On the Invertibility of ReLU Networks

Inverse Problems and Machine Learning, Caltech

Jens Behrmann
joint work with: Sören Dittmer, Pascal Fernsel, Peter Maass

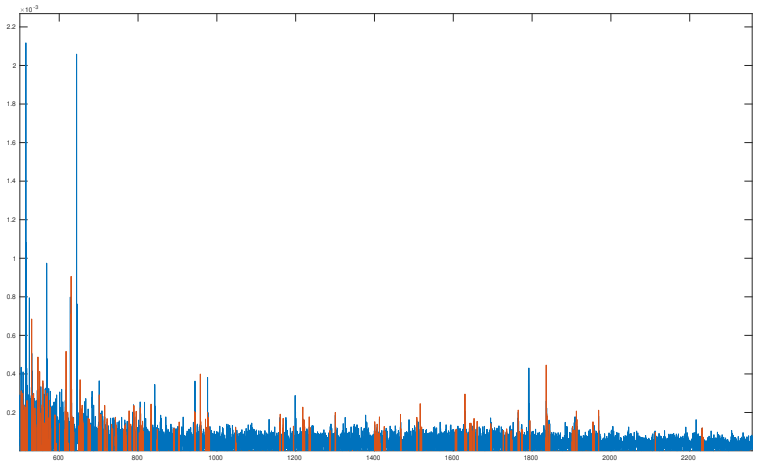
February 09. 2018

Motivation: Inverting a network

Reconstruct input x from features [1]

$$z^* \approx F(x),$$

$F : \mathbb{R}^d \to \mathbb{R}^D$, MLP or CNN
$x^* \in \mathbb{R}^d$ input
$z^* \in \mathbb{R}^D$ features, $z^* = F(x^*)$



Further applications:

- Inverse problems with learned forward operators
- Theoretical understanding
- ...

[1] Mahendran et al. 2015: Understanding deep image representations by inverting them

Main Questions

1. How is information lost during propagation?
   - Pre-images of ReLU layers
2. Is the inverse mapping stable/ instable?
   - Singular values of linearization

Related work:

- Invertibility via assumptions of random weights[2],[3]
- Injectivity and stability of ReLU and pooling[4]

[2] Giryes et al. 2016: DNN with Random Gaussian Weights: A Universal Classification Strategy?
[3] Arora et al. 2015: Why are deep nets reversible: a simple theory, with implications for training
[4] Bruna et al. 2014: Signal Recovery from Pooling Representations

Injectivity, Pre-images, Activation functions

- Combinatorial conditions for injectivity under ReLU [5]

Definition (Retrieval, singleton pre-images)

$A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.
Then, (A, b) does retrieval under ReLU for $x \in \mathbb{R}^n$ if the pre-image of $\mathrm{ReLU}(Ax + b)$ is a singleton.

Remark:

- Other activation functions like ELU, leakyReLU, tanh injective
- cReLU injective if A is frame [6]

[5] Bruna et al. 2014: Signal Recovery from Pooling Representations
[6] Shang et al. 2016: Understanding and Improving Convolutional Neural Networks via Concatenated Rectified Linear Units

Equality and Inequality Systems

$$A|_{y>0}x + b|_{y>0} = y|_{y>0}$$
$$A|_{y=0}x + b|_{y=0} \leq 0.$$

Consider the two cases

$$\mathcal{N}(A|_{y>0}) = \{0\} \text{ and } \mathcal{N}(A|_{y>0}) \neq \{0\}$$

$$A|_{y\leq 0}(P_{\mathcal{N}(A|_{y>0})^{\perp}}x + P_{\mathcal{N}(A|_{y>0})}x) + b|_{y\leq 0} \leq 0$$

Rewrite it into:

$$\overline{A}\overline{x} + \overline{b} \leq 0$$

Definition (Omnidirectional)

