Globally Injective (ReLU) Neural Networks

Maarten V. de Hoop[1]

Michael Puthawala[1], Konik Kothari[2], Matti Lassas[3], Ivan Dokmanić[42]

[1]Rice University, [2]University of Illinois at Urbana-Champaign, [3]University of Helsinki, [4]University of Basel

Simons Foundation MATH + X, DOE, NSF-DMS, Geo-Mathematical Imaging Group, Nvidia

UCI, July 10, 2020

deep learning

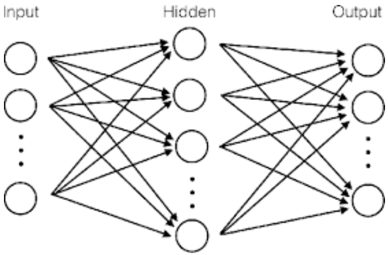
deep neural networks – train set of parameters θ so that

$$N_\theta \colon \mathcal{Z} \to \mathcal{X}$$

maps some given $\mathcal{Z} \supset \{z_i\}_{i=1,\dots,I}$ to given $\{x_i\}_{i=1,\dots,I} \subset \mathcal{X}$

functionality

- classification
- regression
- generation
- encoding, decoding (autoencoder)
- inference



trained deep neural networks vs inverse problems

natural questions

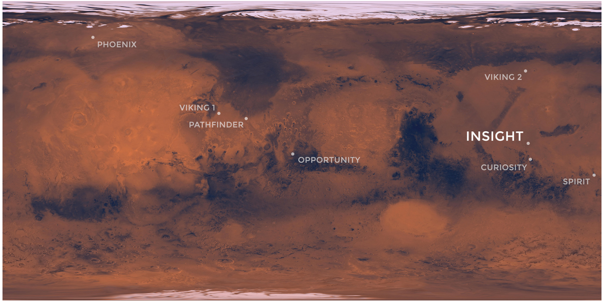
- injectivity (uniqueness)
- stability, quantitative estimate
- reconstruction

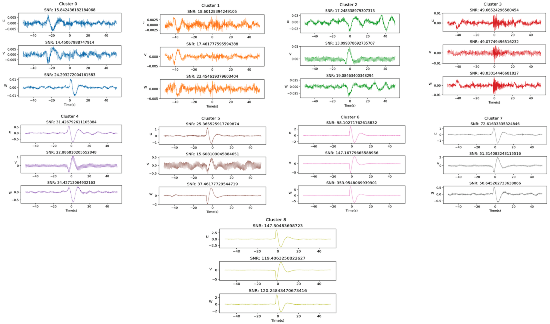
trained deep neural networks vs inverse problems

natural questions

- injectivity (uniqueness)
- stability, quantitative estimate
- reconstruction

implied properties: *approximation*, topological, probabilistic

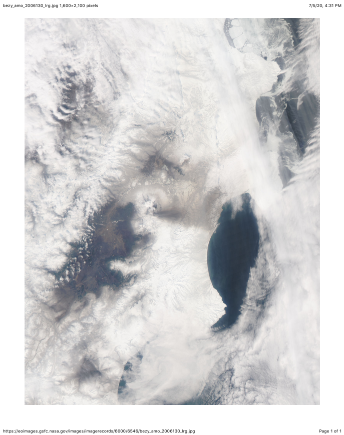


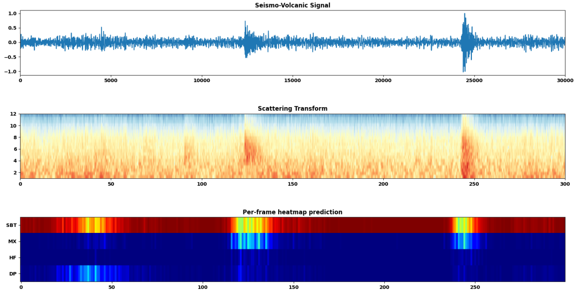


unsupervised learning, clustering – separation

(Barkaoui, Lognonné, Kawamura, Stutzmann, Seydoux, dH, Balestriero, Schloz, Clinton, Stahler, Van Driel, Ceylan, Sainton & Banerdt)

injectivity and data driven discovery





(semi)supervised learning, polyphonic detection, segmentation and classification

(Bueno, Balestriero, dH, Baraniuk, Benítez & Ibáñez)

deep neural networks

feed-forward network, $N \colon \mathbb{R}^n \to \mathbb{R}^m$ affine transformations

$$N(z) = W_{L+1}\phi_L(W_L \cdots \phi_2(W_2\phi_1(W_1 z + b_1) + b_2) \cdots + b_L)$$

