution of 4 kernels of width 3

$$c_1 = [1 \ 0 \ -1], \quad c_2 = [1 \ 0 \ 1], \quad c_3 = [-1 \ 0 \ 1], \quad c_4 = [-1 \ 0 \ -1]. \quad (65)$$

If we choose $P = (3)$, then $W|_{\mathcal{Z}_{(3)}} = \bigcup_{k=1}^4 \{c_k\}$. only has four elements, and so cannot have a DSS w.r.t. every $x \in \mathbb{R}^3$ by Corollary 2. If we however choose $P = (4)$, then $\mathcal{Z}_{(4)}(c_k) = \{[c_k \ 0], [0 \ c_k]\}$ and $W|_{\mathcal{Z}_{(4)}}$ has a DSS of \mathbb{R}^4 w.r.t. all $x \in \mathbb{R}^4$ (from Corollary 2), so W has a DSS w.r.t. all $x \in \mathbb{R}^N$.

This example suggests that to apply Theorem 4 to a convolution layer with q kernels of width O , we must choose P so that $W|_{\mathcal{Z}_P}$ has at least the minimal number $2 \prod_{j=1}^p P_j$ of vectors to have a DSS of a vector space of dimension $|P| = \prod_{j=1}^p P_j$. Some algebra gives that $q \geq 2 \prod_{j=1}^p \frac{1}{1 - O_j/P_j}$ and $P_j \geq \frac{O_j}{1 - (2/q)^{1/p}}$, where the last inequality holds only when $\frac{P_j}{O_j}$ is independent of j .

Before we can commence with the proof of Theorem 4, we prove the following results.

Lemma 4 (Domain Decomposition). *Suppose that $\mathbb{R}^n = \text{span}\{\Omega_1, \dots, \Omega_K\}$ ³ where each Ω_k is a subspace and for each $k = 1, \dots, K$ we have a*

$$W_k = [w_{k,1}^T, \dots, w_{k,N_k}^T]^T \quad \text{and} \quad W = [W_1^T, \dots, W_K^T] \quad (66)$$

such that $w_{k,\ell} \in \Omega_k$ and W_k has a DSS of Ω_k w.r.t. every $x \in \Omega_k$. Then W has a DSS of \mathbb{R}^n w.r.t. every $x \in \mathbb{R}^n$.

Proof. For every $k = 1, \dots, K$ $W_k|_{S(x, W_k)}$ has a DSS of Ω_k with respect to $P_{\Omega_k}(x)$ where P_{Ω_k} is the orthogonal projection of \mathbb{R}^n onto Ω_k . For every $w_{k,\ell}$,

$$\langle w_{k,\ell}, P_{\Omega_k}(x) \rangle = \langle w_{k,\ell}, x \rangle \quad (67)$$

thus $S(x, W_k) \subset S(x, W)$. From $\Omega_k \subset \text{span}(W_k|_{S(x, W_k)})$ for each k we have a set spanning Ω_k that lie in $W|_{S(x, W)}$, hence

$$\mathbb{R}^n = \text{span}(\Omega_1, \dots, \Omega_K) \subset \text{span}\left(\bigcup_{k=1}^K \bigcup_{\ell=1}^{N_k} \{w_{k,\ell}\}\right) = \text{span}(W) \quad (68)$$

contains a DSS of \mathbb{R}^n w.r.t. x . The set $\bigcup_{k=1}^K \bigcup_{\ell=1}^{N_k} \{w_{k,\ell}\}$ has no dependence on x , thus it is true for all $x \in \mathbb{R}^n$. \square

Lemma 5. *Given a convolution operator $C \in \mathbb{R}^{N \times N}$. Let $0 \leq V$ be such that $V + O \leq N$. For each $x \in C$,*

$$\text{aug}_{V:V+O}(x) \in C, \quad \text{where} \quad (\text{aug}_{V:V+O}(x))_J = \begin{cases} (x)_{J-V} & \text{if } 1 + V \leq J \leq V + O \\ 0 & \text{otherwise} \end{cases}. \quad (69)$$

Note that $(\text{aug}_{V:V+O}(x))_J$ restricted to the indices $1 + V$ through $V + P$ is exactly x .

³This is a slight abuse of traditional spanning notation. Here we mean that there is a set of vectors of $\Omega_1, \Omega_2, \dots$ such that their union spans \mathbb{R}^n .

Definition 4 (Zero-Padded Kernel). Let c be a convolution kernel of width O , and let P be another multi-index. We define the set of zero-padded kernels⁴ of c as the set

$$\mathcal{Z}_P(c) = \left\{ \hat{c} \in \mathbb{R}^P : \exists Q, 0 \leq Q \leq P - O \quad \hat{c}_T = \begin{cases} c_{Q-T+O+1} & \text{if } 1 \leq T - Q - \leq O \\ 0 & \text{otherwise} \end{cases} \right\} \quad (70)$$

Note that the above notation implies that $\mathcal{Z}_P(c) = \{\}$ if $O \not\leq P$.

The above is always well defined, as $c_{Q-T+O+1}$ is well defined iff $1 \leq Q - T + O + 1 \leq O$ (recall that the kernel c is of width O), and so

$$1 \leq Q - T + O + 1 \leq O \iff -O \leq Q - T \leq -1 \iff 1 \leq T - Q \leq O \quad (71)$$

Now we proceed with the proof of Theorem 4.

Proof of Theorem 4. The strategy for this proof will be to use Lemma 4 by decomposing \mathbb{R}^n into some number of different domains $\{\Omega_v\}_{v=1}^{n_v}$, each of which are a restriction of \mathbb{R}^n to P non-zero components. For each Ω_v , the elements of W that lie in Ω_v can be identified as the elements in $W|_{\mathcal{Z}_P}$. If for one v the components of $W|_{\mathcal{Z}_P}$ form a DSS of Ω_v w.r.t. every $x \in \Omega_v$, then it **has** a DSS for every such Ω_k , so we can apply Lemma 4 and get that W has a DSS of \mathbb{R}^n w.r.t. all $x \in \mathbb{R}^n$.