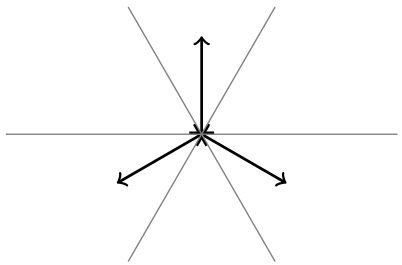
$A \in \mathbb{R}^{m \times n}$ is called omnidirectional if

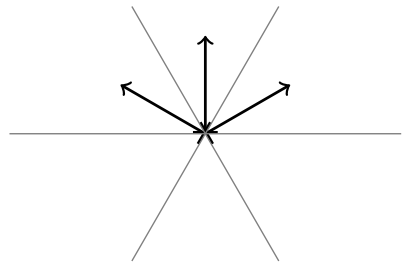
$$\exists! x : Ax \leq 0.$$

Corollary

The following statements are equivalent:

1. $A \in \mathbb{R}^{m \times n}$ *is omnidirectional.*
2. *Every linear open halfspace contains a row of* A.
3. $Ax \leq 0$ *implies* $x = 0$, *where* $x \in \mathbb{R}^n$.



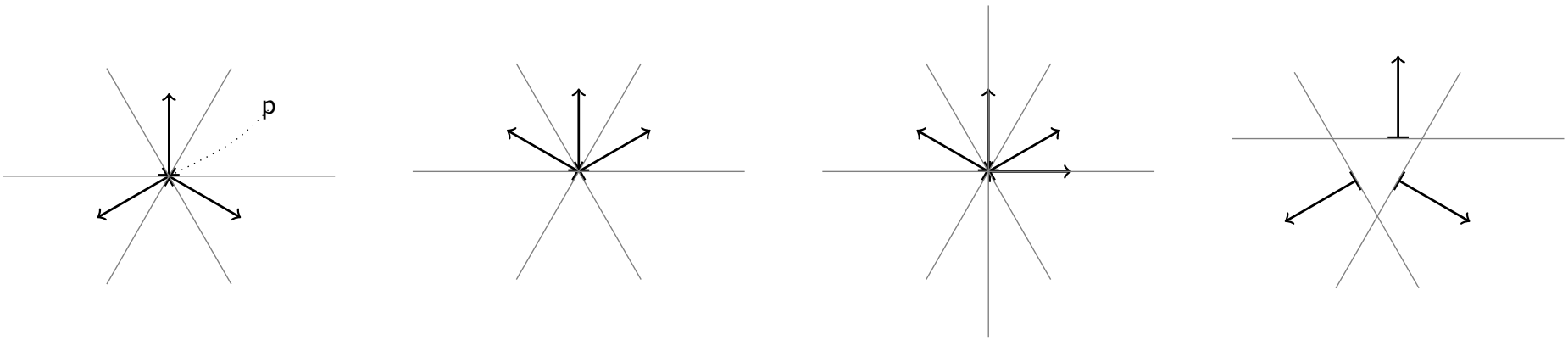


Definition (Omnidirectional for point)

- $A \in \mathbb{R}^{m \times n}$ is called omnidirectional if

$$\exists! x : Ax \leq 0.$$

- $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ is called omnidirectional for the point $p \in \mathbb{R}^n$ if $b = -Ap$ and A omnidirectional.



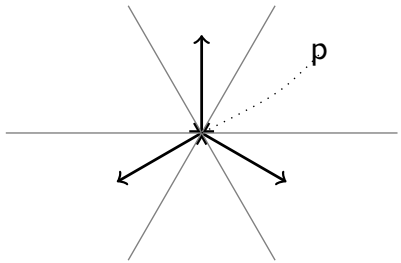
Theorem (Unique solutions of inequality system)