- $\ell = 1, \ldots, L$ index the network layers
- $b_\ell \in \mathbb{R}^{n_{\ell+1}}$ are the bias vectors
- $W_\ell \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$ are the weight matrices with $n_1 = n$, $n_{L+1} = m$
- ϕ_ℓ are the nonlinear activation functions

deep neural networks

feed-forward network, $N\colon \mathbb{R}^n \to \mathbb{R}^m$

affine transformations

$$N(z) = W_{L+1}\phi_L(W_L \cdots \phi_2(W_2\phi_1(W_1 z + b_1) + b_2) \cdots + b_L)$$

- $\ell = 1, \ldots, L$ index the network layers
- $b_\ell \in \mathbb{R}^{n_{\ell+1}}$ are the bias vectors
- $W_\ell \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$ are the weight matrices with $n_1 = n$, $n_{L+1} = m$
- ϕ_ℓ are the nonlinear activation functions

$\mathcal{NN}(n, m)$

$\theta = (W_1, b_1, \ldots, W_L, b_L, W_{L+1})$

deep neural networks

feed-forward network, $N\colon \mathbb{R}^n \to \mathbb{R}^m$ affine transformations

$$N(z) = W_{L+1}\phi_L(W_L \cdots \phi_2(W_2\phi_1(W_1 z + b_1) + b_2) \cdots + b_L)$$

- $\ell = 1, \ldots, L$ index the network layers
- $b_\ell \in \mathbb{R}^{n_{\ell+1}}$ are the bias vectors
- $W_\ell \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$ are the weight matrices with $n_1 = n$, $n_{L+1} = m$
- ϕ_ℓ are the nonlinear activation functions

$\mathcal{N}\mathcal{N}(n, m)$ $\theta = (W_1, b_1, \ldots, W_L, b_L, W_{L+1})$

layerwise analysis: $\phi_\ell(W_\ell x + b_\ell)$ *injectivity*

intermezzo: skip connections

$\tilde{N}\colon \mathbb{R}^n \to \mathbb{R}^m$

$$
\begin{aligned}
\tilde{N}(z) &= h_{L+1} \\
h_{\ell+1} &= \left(\sum_{p=1}^{\ell} \tilde{A}_\ell^p h_p + \tilde{b}_\ell\right) + \phi_\ell\left[\sum_{p=1}^{\ell} A_\ell^p h_p + b_\ell\right] \\
h_0 &= z
\end{aligned}
$$

- $\ell = 1, \ldots, L$ index the network layers
- $\tilde{b}_\ell, b_\ell \in \mathbb{R}^{n_{\ell+1}}$ are the bias vectors
- $\tilde{A}_\ell^p, A_\ell^p \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$, $p \leq \ell$ are the weight matrices with $n_1 = n$, $n_{L+1} = m$

$\mathcal{N}\mathcal{N}_{\text{skip}}(n, m)$

nonlinear activation functions

if ϕ is a one-to-one activation function ($\mathrm{L\,ReLU}_\alpha, \sigma, \tanh$) then there is not much to be done: the layer is injective iff W is injective

focus exclusively on

$$\phi(x) = \mathrm{ReLU}(x) := \max(x, 0)$$

nonlinear activation functions

$\phi := \phi_\ell$

if ϕ is a one-to-one activation function ($\mathrm{L\,ReLU}_\alpha, \sigma, \tanh$) then there is not much to be done: the layer is injective iff W is injective

focus exclusively on

$$\phi(x) = \mathrm{ReLU}(x) := \max(x, 0)$$

- unrolling: ISTA (sparse code inference) (Gregor & Lecun, 10)
- interpolation: optimal activation function regularized by TV^2 norm (linear spline, unknown knots) (Unser, '19)
- expressivity: exponential in number of layers .. (Balestriero & Baraniuk, '20)

$$W = \{w_i\}_{i=1}^m, \quad w_i \in \mathbb{R}^n$$

weight matrices

$$W = \{w_i\}_{i=1}^m, \quad w_i \in \mathbb{R}^n$$

- dense: fully connected

$$W = \{w_i\}_{i=1}^m, \quad w_i \in \mathbb{R}^n$$

- dense: fully connected
- convolutional (CNNs), multi-indices $N = (N_1, \ldots, N_p)$, ..