Given an offset $V' \geq 0$ such that $V' + P \leq N$, define $\Omega_{V'}$ as the subspace of all vectors $x \in \mathbb{R}^N$ such that

$$\Omega_{V'} = \{x \in \mathbb{R}^N : x_J = 0 \text{ if } 1 + V' \not\leq J \text{ or } J \not\leq V' + P\}. \quad (72)$$

From Lemma 5, for any $k = 1, \dots, n_v$, if $x^{k,P} \in \mathcal{Z}_P(c_k)$, C_k is a submatrix of W and

$$\text{aug}_{V':V'+P}(x^{k,P}) \in C_k. \quad (73)$$

Further, for any such $x^{k,P}$,

$$\text{aug}_{V':V'+P}(x^{k,P}) \in \Omega_{V'}. \quad (74)$$

If $W|_{\mathcal{Z}_P}$ contains a DSS for \mathbb{R}^P w.r.t. all $x \in \mathbb{R}^P$, then

$$\text{aug}_{V':V'+P}(W|_{\mathcal{Z}_P}) \text{ contains a DSS for } \Omega_{V'} \text{ w.r.t. all } x \in \Omega_{V'}. \quad (75)$$

This follows from Lemma 4. From Lemma 5 for any V' such that

$$\text{aug}_{V':V'+P}(W|_{\mathcal{Z}_P}) \in W, \quad (76)$$

if $W|_{\mathcal{Z}_P}$ **has** a DSS of \mathbb{R}^P w.r.t. all $x \in \mathbb{R}^P$, then W contains a DSS of $\Omega_{V'}$ for all $0 \leq V' \leq N - P$. Finally, note that $\text{span}(\{\Omega_{V'}\}_{V'=0}^{N-P})$, and so using (75), (76) we can apply Lemma 4 and find that W contains a DSS of \mathbb{R}^N w.r.t. all $x \in \mathbb{R}^N$. \square

For a multi-channel input (with n_c channels) $x \in \underbrace{\mathbb{R}^N \times \mathbb{R}^N \times \dots \times \mathbb{R}^N}_{n_c \text{ times}}$, a multi-channel con-

volution C on x is given by $Cx = \sum_{q=1}^{n_c} C_q x_q$ where C_o is a convolution on \mathbb{R}^N (defined by Definition 2) and $x_o \in \mathbb{R}^N$ is the restriction of x to the o 'th channel. Because of the additive structure of multi-channel convolutions a n_c over $\mathbb{R}^N = \mathbb{R}^{N_1} \times \dots \times \mathbb{R}^{N_p}$ dimensional domain of width $O = (O_1, \dots, O_p)$ with kernels c_1, \dots, c_{n_c} is equivalent to a single convolution of width $(O, n_c) = (O_1, \dots, O_p, n_c)$ over $\mathbb{R}^{(N,p)} = \mathbb{R}^{N_1} \times \dots \times \mathbb{R}^{N_p} \times \mathbb{R}^{n_c}$. This follows from

$$(Cx)_J = \sum_{q=1}^{n_c} (C_q x_q)_J = \sum_{q=1}^{n_c} \sum_{I=1}^O (c_q)_{O-I-1} (x_q)_{J+I} = \sum_{(I,n_c)=1}^{(O,q)} c_{(O,n_c)-(I,q)-1} x_{(J,q)+(I,q)}.$$

⁴With many convolutional neural networks, padding refers to the act of padding the image (with e.g. zeroes), but the convolutional kernels are not padded. For our results, the variable P refers to the padding of the kernels, not the image.

APPENDIX B ROBUSTNESS TO BATCH, WEIGHT, AND SPECTRAL NORMALIZATION

Normalization strategies such as batch (Ioffe & Szegedy, 2015), layer (Ba et al., 2016), instance (Ulyanov et al., 2016), group (Wu & He, 2018), weight (Salimans & Kingma, 2016) and spectral (Miyato et al., 2018) normalization promote convergence during training and encourage low generalization error. Normalization is a many-to-one operation. In this section we show that batch, weight, and spectral normalization do not interfere with injectivity provided, of course, that the network is injective without normalization. We ask if injectivity is compatible with normalization in a *trained* network.

Let $\{x_i\}_{i=1,\dots,m}$ represent the inputs to a given layer over a mini-batch. The batch normalization adds two learnable parameters (γ, β) and transforms x_i to y_i as

$$\mu_{\mathcal{B}} = \frac{1}{m} \sum_{i=1}^m x_i, \quad \sigma_{\mathcal{B}}^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2, \quad \hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}, \quad y_i = \gamma \hat{x}_i + \beta. \quad (77)$$

Although for a given γ, β the relationship between x_i and y_i is not injective, batch normalization is usually present during training but not at test time. Provided that the learned weights satisfy Theorem 1 and Lemma 1, batch normalization does not spoil injectivity.

In batch renormalization (Ioffe, 2017) it is still desirable to whiten the input into each layer. During run time there may be no mini-batch, so running averages of the σ_i and μ_i (denoted $\hat{\sigma}, \hat{\mu}$) computed during training are used. Batch renormalization at test time is then $\hat{x} = \frac{x - \hat{\mu}}{\sqrt{\hat{\sigma}^2 + \epsilon}}, y = \gamma \hat{x} + \beta$. Since, importantly, $\hat{\sigma}$ and $\hat{\mu}$ are not functions of x , the mapping between x and y is one-to-one (when $\gamma \neq 0$), thus such normalization in an injective network does not spoil the injectivity.

In weight normalization (Salimans & Kingma, 2016) the coefficients of the weight matrices are normalized to have a given magnitude. This normalization is not a function of the input or output signals, but rather of the weight matrices themselves. This plays no role in injectivity.

B.1 LAYER, INSTANCE AND GROUP NORMALIZATION

layer, instance and group normalization, unlike batch normalization, take place during both training and execution and, unlike weight and spectral normalization, the normalization is done on the input/outputs of layers instead of on the weight matrices. For these normalizations (2) is modified so that it becomes

$$N(z) = \phi_L(W_L M_L(\dots \phi_2(W_2 M_2(\phi_1(W_1 z + b_1)) + b_2) \dots + b_L)) \quad (78)$$

where $M_\ell: \mathbb{R}^{n_{i+1}} \rightarrow \mathbb{R}^{n_{i+1}}$ are normalization functions that are many-to-one. In general $\phi_\ell(W_\ell M_\ell(\cdot) + b_\ell)$ will not be injective for any $\phi_\ell, W_\ell, b_\ell$ on account of M_ℓ , but for all of the mentioned normalization techniques we can get near injectivity. Before we descend into the particular we make the following observation about normalization methods that obey a certain structure.