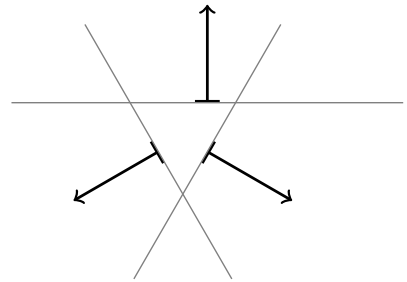
Let

$$\overline{A}\overline{x} + \overline{b} \leq 0$$

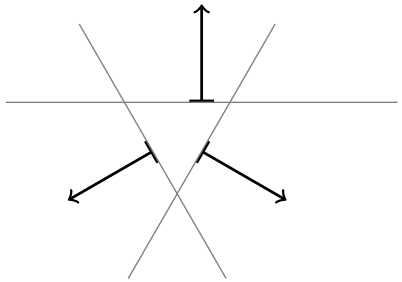
have a solution $\overline{x}_0$.
Then this solution is unique iff there exists an index set, I, *for the rows s.t.* $(\overline{A}|_I, \overline{b}|_I)$ *is omnidirectional for* x_0.

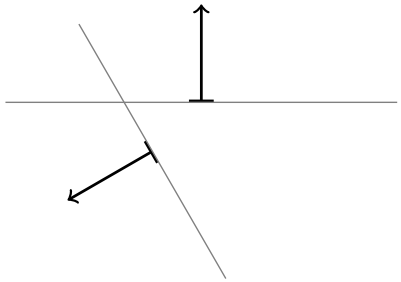
Realistic?



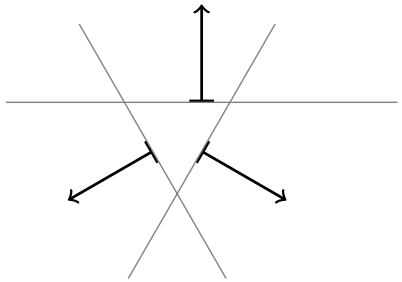


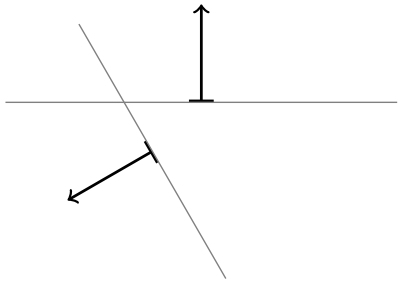
Pre-Image finite or infinite?





Pre-Image finite or infinite?





Theorem (Convex hull)

$A \in \mathbb{R}^{m \times n}$ is omnidirectional iff

$$0 \in \mathrm{Conv}(A)^{\mathrm{o}},$$

where $\mathrm{Conv}(A)^{\mathrm{o}}$ is the interior of the convex hull, spanned by the rows of A.

Singleton / Finite / Infinite

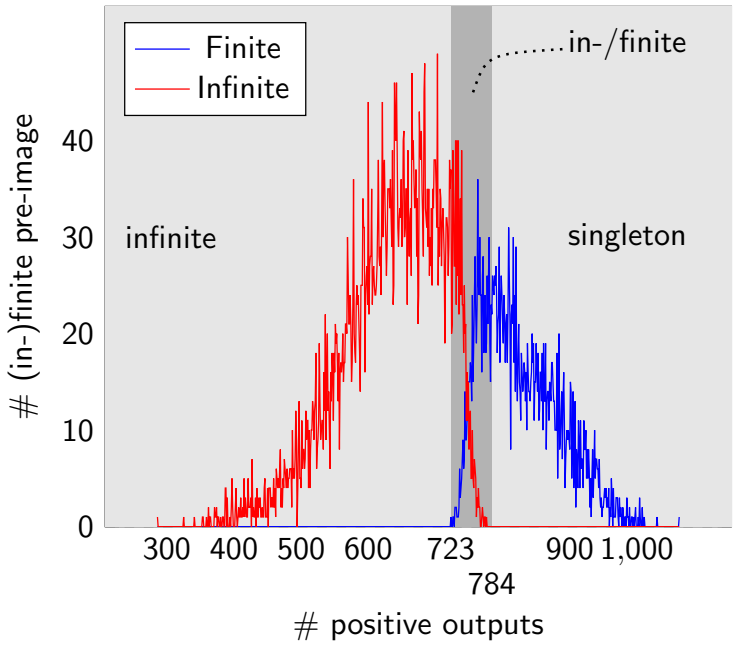
Setup: 2 layer MLP on MNIST, (3500, 784) neurons