$C \in \mathbb{R}^{N \times N}$, stride 1

$$\mathbb{R}^{M \times N} \ni W = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_{n_Q} \end{bmatrix} \text{ where for each } C,\ J\text{: } (Cx)_J = \sum_{I=1}^{O} c_{O-I+1} x_{J+I} = \sum_{I'=1+J}^{O+J} c_{O+J-I'+1} x_{I'}$$

$x \in \mathbb{R}^N$, $c \in \mathbb{R}^O$ (kernels, width O), n_Q convolutions

$$W = \{w_i\}_{i=1}^m, \quad w_i \in \mathbb{R}^n$$

- dense: fully connected
- convolutional (CNNs), multi-indices $N = (N_1, \ldots, N_p)$, .. $C \in \mathbb{R}^{N\times N}$, stride 1

$$\mathbb{R}^{M\times N} \ni W = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_{n_Q} \end{bmatrix} \text{ where for each } C,\ J\text{: } (Cx)_J = \sum_{I=1}^{O} c_{O-I+1} x_{J+I} = \sum_{I'=1+J}^{O+J} c_{O+J-I'+1} x_{I'}$$

$x \in \mathbb{R}^N$, $c \in \mathbb{R}^O$ (kernels, width O), n_Q convolutions

- iid Gaussian elements: weight matrices follow a Gaussian distribution with zero mean

permutations

layer-wise injectivity is clearly not dependent on the arrangement of the rows of W; thus we often refer to the rows of $W \in \mathbb{R}^{m \times n}$ using set notation

$$w \in W : \ w \text{ is a row vector of } W$$

prior work

(Bruna, Salzam & LeCun, '13): injectivity of pooling motivated by the problem of signal recovery from feature representations

(Hand & Voroninski, '18): optimization landscape for inverting ReLU generative priors

(Mallat, Zhang & Rochette, '18): ReLU activation function acts as a phase filter and that the layer is bi-Lipschitz, and hence injective, provided that the filters have a sufficiently diverse phase and form a frame

(Lei, Jalal, Dhillon & Dimakis, '19): with high probability a layer of a neural network can be inverted about a fixed point provided that the weights are normally distributed and that the network is expansive by a factor of at least 2.1

prior work

(Bruna, Salzam & LeCun, '13): injectivity of pooling motivated by the problem of signal recovery from feature representations

(Hand & Voroninski, '18): optimization landscape for inverting ReLU generative priors

(Mallat, Zhang & Rochette, '18): ReLU activation function acts as a phase filter and that the layer is bi-Lipschitz, and hence injective, provided that the filters have a sufficiently diverse phase and form a frame

(Lei, Jalal, Dhillon & Dimakis, '19): with high probability a layer of a neural network can be inverted about a fixed point provided that the weights are normally distributed and that the network is expansive by a factor of at least 2.1

injectivity is automatic from invertible neural networks such as normalizing flows (Kingma & Dhariwal, '18)

injectivity seems to be a natural heuristic to increase latent space capacity without increasing its dimension (Brock, Lim, Ritchie & Weston,' 16)

Directed Spanning Set (DSS)

fundamental notion in our analysis

Definition (Directed Spanning Set)

Let $Y = \{y_i\}_{i=1,\dots,m}$ be a set of vectors, $y_i \in \mathbb{R}^n$. We say that Y has a Directed Spanning Set (DSS) of $\Omega \subset \mathbb{R}^n$ with respect to a vector $x \in \mathbb{R}^n$ if there exists a $\hat{Y}_x \subset Y$ such that for all

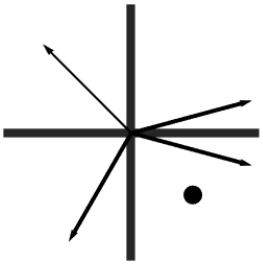
$$\forall y_i \in \hat{Y}_x, \quad \langle y_i, x\rangle \geq 0$$

and $\Omega \subset \operatorname{span}(\hat{Y}_x)$. Equivalently, Y is a DSS w.r.t. x if $\hat{Y}_x$ spans Ω and all elements of $\hat{Y}_x$ lie on the same (closed) side of the plane with normal x as x does (or all of Y if $x = 0$).

$$S(x, W) = \{i \in [[m]] \colon \langle w_i, x\rangle \geq 0\}\,, \quad [[m]] = \{1, \dots, m\} \qquad \text{complement } S^c(x, W)$$

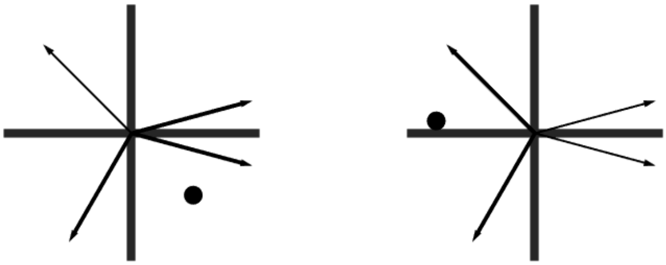
DSS illustration: $\hat{Y}_x$ (bold)