Definition 5 (Scalar-Augmented Injective Normalization). Let $M_\ell(x): \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a normalization function that is understood to be many-to-one. We say that $M_\ell(x)$ is scalar-augmented injective if there exists a function $m_\ell(x): \mathbb{R}^n \rightarrow \mathbb{R}^k$ where $k \ll n$ and $\tilde{M}_\ell: \mathbb{R}^n \times \mathbb{R}^k \rightarrow \mathbb{R}^n$ such that

$$M_\ell(x) := \tilde{M}_\ell(x; m_\ell(x)) \quad (79)$$

and $\tilde{M}_\ell(x; m_\ell(x))$ is injective on x given $m_\ell(x)$.

An example of a normalization function that is scalar-augmented injective is

$$M_\ell(x) = \frac{x}{\|x\|_2}. \quad (80)$$

For this choice of M_ℓ , $k = 1$, and

$$\tilde{M}_\ell(x; c) = \frac{x}{c} \quad m_\ell(x) = \|x\|_2. \quad (81)$$

With this definition, we can prove the following trivial but useful result

Lemma 6 (Restricted Injectivity of Scalar-Augmented Normalized Networks). *Let N be a deep network of the form in (78) and let each $\phi_\ell(W_\ell \cdot)$ be layer-wise injective. Let the normalization functions $\{M_\ell\}_{\ell=1,\dots,L}$ each be scalar-augmented injective. Then given $\{m_\ell(x)\}_{\ell=1,\dots,L}$, the network*

$$\tilde{N}(z; m_1, \dots, m_\ell) = \phi_L(W_L \tilde{M}_L(\dots \phi_2(W_2 \tilde{M}_2(\phi_1(W_1 z + b_1); m_2) + b_2) \dots + b_L; m_L)) \quad (82)$$

is injective.

Proof. The proof of Lemma 6 follows from a straightforward application of induction, combined with Definition 5. \square

Remark 5. Note that Lemma 6 implies that for a fixed $\{m_\ell(x)\}_{\ell=1,\dots,L}$, there is at most one value of z such that

$$\tilde{N}(z; m_1, \dots, m_\ell) = N(z), \quad (83)$$

where $N(z)$ is given by (78), and that the z 's on both sides of (83) are the same. It is still entirely possible that there are is another choice of z' , $\{m_\ell(x)\}_{\ell=1,\dots,L}$ such that

$$\tilde{N}(z; m_1, \dots, m_\ell) = \tilde{N}(z'; m'_1, \dots, m'_\ell). \quad (84)$$

An example of this would be if M_ℓ is of the form in (80), then

$$\tilde{M}_\ell(x; m_\ell(x)) = \tilde{M}_\ell(2x; 2m_\ell(x)). \quad (85)$$

In other words, Lemma 6 implies that the deep network is injective (in z) for a fixed $\{m_\ell(x)\}_{\ell=1,\dots,L}$, but it may still not be injective for all z and $\{m_\ell(x)\}_{\ell=1,\dots,L}$.

With Lemma 6 in tow, we can show that layer, instance, and group normalization are all scalar-augmented injective normalizations, so Lemma 6 applies and yields a kind of injectivity. Layer, instance and group normalization are all related insofar as they can all be expressed in the same abstract form. For a given input x , all three break x up into K parts denoted $\{x|_{S_k}\}_{k=1,\dots,K}$ such that for each $k = 1, \dots, K$

$$\mu_k = \frac{1}{m} \sum_{i=1}^m (x|_{S_k})_i \quad \sigma_k^2 = \frac{1}{m} \sum_{i=1}^m ((x|_{S_k})_i - \mu_k)^2 \quad (86)$$

$$(\hat{x}|_{S_k})_i = \frac{(x|_{S_k})_i - \mu_k}{\sqrt{\sigma_k^2 + \epsilon}} \quad M(x)|_{S_k} = \gamma_k (\hat{x}|_{S_k})_i + \beta_k. \quad (87)$$

The differences between the three normalization are how $\{S_k\}_{k=1,\dots,K}$ is chosen. For layer normalization $K = 1$ and the normalization is applied to the entire input signal. For instance normalization, there is one S_k for each channel, and the $x|_{S_k}$ restricts x to just one channel of inputs, that is the normalization is done channel-wise. Group normalization is part way between these two, where k is less than the number of channels, and channels are batched together.

In any case, for any of these normalization methods, they are all scalar-augmented injective normalization where

$$M_\ell(x) = \tilde{M}_\ell(x; \{\sigma_{k,\ell}, \mu_{k,\ell}\}_{k=1,\dots,K}). \quad (88)$$

Thus, by Lemma 6 their corresponding deep networks are all injective, provided that for each ℓ , $\{\sigma_{k,\ell}, \mu_{k,\ell}\}_{k=1,\dots,K}$ is saved.

B.2 POOLING OPERATIONS

Although pooling may have a different aim than typical normalization, we consider it in this section, as it is mathematically similar to (78). Pooling is similar to layer, instance and group normalization in the sense that they partition the input space into K disjoint pieces, and then output a weighted average upon each piece. Specifically, if $M_p(x): \mathbb{R}^n \rightarrow \mathbb{R}^K$ where for $k = 1, \dots, K$,

$$M_p(x)|_{S_k} = \|x|_{S_k}\|_p \quad (89)$$

where $\|\cdot\|_p$ is the discrete p norm of x restricted to the set S_k . For $p = 1$ this is the mean of the absolute value, for $p = 2$ this is the Euclidean mean and for $p = \infty$ it is the maximum of the absolute value. The injectivity of this operation in the cases where $p = 1, 2, \infty$ is considered in the work [Bruna et al. \(2013\)](#).