1. Count number of positive outputs (> 784 singleton)
2. Projection onto Null-Space of equality system
3. Check for omnidirectionality via *linear programming* (convex hull as side-condition)

Singleton / Finite / Infinite

Setup: 2 layer MLP on MNIST, (3500, 784) neurons

1. Count number of positive outputs (> 784 singleton)
2. Projection onto Null-Space of equality system
3. Check for omnidirectionality via *linear programming* (convex hull as side-condition)

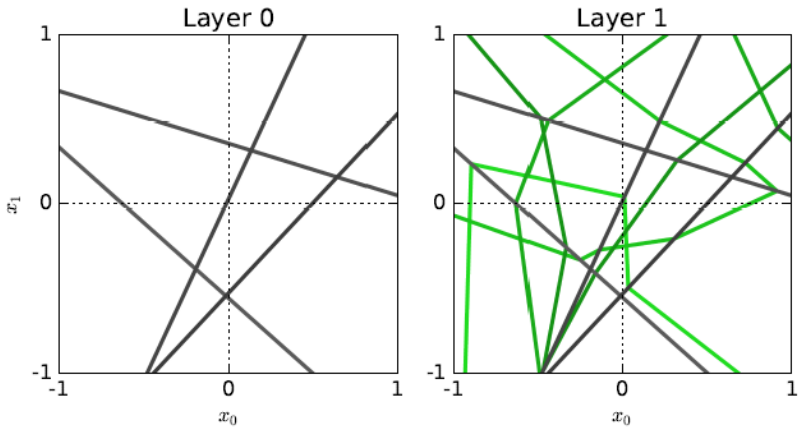


Stability - Locally Linear

Theorem (Linear functions on convex polytopes [7])

The input space $\mathbb{R}^d$ *of a* ReLU *network* F *is partitioned into convex polytopes* P_F*, where for* $P \in P_F$

$$F(x) = A_P x + b_P, \quad \forall x \in P. \tag{1}$$

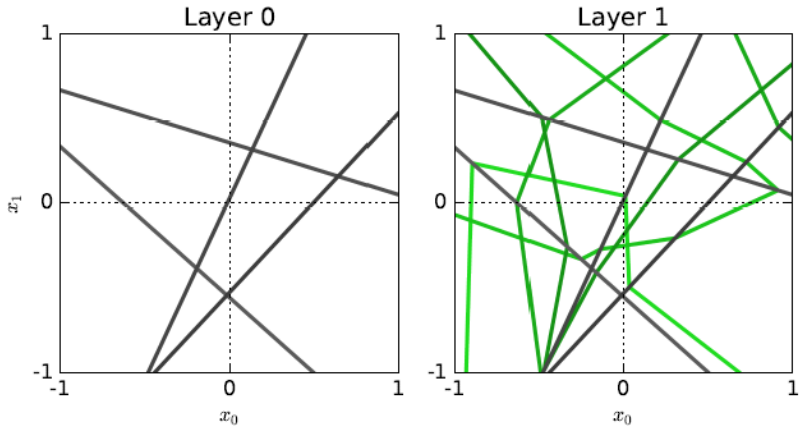


[7] Raghu et al. 2017: On the Expressive Power of Deep Neural Networks

Stability - Simplifications

- Assume: $x \in P$ known
  (for reconstruction of x given a output z^* of the network F)
- Analyze: Stability of linearization using singular values $\sigma_{min}, \sigma_{max}$:

$$\sigma_{min}\|x - x'\|_2 \leq \|A_P(x - x')\|_2 \leq \sigma_{max}\|x - x'\|_2, \quad x, x' \in P \cap \mathcal{N}(A_P)^{\perp}$$