● : x



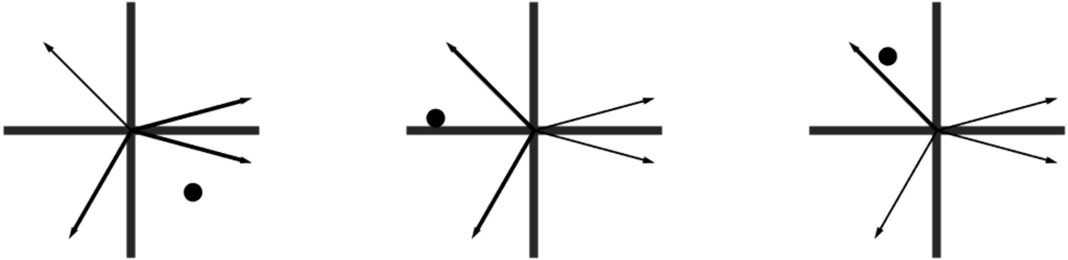
DSS illustration: $\hat{Y}_x$ (bold)

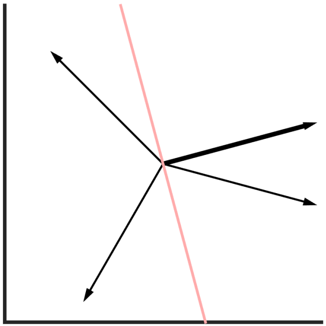
●: x

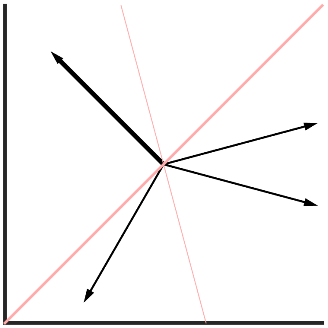


DSS illustration: $\hat{Y}_x$ (bold)

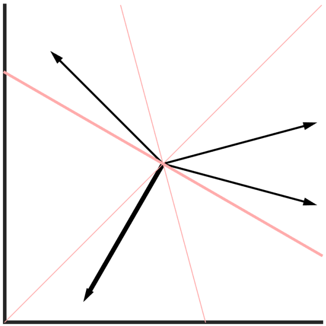
● : x

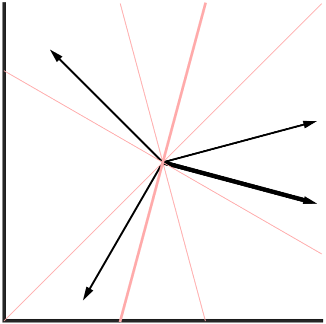


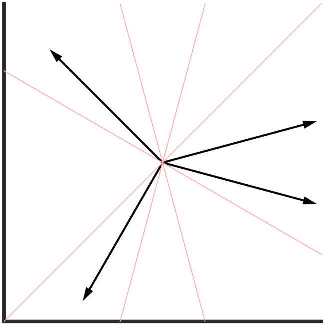


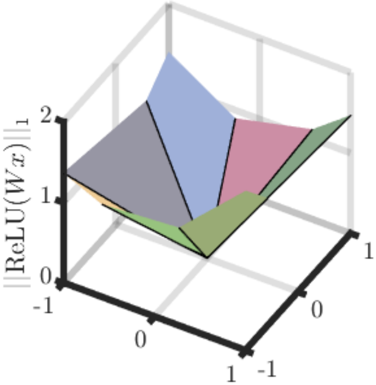


wedges, division of input space









$\|\mathrm{ReLU}(Wx)\|_1$
2
1
0
-1
0
1
-1
0
1

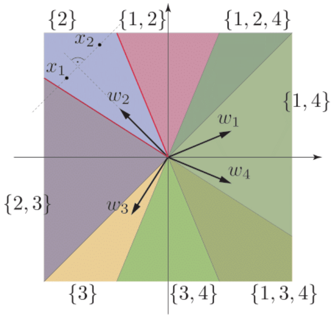
wedges, division of input space

W partitions $\mathbb{R}^n$ into open wedges S_k, $\mathbb{R}^n = \bigcup_k S_k$, with constant sign patterns:

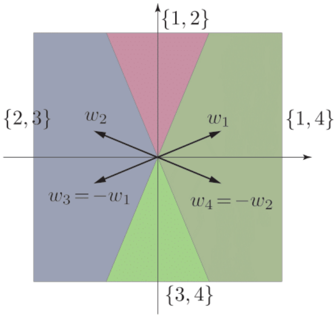
$$\text{for } x_1, x_2 \in S_k, \;\; \text{sign}(Wx_1) = \text{sign}(Wx_2)$$