APPENDIX C PROOFS FROM SECTION 3

C.1 PROOF OF THEOREM 5

To prove Theorem 5 we combine the approximation results for neural networks and the low regularity version of the generic orthogonal projector technique used to prove the easy version of the Whitney’s embedding theorem (Hirsch, 2012, Chapter 2, Theorem 3.5) that shows that a C^2 -smooth manifold of dimension n can be embedded in \mathbb{R}^{2n+1} with an injective, C^2 -smooth map. To prove the result, we first approximate $f(x)$ by a ReLU-type neural network that is only Lipschitz-smooth, so the graph of the map F_θ is only a Lipschitz-manifold. We note that limited regularity often causes significant difficulties for embedding results, as for example for the Lipschitz-smooth manifolds it is presently known only that a n -dimensional manifold can be embedded (without preserving distances) in the Euclidean space \mathbb{R}^N of dimension $N = (n + 1)^2$, and the classical Whitney problem, whether a n -dimensional manifold can be embedded in \mathbb{R}^{2n+1} , is still open Luukkainen & Väisälä (1977); Cobzaş et al. (2019). Due to this lack of smoothness, we recall the details how this generic projector technique works.

Proof of Theorem 5. Let $\varepsilon > 0$ and $\mathcal{Z} \subset \mathbb{R}^n$ be a compact set. As a continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ can be uniformly approximated in a compact set by a C^∞ -smooth function (see e.g. (Adams, 1975, Thm. 2.29)), we can without loss of generality assume that f smooth and therefore a locally Lipschitz function.

By classical results of approximation theory for shallow neural networks, see Hornik (1991); Leshno et al. (1993); Pinkus (1999), for any $L \geq 1$ there are m and a neural network $F_\theta \in \mathcal{NN}(n, m, L, m)$ such that

$$|f(x) - F_\theta(x)| \leq \frac{1}{2}\varepsilon, \quad \text{for all } x \in \mathcal{Z}. \quad (90)$$

We note that by using recent results for deep neural networks, e.g. by Yarotsky (2017), one can obtain efficient estimates on how a given accuracy ε can be obtained using sufficiently large L and m . Our aim is the perturb $F_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^m$ so that it becomes injective.

We assume that $\mathcal{Z} \subset B^n(0, r_1)$, where $B^n(0, r_1) \subset \mathbb{R}^n$ is an open ball having centre 0 and radius $r_1 > 0$. We denote the closure of this ball by $\overline{B}^n(0, r_1)$.

Let $D = m + n$, $\alpha > 0$, and define a map $H_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^{m+n}$,

$$H_\theta(x) = (\alpha x, F_\theta(x)) \in \mathbb{R}^n \times \mathbb{R}^m = \mathbb{R}^D. \quad (91)$$

Observe that the map $H_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^D$ is injective.

Let $\mathcal{V}(k, D)$ denote the set of k -tuples (v_1, v_2, \dots, v_k) where v_j are orthonormal vectors in \mathbb{R}^D . Such vectors span a k -dimensional linear space. Furthermore, let $G(k, D)$ denote the set of k -dimensional linear subspaces of \mathbb{R}^D , and for $V \in G(k, D)$, $V = \text{span}(v_1, v_2, \dots, v_k)$, let $P_V = P_{(v_1, v_2, \dots, v_k)} : \mathbb{R}^D \rightarrow \mathbb{R}^D$ be an orthogonal projection which image is the space V . As the dimension of the orthogonal group $O(k)$ is $k(k - 1)/2$ and by Milnor & Stasheff (1974), the set $G(k, D)$, called the Grassmannian, is a smooth algebraic variety, of dimension $k(D - k)$ and the dimension of $\mathcal{V}(k, D)$ is $k(D - k) + k(k - 1)/2 = k(2D - k - 1)/2$.

To prove Theorem 5, we need the following lemma.

Lemma 7. *Let $H_\theta \in \mathcal{NN}(n, D)$, $D > d \geq 2n + 1$ be a neural network such that $H_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^D$ is injective. Let $X_\theta = \{V \in G(d, D) : P_V \circ H_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^D \text{ is injective}\}$. Then the set X_θ is an intersection of countably many open and dense subsets of $G(d, D)$, that is, elements of X_θ are generic. Moreover, the $d(D - d)$ dimensional Hausdorff measure of the complement of X_θ in $G(d, D)$ is zero.*

Note that for $V \in X_\theta$, we have $P_V \circ H_\theta \in \mathcal{NN}(n, D)$.

Proof. We use that fact that $H_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^D$ is injective and locally Lipschitz-smooth. Recall that $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is locally Lipschitz if for any compact set $\mathcal{Z} \subset \mathbb{R}^n$ there is $L_{\mathcal{Z}} > 0$ such that $|f(x) - f(y)| \leq L_{\mathcal{Z}}|x - y|$ for all $x, y \in \mathcal{Z}$.

Let $v \in \mathbb{R}^m$ be a unit vector and let $Q_v : \mathbb{R}^D \rightarrow \mathbb{R}^D$ be the projection $Q_v(z) = z - (z \cdot v)v$. Let $A = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : x \neq y\}$.

As $H_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^D$ is an injection, we can define the map

$$s_\theta : A \rightarrow \mathbb{S}^{D-1} = \{w \in \mathbb{R}^D : |v| = 1\}, \quad s_\theta(x, y) = \frac{H_\theta(x) - H_\theta(y)}{|H_\theta(x) - H_\theta(y)|}. \quad (92)$$

As observed in the proof of the Whitney's embedding theorem (Hirsch, 2012, Chapter 2, Theorem 3.5), the map $Q_w \circ H_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^D$ is an injection when $w \in \mathbb{S}^{D-1}$. To see this, assume that $w \in \mathbb{S}^{D-1}$ satisfies $w \notin s_\theta(A)$, that is, w is not in the image of s_θ . Then, if there are $x, y \in \mathbb{R}^n$, $x \neq y$ such that $Q_w H_\theta(x) = Q_w H_\theta(y)$, we see that there is $t \in \mathbb{R}$ such that $H_\theta(x) - H_\theta(y) = tw$. As w is a unit vector, this yields that $t = |H_\theta(x) - H_\theta(y)| \neq 0$ and

$$w = (H_\theta(x) - H_\theta(y))/t = s_\theta(x, y),$$

which is in contradiction with the assumption that $w \notin s_\theta(A)$. Hence, $w \notin s_\theta(A)$ yields that $Q_w \circ H_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^D$ is an injection.