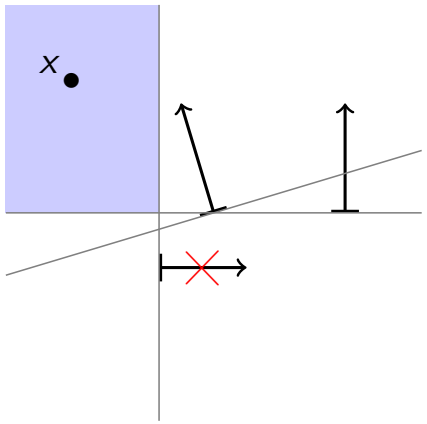
Source: Raghu et al. 2017: On the Expressive Power of Deep Neural Networks

Stability - ReLU as Diagonal Matrix

Linearization A_P of a network with L layers can be written as [8]

$$A_P = A^L D_{I^{L-1}} A^{L-1} \cdots D_{I^1} A^1, \qquad \text{where } D_{ii} = \begin{cases} 1, i \notin I \\ 0, i \in I \end{cases}.$$

$\rightarrow$ **Removal of rows due to** ReLU



[8] Wang et al. 2016: Analysis of deep neural networks with extended data jacobian matrix

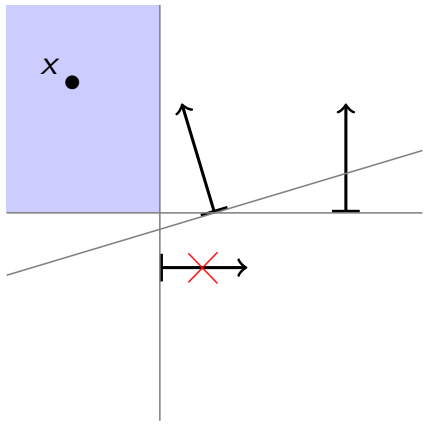
Lemma (Removal of weakly correlated rows)

$A \in \mathbb{R}^{m \times n}$ with rows a_j and $I \subseteq [m]$. For a fixed $k \in I$ let $a_k \in \mathcal{N}(D_I A)^{\perp}$. Moreover, let

$$\forall j \notin I : |\langle a_j, a_k \rangle| \leq c \frac{\|a_k\|_2}{\sqrt{M}},$$

where $M = m - |I|$ and constant $c > 0$. Then for the singular values $\sigma_I \neq 0$ of $D_I A$:

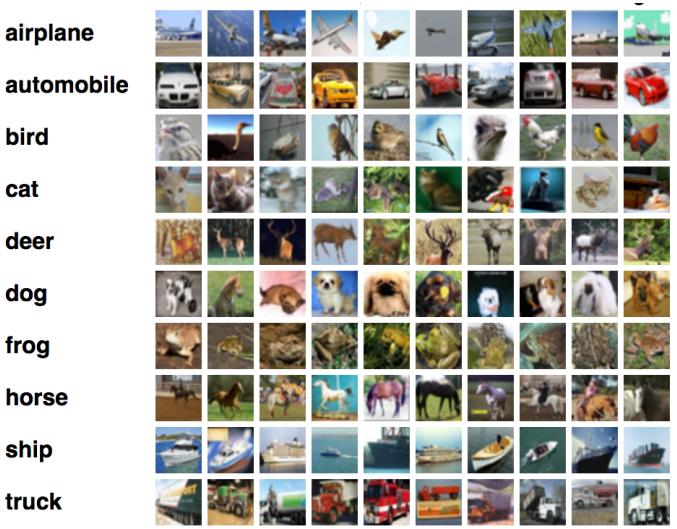
$$0 < \sigma_K = \min\{\sigma_I : \sigma_I \neq 0\} \leq c$$