W partitions $\mathbb{R}^n$ into open wedges S_k, $\mathbb{R}^n = \bigcup_k S_k$, with constant sign patterns:

$$\text{for } x_1, x_2 \in S_k, \quad \text{sign}(Wx_1) = \text{sign}(Wx_2)$$



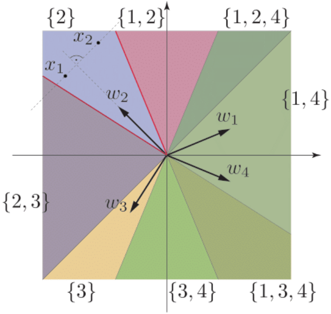
non-DDS



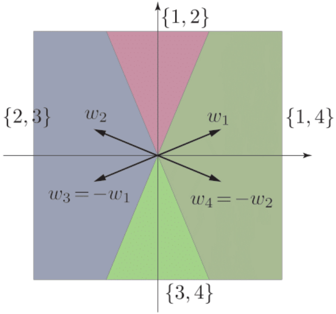
DSS

W partitions $\mathbb{R}^n$ into open wedges S_k, $\mathbb{R}^n = \bigcup_k S_k$, with constant sign patterns:

$$\text{for } x_1, x_2 \in S_k, \quad \text{sign}(Wx_1) = \text{sign}(Wx_2)$$



non-DDS



DSS

the number of wedges can be exponential in m, $\frac{m}{n} = c \geq 2$ fixed (Winder, '66)

layerwise injectivity

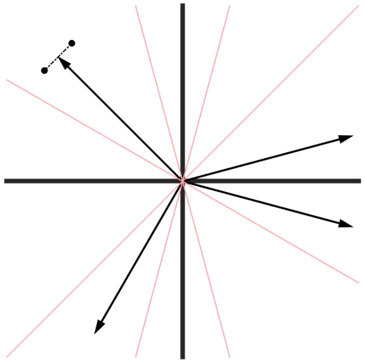
Theorem (ReLU(Wx))

Let $W \in \mathbb{R}^{m\times n}$ where $n > 1$ be a matrix with row vectors $\{w_j\}_{j=1}^m$. The function $\mathrm{ReLU}(W\,\cdot) \colon \mathbb{R}^n \to \mathbb{R}^m$ is injective if and only if W is a DSS w.r.t every $x \in \mathbb{R}^n$.

elements of proof

reverse direction

- suppose there is an x such that W does not contain a DSS w.r.t. x
- let $x^{\perp} \in \ker(W|_{S(x,W)})$, $\alpha \in \mathbb{R}^{+}$ such that $\alpha < \min_{j \in S^c(x,W)} -\frac{\langle x, w_j \rangle}{|\langle x^{\perp}, w_j \rangle|}$
- then $\mathrm{ReLU}(W(x + \alpha x^{\perp})) = \mathrm{ReLU}(Wx)$



Lemma (ReLU($Wx + b$))

Let $W \in \mathbb{R}^{m\times n}$ *and* $b \in \mathbb{R}^m$. *The function* $\mathrm{ReLU}(W \cdot + b)\colon \mathbb{R}^n \to \mathbb{R}^m$ *is injective if and only if* $\mathrm{ReLU}(W|_{b\geq 0} \cdot)$ *is injective, where* $W|_{b\geq 0} \in \mathbb{R}^{m\times n}$ *is row-wise the same as* W *where* $b_i \geq 0$, *and is a row of zeroes when* $b_i < 0$.

layerwise injectivity, bias

Lemma (ReLU($Wx + b$))

Let $W \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. The function $\mathrm{ReLU}(W\cdot + b)\colon \mathbb{R}^n \to \mathbb{R}^m$ *is injective if and only if* $\mathrm{ReLU}(W|_{b\geq 0}\,\cdot)$ *is injective, where* $W|_{b\geq 0} \in \mathbb{R}^{m\times n}$ *is row-wise the same as* W *where* $b_i \geq 0$, *and is a row of zeroes when* $b_i < 0$.

layerwise injectivity implies end-to-end injectivity it is apparent that if $W \in \mathbb{R}^{m \times n}$ the larger the ratio of $c = \frac{m}{n}$, the *expansivity*, the 'more likely' it is that $\mathrm{ReLU}(W\cdot)$ is injective

Corollary

For any $W \in \mathbb{R}^{m \times n}$, $\mathrm{ReLU}(W\cdot)$ *is non-injective if* $m < 2 \cdot n$. *If* $W \in \mathbb{R}^{2n \times n}$ *and satisfies the conditions in the Theorem and Lemma, then (up to row rearrangement)* W *can be written as*

$$W = \begin{bmatrix} B \\ -DB \end{bmatrix}$$

where $B, D \in \mathbb{R}^{n \times n}$, B *is a basis, and* D *a diagonal matrix with strictly positive diagonal entries.*

it is apparent that if $W \in \mathbb{R}^{m \times n}$ the larger the ratio of $c = \frac{m}{n}$, the *expansivity*, the 'more likely' it is that $\mathrm{ReLU}(W\cdot)$ is injective