We consider the image $s_\theta(A) \subset \mathbb{S}^{D-1}$. For $h > 0$, let

$$A_h = \{(x, y) \in \overline{B}^n(0, h^{-1}) \times \overline{B}^n(0, h^{-1}) : |H_\theta(x) - H_\theta(y)| \geq h\}.$$

As $H_\theta : \overline{B}^n(0, h^{-1}) \rightarrow \mathbb{R}^m$ is Lipschitz-smooth with some Lipschitz constant L_h , we see that the map $s_\theta : A_h \rightarrow \mathbb{S}^{D-1}$ is Lipschitz-smooth. Since the set A has the Hausdorff dimension $2n$ and the map $s_\theta : A_h \rightarrow \mathbb{S}^{D-1}$ is Lipschitz-smooth, the Hausdorff dimension of the set $s_\theta(A_h)$ is at most $2n$, see e.g. Morgan (2016, p. 26). The set A is the union of all sets A_{h_j} , where $h_j = 1/j$ and $j \in \mathbb{Z}$. By Mattila (1995, p. 59) the Hausdorff dimension of a countable union of sets S_j is the supremum of the Hausdorff dimension of the sets S_j . Hence $s_\theta(A) = \bigcup_{j=1}^{\infty} s_\theta(A_{h_j})$ has the Hausdorff dimension less or equal $2n$.

Since the dimension $D - 1$ of \mathbb{S}^{D-1} is strictly larger than $2n$, we see that the set $s_\theta(A_{h_j})$ is closed, its complement is an open and dense set, and thus the set $Y_1(\theta) := \mathbb{S}^{D-1} \setminus s_\theta(A)$ is an intersection of countably many open and dense sets.

Observe that as Q_{w_1} is a linear map, the map $Q_{w_1} \circ H_\theta$ is also a neural network that belongs in $\mathcal{NN}(n, D)$, and we can denote $Q_{w_1} \circ H_\theta = H_{\theta_1}$ with some parameters θ_1 . Thus we can repeat the above arguments using the map $Q_{w_1} \circ H_\theta : \mathbb{R}^n \rightarrow \text{span}(w_1)^\perp \equiv \mathbb{R}^{D-1}$ instead of $H_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^D$. Repeating the above arguments $D - d$ times, can choose orthonormal vectors $w_j \in \mathbb{S}^{D-1}$, $j = 1, 2, \dots, D - d$, and sets $Y_j(\theta, w_1, \dots, w_{j-1}) \subset \mathbb{S}^{D-1} \cap (w_1, \dots, w_{j-1})^\perp$, which $D - j$ dimensional Hausdorff measures vanish and which complements are intersections of countably many open and dense sets. Let B_θ to be the set of all n -tuples (w_1, \dots, w_{D-d}) where $w_1 \in Y_1(\theta)$ and $w_j \in Y_j(\theta, w_1, \dots, w_{j-1})$ for all $j = 2, \dots, D - d$. Note that all such vectors w_j , $j = 1, 2, \dots, D - d$ are orthogonal vectors spanning a $D - d$ dimensional vector space V , and the map

$$P_V \circ H_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^D, \quad \text{where } P_V = Q_{w_{D-d}} \circ \dots \circ Q_{w_2} \circ Q_{w_1} \quad (93)$$

is injective. By the above, construction the $(D - (d + 1)/2)d$ dimensional Hausdorff measure of the complement of B_θ in $\mathcal{V}(D, d)$ is zero and B_θ is generic set. As the dimension of the set of the orthogonal basis in a d -dimensional vector space is $(d - 1)d/2$, we obtain the claim. \square

Now we continue with the proof of Theorem 5. Let $V_0 = \{(0, 0, \dots, 0)\} \times \mathbb{R}^m \subset \mathbb{R}^D$. By applying Lemma 7 with $d = m$ we see that any neighborhood of V_0 in $G(d, D)$ contain a m -dimensional vector space $V \in X_\theta$. Then $P_V \circ H_\theta$ is injective. We can choose $V \in X_\theta$ to be so close to V_0 that there is a rotation $R_V \in O(D)$ of the space \mathbb{R}^D , that maps the subspace V to the subspace V_0 , such that

$$\|R_V - \text{Id}\|_{\mathbb{R}^d \rightarrow \mathbb{R}^d} < \frac{1}{2(1 + \alpha + \|F_\theta\|_{C(\mathcal{Z})})} \varepsilon. \quad (94)$$

where Id is the identity function. Let $\pi_0 : \mathbb{R}^D \rightarrow \mathbb{R}^m$ be the map $\pi_0(y', y'') = y''$ be the projection to the last m coordinates. Then

$$\|R_V \circ Q_{P_V} \circ H_\theta - H_\theta\|_{C(\mathcal{Z})} < \frac{1}{2} \varepsilon, \quad \pi_0 \circ H_\theta = F_\theta \quad (95)$$

This and (90) imply that the claim of Theorem 5 holds for the injective neural network $N_\theta = \pi_0 \circ R_V \circ P_V \circ H_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^m$. \square

Lemma 7 used in the above proof yields also Corollary 3.

Proof of Corollary 3. Observe that a measure μ that is absolutely continuous with respect to the Lebesgue $\prod_{j=1}^k \mathbb{R}^{d_{2j} \times d_{2j-1}}$ is absolutely continuous also with respect to the normalized Gaussian distribution, and that if the set $S = \{F_k \text{ is not injective}\}$ has measure zero with respect to the normalized Gaussian distribution, then its μ -measure is also zero. Thus we can assume without loss of generality that the elements of matrices B_j are independent and have normalized Gaussian distributions.