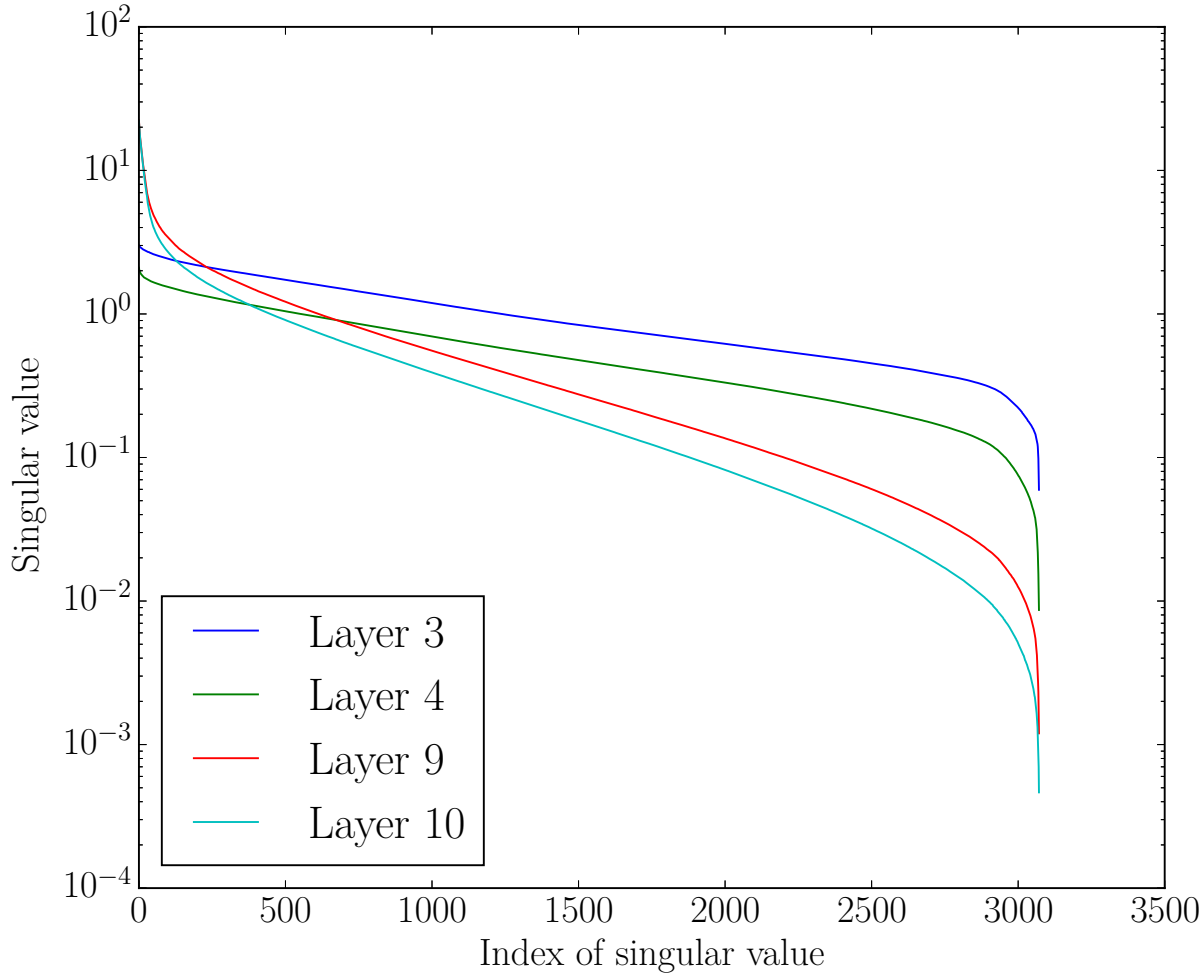
Numerical Experiments

- Convolutional Networks (CNN) fit the theoretical framework
- Linearization via backpropagation w.r.t. input
- Full SVD for different layers/ samples (nonlinear!)
- Small CNN on CIFAR10