Corollary

For any $W \in \mathbb{R}^{m \times n}$, $\mathrm{ReLU}(W\cdot)$ is non-injective if $m < 2 \cdot n$. If $W \in \mathbb{R}^{2n \times n}$ and satisfies the conditions in the Theorem and Lemma, then (up to row rearrangement) W can be written as

$$W = \begin{bmatrix} B \\ -DB \end{bmatrix}$$

where $B, D \in \mathbb{R}^{n \times n}$, B is a basis, and D a diagonal matrix with strictly positive diagonal entries.

in general, no deterministic characterization for $m > 2n$

define $\mathcal{W}_{m,n}$ as the distribution of weight matrices in $\mathbb{R}^{m\times n}$ with iid Gaussian elements,

$$\mathcal{I}(m,n) = \mathbb{P}\left(\text{ReLU}(Wx) \text{ is injective where } W \sim \mathcal{W}_{m,n}\right)$$

consider $\mathcal{I}(m,n)$ as $n \to \infty$ for fixed $c := \frac{m}{n}$

H: binary entropy function

define $\mathcal{W}_{m,n}$ as the distribution of weight matrices in $\mathbb{R}^{m\times n}$ with iid Gaussian elements,

$$\mathcal{I}(m,n) = \mathbb{P}\left(\mathrm{ReLU}(Wx) \text{ is injective where } W \sim \mathcal{W}_{m,n}\right)$$

consider $\mathcal{I}(m,n)$ as $n \to \infty$ for fixed $c := \frac{m}{n}$

Theorem

If c is greater than a certain value c^* (approximately equal to 5.7), then

$$\mathcal{I}(m,n) \geq 1 - \exp(-\Omega(n)) \qquad \Omega(n) : \textit{at least } \mathcal{O}(n)$$

If c is less than a $c^\dagger$ (approximately 3.4), then

$$\mathcal{I}(m,n) \to 0.$$

$-\log_2(ce) + c - 1 - H\left(\frac{1}{c-1}\right) > 0$ for $c \geq c^*$

$\frac{1}{2}\mathrm{erfc}\left(\frac{1}{\sqrt{2c^\dagger}}\right) = \frac{1}{c^\dagger}$

convolutional layers

set of zero-padded kernels for $c \in \mathbb{R}^O$: think of a multi-index as a box (or a hyperrectangle);

set of zero-padded kernels for $c \in \mathbb{R}^O$: think of a multi-index as a box (or a hyperrectangle); let P be a multi-index such that O 'fits' in P, then define

$$\mathcal{Z}_P(c) = \{d \in \mathbb{R}^P \ : \ d \text{ is a shift of } c \text{ within the box } P\}$$

set of zero-padded kernels for $c \in \mathbb{R}^O$: think of a multi-index as a box (or a hyperrectangle); let P be a multi-index such that O 'fits' in P, then define

$$\mathcal{Z}_P(c) = \{d \in \mathbb{R}^P \ : \ d \text{ is a shift of } c \text{ within the box } P\}$$

Theorem

Suppose that $W \in \mathbb{R}^{M \times N}$ is a convolution layer with convolutions $\{C_k\}_{k=1}^{n_Q}$, and corresponding kernels $\{c_k\}_{k=1}^{n_Q}$. If for any P,

$$W|_{\mathcal{Z}_P} := \bigcup_{k=1}^{n_Q} \mathcal{Z}_P(c_k)$$

is a DSS for $\mathbb{R}^P$ with respect to all $x \in \mathbb{R}^P$, then $\mathrm{ReLU}(W\cdot)$ is injective.

d and domain decomposition

$\mathcal{Z}_P(c)$

P vs Ω_k, $\operatorname{span}\{\Omega_1, \ldots, \Omega_K\} = \mathbb{R}^n$

the proof relies on

Lemma

Suppose that $\mathbb{R}^n = \text{span}\{\Omega_1, \ldots, \Omega_K\}$ *where each* Ω_k *is a subspace and for each* $k = 1, \ldots, K$ *we have a*

$$W_k = \left[w_{k,1}^T, \ldots, w_{k,N_k}^T\right]^T \quad \text{and} \quad W = \left[W_1^T, \ldots, W_K^T\right]^T$$