Assume next that we have shown that $F_{j-1} : \mathbb{R}^n \rightarrow \mathbb{R}^{d_{2j-2}}$ is injective almost surely. Then, $f_\theta^{(j)} \circ F_{j-1} : \mathbb{R}^n \rightarrow \mathbb{R}^{d_{2j-1}}$ is injective almost surely. If $d_{2j} > d_{2j-1}$, the matrix B_j is almost surely injective and so is $F_j = B_j \circ f_\theta^{(j)} \circ F_{j-1} : \mathbb{R}^n \rightarrow \mathbb{R}^{d_{2j}}$. Thus, it is enough to consider the case when $d_{2j-1} \geq d_{2j} \geq 2n + 1$. Then the matrix $B_j : \mathbb{R}^{d_{2j-1}} \rightarrow \mathbb{R}^{d_{2j}}$ has almost surely rank d_{2j} . By using the singular value decomposition, we can write $B_j = R_j^1 D_j R_j^2$ where $R_j^1 \in O(d_{2j})$, $R_j^2 \in O(d_{2j-1})$, $D_j \in \mathbb{R}^{d_{2j} \times d_{2j-1}}$ is matrix which principal diagonal elements are almost surely strictly positive and the other elements are zeros. Let $V_j = (R_j^2)^{-1}(\mathbb{R}^{d_{2j}} \times \{(0, 0, \dots, 0)\}) \subset \mathbb{R}^{d_{2j-1}}$. Let P_{V_j} be an orthogonal projector in $\mathbb{R}^{d_{2j-1}}$ onto the space V_j of dimension d_{2j} . As the distribution of the matrix B_j is invariant in rotations of the space, so is the distribution of linear space V_j in $G(d_{2j-1}, d_{2j})$. By Lemma 7, we see that the map $P_{V_j} \circ f_\theta^{(j)} \circ F_{j-1} : \mathbb{R}^n \rightarrow \mathbb{R}^{d_{2j-1}}$ is injective almost surely. This implies that the map $B_j \circ f_\theta^{(j)} \circ F_{j-1} : \mathbb{R}^n \rightarrow \mathbb{R}^{d_{2j}}$ is injective almost surely, that is, the map $F_j : \mathbb{R}^n \rightarrow \mathbb{R}^{d_{2j}}$ is injective almost surely. The claim follows by induction. \square

In Baraniuk & Wakin (2009), see also Hegde et al. (2008); Iwen & Maggioni (2013), Broomhead & Kirby (2001; 2000), the authors study manifold learning using random projectors. These results are related to the proof of Theorem 5 above. Let H_θ be given by (91), a ReLU-based neural network whose graph $M \subset \mathbb{R}^d$, $d = 2n + m$. When P_V is a random projector in \mathbb{R}^d onto a m -dimensional linear subspace V , the injectivity of the neural network $P_V \circ H_\theta$ is closely related to the property that $P_V(M)$ is an n -dimensional submanifold with a large probability. In Broomhead & Kirby (2001; 2000) the authors use Whitney embedding results for C^2 -smooth manifold for dimension reduction of data. Our proof applies similar techniques for Lipschitz-smooth maps. In Baraniuk & Wakin (2009); Hegde et al. (2008); Iwen & Maggioni (2013), the authors apply the result that when $M \subset \mathbb{R}^D$ is a submanifold and D is large enough, a random m -dimensional projector P_V satisfies on M the restricted isometry property with a large probability. In this case, $P_V \circ H_\theta$ is not only an injection but its inverse map is also a local Lipschitz map. In this sense, the techniques in Baraniuk & Wakin (2009); Hegde et al. (2008); Iwen & Maggioni (2013) would give improved results to the generic projection technique used in this paper. The results in Baraniuk & Wakin (2009); Hegde et al. (2008); Iwen & Maggioni (2013), however, require that the dimension m of image space of the map $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ satisfies $m \geq C \log(\varepsilon^{-1})$, where ε is the precision parameter in the inequality (10). Our result, Theorem 5 requires only that $m \geq 2n + 1$.

APPENDIX D MISCELLANEOUS LEMMAS

Lemma 8. *Let $x, y, z \in \mathbb{R}$ and $z \geq 0$. Then*

$$\text{ReLU}(x + z) = \text{ReLU}(y + z) \implies \text{ReLU}(x) = \text{ReLU}(y). \quad (96)$$

Lemma 9 (Useful Inequalities). *The following geometric inequalities are useful and have straight forward proofs. Suppose that $a, b, c \in \mathbb{R}^n$, then*

1. *For $i = 1, \dots, k$, let $a_i \in \mathbb{R}^m$. If*

$$\sum_{i=1}^k \|a_i\|_2^2 \leq \left\| \sum_{i=1}^k a_i \right\|_2^2, \quad (97)$$

then if for each $j = 1, \dots, m$, $a_i|_j \cdot a_{i'}|_j \geq 0$ for each pair $i, i' \in [[k]]$, then

$$\sum_{i=1}^k \|a_i\|_2 \leq \sqrt{k} \left\| \sum_{i=1}^k a_i \right\|_2. \quad (98)$$

2. If $a, b \in \mathbb{R}^n$ and $\langle a, b \rangle \geq 0$, then

$$\|a\|_2^2 + \|b\|_2^2 \leq \|a + b\|_2^2. \quad (99)$$

Lemma 10 (Co-linear Additivity of $\text{ReLU}(W \cdot)$). *Let $W \in \mathbb{R}^{m \times n}$, $x_1, x_2 \in \mathbb{R}^n$, $\ell^{x_1, x_2}(t) = (1-t)x_1 + tx_2$. Let there be*

$$0 = t_1 \leq t_2 \leq \dots \leq t_{n_t-1} \leq t_{n_t} = 1 \quad (100)$$

then

$$\sum_{k=1}^{n_t-1} \|\text{ReLU}(Wx_{t_k}) - \text{ReLU}(Wx_{t_{k+1}})\|_2^2 \leq \|\text{ReLU}(Wx_1) - \text{ReLU}(Wx_2)\|_2^2. \quad (101)$$

and

$$\sum_{k=1}^{n_t-1} \|\text{ReLU}(Wx_{t_k}) - \text{ReLU}(Wx_{t_{k+1}})\|_2 \leq \sqrt{n_t} \|\text{ReLU}(Wx_1) - \text{ReLU}(Wx_2)\|_2. \quad (102)$$