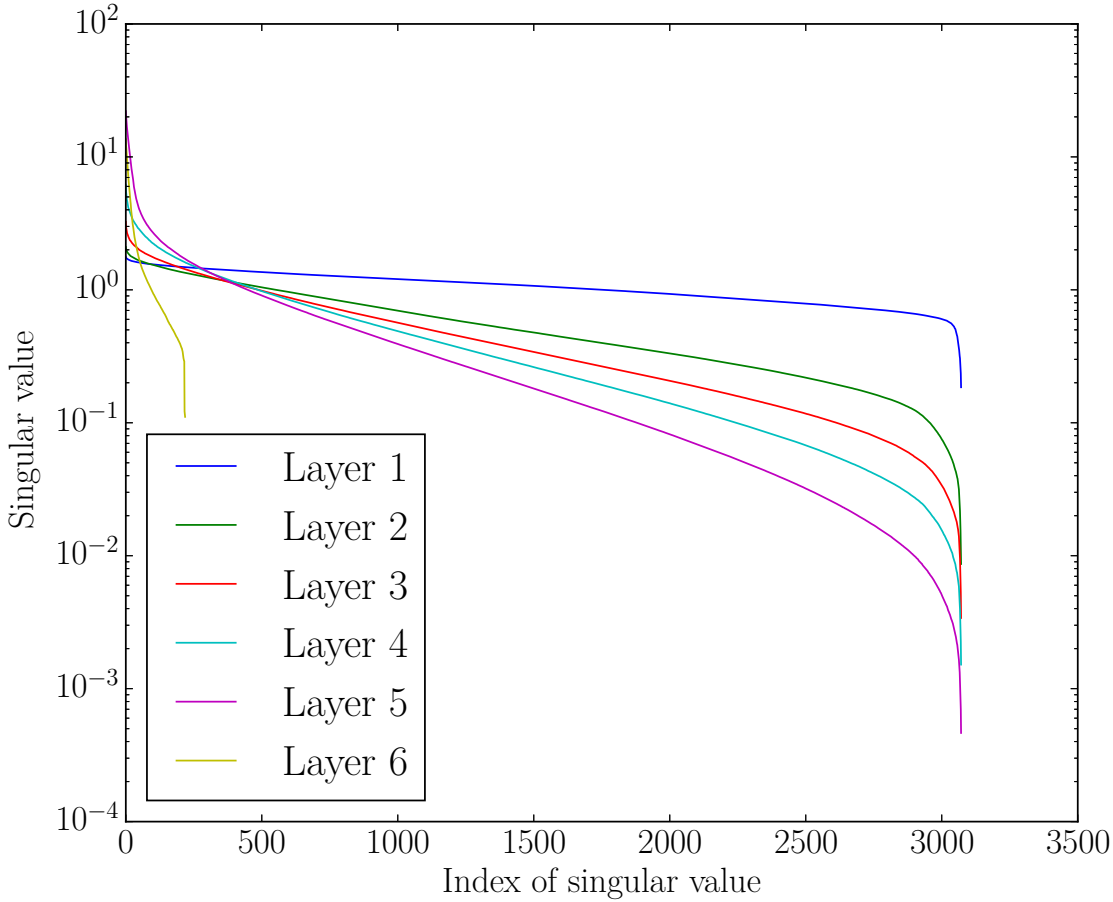
Type	kernel size	stride	# feature maps	# output units
Conv layer	(3,3)	(1,1)	32	-
Conv layer	(3,3)	(2,2)	64	-
Conv layer	(3,3)	(1,1)	64	-
Conv layer	(3,3)	(1,1)	32	-
Conv layer	(3,3)	(1,1)	32	-
Conv layer	(3,3)	(2,2)	64	-
Dense layer	-	-	-	512
Dense layer	-	-	-	10