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

the support of each of the elements of W_k must be contained in the corresponding Ω_k (W_k must be a block matrix w.r.t. a basis of Ω_k)

global inverse Lipschitz constant

by the piecewise linear nature of the ReLU operator, it is clear that

$$\|\mathrm{ReLU}(Wx_0) - \mathrm{ReLU}(Wx_1)\| \leq \|W\| \, \|x_0 - x_1\|$$

global inverse Lipschitz constant

by the piecewise linear nature of the ReLU operator, it is clear that

$$\|\mathrm{ReLU}(Wx_0) - \mathrm{ReLU}(Wx_1)\| \le \|W\| \, \|x_0 - x_1\|$$

inverse on the range

Theorem

Let $W \in \mathbb{R}^{m \times n}$ be a DSS w.r.t. every $x \in \mathbb{R}^n$. Then, for any $x_0, x_1 \in \mathbb{R}^n$,

$$\|\mathrm{ReLU}(Wx_0) - \mathrm{ReLU}(Wx_1)\|_2 \ge \left[\frac{1}{\sqrt{2m}} \min_{x \in \mathbb{R}^n} \sigma(W|_{S(x,W)}) \right] \|x_0 - x_1\|_2$$

where σ denotes the smallest singular value.

$$S(x, W) = \{ i \in [[m]] : \ \langle w_i, x \rangle \ge 0 \}$$

adjacent wedges

the proof relies on

Lemma

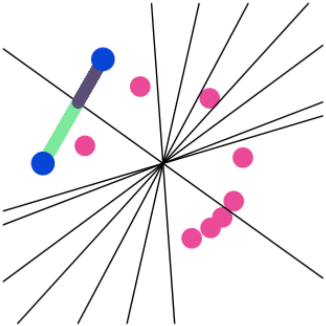
*Let $W \in \mathbb{R}^{m\times n}$ have a DSS w.r.t. every $x \in \mathbb{R}^n$ and $x_0, x_1 \in \mathbb{R}^n$. If x_0 and x_1 are in **adjacent wedges** and the line that connects them is **nonexceptional**, then*

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

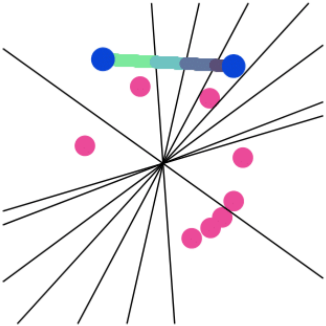
where $\sigma(M)$ is the smallest singular value of the matrix M.

line **nonexceptional** if it passes faces but not corners ($n \geq 3$)

adjacent and non-adjacent wedges



(a) two points that are in adjacent wedges



(b) two points that are in non-adjacent wedges

blue points are x_0 and x_1; pink points are elements (rows) of W

non-adjacent wedges, connected through faces

Lemma

*Let $W \in \mathbb{R}^{m\times n}$ be a DSS w.r.t. every $x \in \mathbb{R}^n$. Let x_0, x_1 be such that the line connecting them is **nonexceptional** and passes through $\mathcal{N} = \# S(x_0, W) \setminus S(x_1, W)$ points, then*

$$\|\mathrm{ReLU}(Wx_0) - \mathrm{ReLU}(Wx_1)\|_2 \geq \frac{1}{\sqrt{2\mathcal{N}}} \min_{t\in[0,1]} \sigma(W|_{S(\ell^{x_0,x_1}(t),W)}) \, \|x_0 - x_1\|_2 \, .$$

need to prove that any two points are ϵ close to two nonexceptional points

‘reconstruction’

linear programming, layerwise: let $y \in \mathrm{Range}\,(\mathrm{ReLU}(W \cdot + b))$, then the solution, x, of $\mathrm{ReLU}(Wx + b)) = y$ is given as the solution of

$$\operatorname*{argmin}_{x \in \mathbb{R}^n} \|(W|_{y>0} x + b|_{y>0}) - y|_{y>0}\|_2^2, \quad W|_{y \leq 0} x + b|_{y \leq 0} \leq 0$$

simplex method

> if the solution, x, is such that $\langle w_j, x \rangle \neq 0$, $j = 1, \ldots, m$, then the inequality constraint is unnecessary

(Lei, Jalal, Dhillon & Dimakis, ’19)

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a continuous function, where $m \geq 2n+1$. Then for any $\varepsilon > 0$ and compact subset $\mathcal{Z} \subset \mathbb{R}^n$ there exists a neural network $N_\theta \in \mathcal{NN}(n,m)$ of depth L such that $N_\theta : \mathbb{R}^n \to \mathbb{R}^m$ is injective and

$$|f(z) - N_\theta(z)| \leq \varepsilon, \quad \text{for all } z \in \mathcal{Z}.$$