Proof. Let $x_t = \ell^{x_1, x_2}(t)$, then as a function of t , the j 'th component of $\text{ReLU}(Wx_t)$ is either increasing (if $\langle w_j, x_2 - x_1 \rangle \geq 0$) or decreasing (if $\langle w_j, x_2 - x_1 \rangle \leq 0$). In either case, it is clear that for each $j = 1, \dots, m$

$$(\text{ReLU}(Wx_1) - \text{ReLU}(Wx_2))|_j \cdot (\text{ReLU}(Wx_2) - \text{ReLU}(Wx_3))|_j \geq 0 \quad (103)$$

hence clearly

$$\langle (\text{ReLU}(Wx_1) - \text{ReLU}(Wx_2)), (\text{ReLU}(Wx_2) - \text{ReLU}(Wx_3)) \rangle \geq 0 \quad (104)$$

thus we can apply (99) and we obtain (101). Applying (98) then yields (102). \square

APPENDIX E DETAILED COMPARISON TO PRIOR WORK

E.1 COMPARISON TO BRUNA *et al.*

In [Bruna et al. \(2013, Proposition 2.2.\)](#) the authors give a result invoking a condition similar to our DSS condition ([Definition 1](#)). It also concerns injectivity of a ReLU layer in terms of the injectivity of the weight matrix restricted to certain rows. The authors also compute a bi-Lipschitz bound for a layer (similar to our [Theorem 3](#)), though as we show in the following examples their analysis is in some cases not precisely aligned with injectivity.

Their criterion is given in two parts. For a weight matrix, they first define a notion of admissible set which indicates the points where the weight matrix's injectivity must be tested. Injectivity follows provided that the weight matrix is non-singular when restricted to each admissible set. Given a weight matrix $W \in \mathbb{R}^{M \times N}$ and bias b , the authors say that $\Omega \subset \{1, \dots, M\}$ is admissible if

$$\bigcap_{i \in \Omega} \{x: \langle x, w_i \rangle > b_i\} \cap \bigcap_{i \notin \Omega} \{x: \langle x, w_i \rangle < b_i\} \quad (105)$$

is not empty. For our analysis we focus on the case when $b \equiv 0$. In this case Ω is admissible if and only if

$$\exists x \in \mathbb{R}^n \quad \text{such that} \quad \langle x, w_i \rangle \begin{cases} > 0 & \text{if } i \in \Omega \\ < 0 & \text{if } i \notin \Omega \end{cases}. \quad (106)$$

Note that the inequality in (106) is strict, unlike (2). If, for example, W has a column that is the zero vector, then there are no admissible Ω . The authors use the notation $\bar{\Omega}$ to denote all admissible sets for a given weight matrix. In their notation F is the transpose of our weight matrix W , F_Ω are the Ω rows of the weight matrix, $F_\Omega|_{V_\Omega}$ is the subspace generated by the Ω rows of W . The authors also call the ReLU function the half-rectification function. $\lambda_-(F)$ and $\lambda_+(F)$ denote the lower and upper frame bounds of F respectively. The injectivity criterion from [Bruna et al. \(2013\)](#) is

Proposition 1. *Let $A_0 = \min_{\Omega \in \bar{\Omega}} \lambda_-(F_\Omega|_{V_\Omega})$. Then the half-rectification operator $M_b(x) = \text{ReLU}(F^T x + b)$ is injective if and only if $A_0 > 0$. Moreover, it satisfies*

$$\forall x, x', A_0 \|x - x'\| \leq \|M_b(x) - M_b(x')\| \leq B_0 \|x - x'\| \quad (107)$$

with $B_0 = \max_{\Omega \in \bar{\Omega}} \lambda_+(F_\Omega) \leq \lambda_+(F)$.

We now show that Proposition 1 does not precisely align with injectivity of $\text{ReLU}(W(\cdot))$. We construct a weight matrix for which $A_0 > 0$, but does not yield an injective $\text{ReLU}(W(\cdot))$. If

$$W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix} \quad (108)$$

then clearly $\text{ReLU}(Wx)$ is not injective (for all $\alpha < 0$, $\text{ReLU}(W \begin{bmatrix} 0 \\ \alpha \end{bmatrix}) = \bar{0}$). The only admissible sets are $\bar{\Omega} = \{\{1\}, \{3\}, \{1, 2\}, \{2, 3\}\}$ (notably $\{2\}$ is not admissible). W is full rank on all $\Omega \in \bar{\Omega}$, so $A_0 > 0$ so Proposition 1 implies that $\text{ReLU}(W(\cdot))$ is injective. Now consider the case when

$$W = \begin{bmatrix} B \\ -DB \\ 0 \end{bmatrix} \quad (109)$$

where B is a basis of \mathbb{R}^n , D is a strictly positive diagonal matrix, and 0 is the zero row vector. From Corollary 2, W satisfies Theorem 1, and so $\text{ReLU}(W(\cdot))$ is injective. On account of the zero row vector in (109), $\forall x \in \mathbb{R}^n$, $\langle x, 0 \rangle = 0$ so there are no Ω that are admissible Ω according to (105). Thus A_0 is undefined.

Now we construct an example of a W and $x, x' \in \mathbb{R}^n$ for which $A_0 \|x - x'\| > \|\text{ReLU}(Wx) - \text{ReLU}(Wx')\|$. Let

$$W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad x = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad x' = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}. \quad (110)$$

Clearly on every admissible set $\lambda_-(W_\Omega|_{V_\Omega}) = 1$, so

$$\begin{aligned} \text{ReLU}(Wx) &= \frac{1}{\sqrt{2}} \text{ReLU} \left(\begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} \right) = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \\ \text{ReLU}(Wx') &= \frac{1}{\sqrt{2}} \text{ReLU} \left(\begin{bmatrix} -1 \\ 1 \\ 1 \\ -1 \end{bmatrix} \right) = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \end{aligned} \quad (111)$$

hence,

$$\|\text{ReLU}(Wx) - \text{ReLU}(Wx')\| = \frac{1}{\sqrt{2}} \left\| \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix} \right\| = 1 \quad (112)$$

and

$$\|x - x'\| = \frac{1}{\sqrt{2}} \left\| \begin{bmatrix} 2 \\ 0 \end{bmatrix} \right\| = \sqrt{2}. \quad (113)$$

From this we have

$$\sqrt{2} = A_0 \|x - x'\| > \|\text{ReLU}(Wx) - \text{ReLU}(Wx')\| = 1. \quad (114)$$

On the other hand, substituting this into (6) yields

$$\frac{C}{\sqrt{m}} \|x - x'\| = \frac{1}{\sqrt{8}} \sqrt{2} = \frac{1}{2} \leq 1, \quad (115)$$

which does hold, suggesting that the lower bound in Proposition 1 is not pessimistic enough.

E.2 RELATIONSHIP TO MALLAT ET AL. (2018)

In Mallat et al. (2018) the authors consider a construction analogous to our convolutional construction (in Definition 3) defined on a continuum (i.e. infinite-dimensional function defined on an interval) rather than on a vector (i.e. discrete finite dimensional function) defined on a subset of \mathbb{R}^n . The authors posit that CNNs first learn a layer of filters localized in frequency varied in phase. The authors also show that a ReLU activation function acts as a filter on the phase of the convolution of the filters against the input signal and that, provided that the filters are sufficiently different in phase and satisfy a frame condition then the layer is bi-Lipschitz, and hence is injective. Their analysis a particularized version of ours, and can be straight forwardly subsumed by our work.

The frame condition is given by Proposition 2.6 in Mallat et al. (2018) that the weight matrix must satisfy in order ensure that W is invertible and stable. In the notation of Mallat et al. (2018) the filters $\hat{\psi}_\lambda$ are analogous to the Fourier transform of the kernels in Definition 2, and the condition in (2.25) in Mallat et al. (2018) is one natural way to generalize the notion of a basis to a continuous signal. Hence, Proposition 2.6 in Mallat et al. (2018) can be loosely interpreted as a statement that the kernels in a given layer of width P form a basis of \mathbb{R}^P .

The second condition is that kernels are given in terms of belonging to a family, and members of this family are related to each other in the sense that members of the same family are centered in the same Fourier domain, and act as phase offsets of differing phase. Equation 2.14 of Mallat et al. (2018) describes that a phase filter $H: \mathbb{C} \times [0, 2, \pi] \rightarrow \mathbb{C}$ is defined by

$$\forall z \in \mathbb{C}, \alpha \in [0, 2\pi], \quad Hz(\alpha) = |z|h(\alpha - \varphi(z)) \quad (116)$$

where $|z|$ standard modulus of a complex number, $\varphi(z)$ is the complex phase, and $h(\alpha) = \text{ReLU}(\cos(\alpha))$. If we consider just the value of $Hz(0)$ and $Hz(\pi)$, then we find

$$Hz(0) = |z| \text{ReLU}(\cos(\varphi(z))) = \text{ReLU}(|z| \cos(\varphi(z))) = \text{ReLU}(\mathcal{R}z), \quad (117)$$

$$Hz(\pi) = |z| \text{ReLU}(-\cos(\varphi(z))) = \text{ReLU}(-|z| \cos(\varphi(z))) = \text{ReLU}(-\mathcal{R}z), \quad (118)$$

where $\mathcal{R}z$ is the real part of z . If (as in Mallat et al. (2018)) z is given by $z = x \star c_k$, where x is a real signal and $\star c_k$ denotes the convolution against a kernel c_k , then (117) and (118) imply that

$$Hz(0) = \text{ReLU}(x \star \mathcal{R}c_k), \quad (119)$$

$$Hz(\pi) = \text{ReLU}(x \star (-\mathcal{R}c_k)), \quad (120)$$

that is, that for every kernel c_k in a layer, the kernel $-c_k$ is also in that layer.-

Combining the two logical conditions above implies that the kernels of width c_k form a basis of \mathbb{R}^P and that for every kernel c_k there is also a kernel $-c_k$. Together these two (by Corollary 2) that the c_k form a DSS of \mathbb{R}^P , and thus by Theorem 4 the entire layer is injective.

APPENDIX F ARCHITECTURE DETAILS FOR EXPERIMENTS

Generator network: We train a generator with 5 convolutional layers. The input latent code is 256-dimensional which is treated by the network as a $1 \times 1 \times 256$ size tensor. The first layer is a transposed convolution with a kernel size of 4×4 with stride 1 and 1024 output channels. This is followed by a leaky ReLU. We follow this up by 3 conv layers each of which halve the number of channels and double the image size (i.e. we go from $N/2 \times N/2 \times C$ to $N \times N \times C/2$ tensor) giving an expansivity of two, the minimum required for injectivity of ReLU networks. Each of these 3 convolution layers has kernel size 3, stride 2 and is followed by the ReLU activation. These layers are made injective by having half the filters as w and the other half as $-s^2w$. Here, w and s are trainable parameters. The biases in these layers are kept at zero. We do not employ any normalization schemes. Lastly, we have a convolution layer at the end to get to 3 channels and required image size. This layer is followed by the sigmoidal activation. We compare this to a regular GAN which has all the same architectural components including nonlinearities except the filters are not chosen as w and $-s^2w$ and we also allow biases (see Figure 9 for a qualitative comparison).



Figure 9: Samples generated with a DCGAN on FFHQ dataset: injective layers (left) vs generic layers (right)

Critic network: The discriminator has 5 convolution layers with 128, 256, 512, 1024 and 1 channels per layer. Each convolution layer has 4×4 kernels with stride 2. Each layer is followed by the leaky-ReLU activation function. The last layer of the network is followed by identity.

Inference network: The inference network has the same architecture as the first 4 convolution layers of the discriminator. This is followed by 3 fully-connected layers of size 512, 256 and 256. The first 2 fully-connected layers have a Leaky ReLU activation while the last layer has identity activation function. The inference net is trained in tandem with the GAN.

We use the Wasserstein loss with gradient penalty (Gulrajani et al., 2017) to train our networks. We train for 40 epochs on a data set of size 80000 samples. We use a batch size of 64 and Adam optimizer for training with learning rate of 10^{-4} .

We report FID (Heusel et al., 2017) and Inception score (Salimans et al., 2016) using 10000 generated samples. The standard deviation was calculated using 5 sets of 10000 generated samples. In order to calculate the mean and covariance of generated distributions, we sample 50000 codes.