- Lipschitz version of the generic orthogonal projector technique; this technique is used, for example, to prove the easy version of the Whitney's embedding theorem
- first approximate function f by a neural network and then apply to it a generic projection to make the neural network injective

controlling expansivity through random projections

if we daisy chain networks together we can control how expansive the final network is by introducing interstitial matrix multiplies, provided that the matrices are 'slightly' random

Corollary

Let $n, m, d_\ell \in \mathbb{Z}_+$, $\ell = 0, 1, \ldots, 2k$ be such that $d_0 = n$, $d_{2k} = m \geq 2n + 1$ and $d_{2j} \geq 2n + 1$ for $j \geq 1$. Let

$$F_k = B_k \circ f^{(k)} \circ B_{k-1} \circ f^{(k-1)} \circ \cdots \circ B_1 \circ f^{(1)}$$

where $f^{(j)} : \mathbb{R}^{d_{2j-2}} \to \mathbb{R}^{d_{2j-1}}$ are injective neural networks and $B_j : \mathbb{R}^{d_{2j-1}} \to \mathbb{R}^{d_{2j}}$ are random matrices whose joint distribution is absolutely continuous with respect to the Lebesgue measure of $\prod_{j=1}^{k}(\mathbb{R}^{d_{2j} \times d_{2j-1}})$. Then the neural network $F_k : \mathbb{R}^n \to \mathbb{R}^m$ is injective almost surely.

heuristics made rigorous

- $N : \mathcal{Z} \to N(\mathcal{Z}) \subset \mathcal{X}$ is a homeomorphism

heuristics made rigorous

- $N : \mathcal{Z} \to N(\mathcal{Z}) \subset \mathcal{X}$ is a homeomorphism
- N produces points that are on a topological (or Lipschitz-smooth) manifold
- if $\mathcal{Z}_1 \subset \mathcal{Z}$ then $N(\mathcal{Z}_1)$ has the same topology as $\mathcal{Z}_1$

outlook

verification of DSS
beyond layerwise viewpoint

outlook

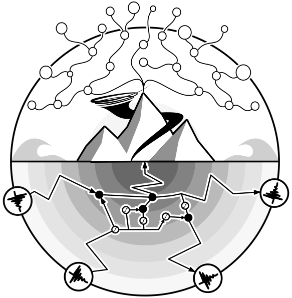
verification of DSS
beyond layerwise viewpoint

connection with inverse problems

- inference (networks)
- inductive bias (Kothari, dH & Dokmanić, ArXiv)
- learning/training dynamics

analysis and guarantees

classification, unsupervised



normalization of output layerwise

$$W_{L+1}\phi_L(W_L M_L(\ldots\phi_2(W_2 M_2(\phi_1(W_1 z + b_1)) + b_2)\ldots + b_L))$$

where $M_\ell \colon \mathbb{R}^{n_{\ell+1}} \to \mathbb{R}^{n_{\ell+1}}$, understood to be many-to-one

Definition (Scalar-Augmented Injective Normalization)

We say that $M_\ell(x) \colon \mathbb{R}^n \to \mathbb{R}^n$ is scalar-augmented injective if there exists a functions $m_\ell(x) \colon \mathbb{R}^n \to \mathbb{R}^k$ where $k \ll n$ and $\tilde{M}_\ell \colon \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ such that

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

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

normalization of output layerwise

$$W_{L+1}\phi_L(W_L M_L(\ldots\phi_2(W_2 M_2(\phi_1(W_1 z + b_1)) + b_2)\ldots + b_L))$$

where $M_\ell \colon \mathbb{R}^{n_{\ell+1}} \to \mathbb{R}^{n_{\ell+1}}$, understood to be many-to-one

Definition (Scalar-Augmented Injective Normalization)

We say that $M_\ell(x) \colon \mathbb{R}^n \to \mathbb{R}^n$ is scalar-augmented injective if there exists a functions $m_\ell(x) \colon \mathbb{R}^n \to \mathbb{R}^k$ where $k \ll n$ and $\tilde{M}_\ell \colon \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ such that

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

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

$$m_\ell(x) = \|x\|_2\,, \;\; \tilde{M}_\ell(x;\alpha) = \frac{x}{\alpha} \qquad M_\ell(x) = \frac{x}{\|x\|_2}$$

robustness to normalization

Lemma

Let N be a deep ReLU network that is layer-wise injective. Let the normalization functions $\{M_\ell\}_{\ell=1,\ldots,L}$ each be scalar-augmented injective. Then, given $\{m_\ell(x)\}_{\ell=1,\ldots,L}$, the network

$$\tilde{N}(z; m_1, \ldots, m_\ell) = W_{L+1}\phi_L(W_L\tilde{M}_L(\ldots \tilde{M}_2(\phi_1(W_1 z + b_1); m_1)\cdots + b_L; m_L))$$

is injective.

includes batch, weight normalizations

pooling