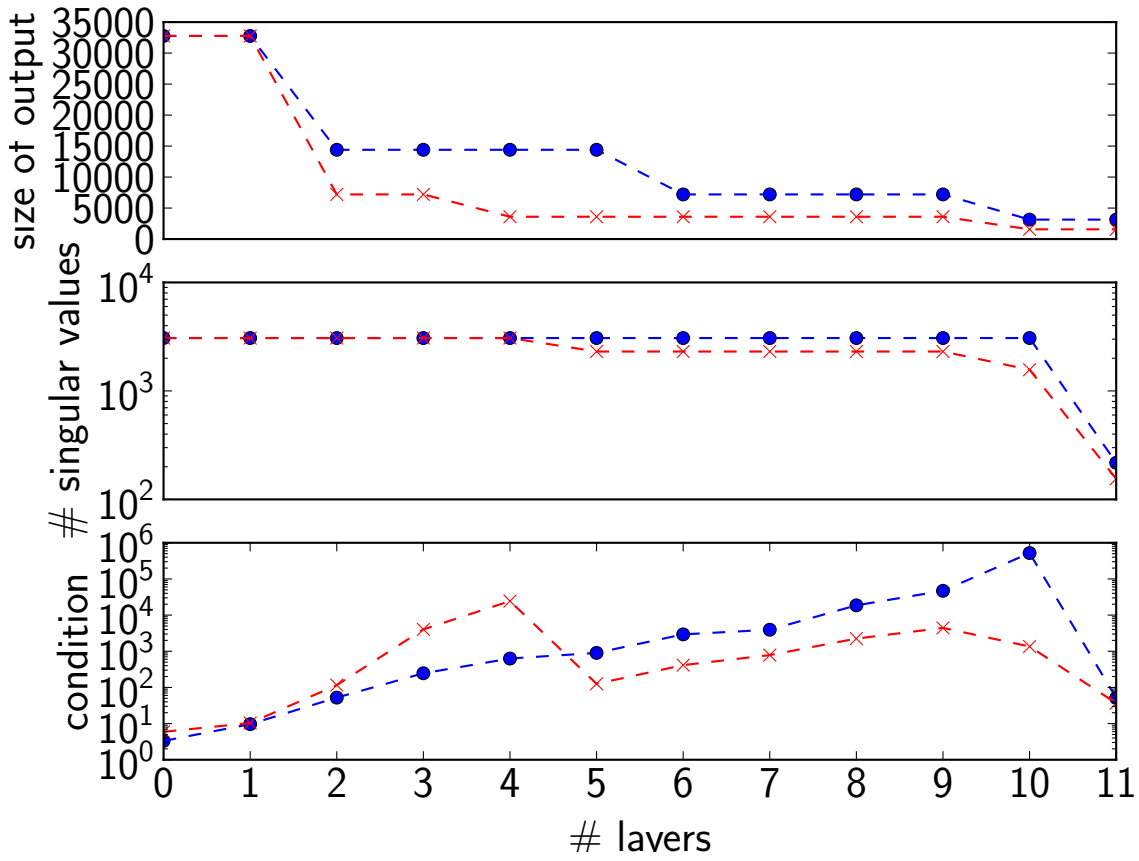
Effect of ReLU



Decay over Layers



Trade-off: Stability vs. Information Loss



Conclusion and Outlook

- Approach to better understand invertibility of deep ReLU networks
- Condition if pre-image of a layer is singleton, finite or infinite
- Stability analysis via SVD of linearization

Conclusion and Outlook

- Approach to better understand invertibility of deep ReLU networks
- Condition if pre-image of a layer is singleton, finite or infinite
- Stability analysis via SVD of linearization

Next steps:

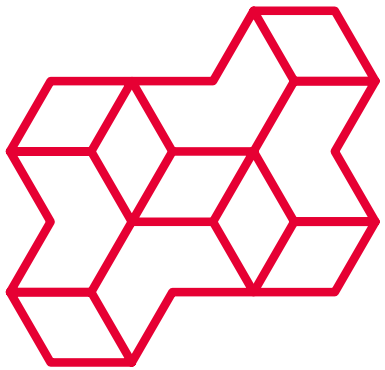
- Theory for CNNs, residual connections, ...
- Dropping linearity assumption
- Connection to stability analysis in context of adversarial examples
- ...

Thank you for your attention!

Joint work with:

- Sören Dittmer
- Pascal Fernsel
- Peter Maass

Universität Bremen



Boskamp, T. et al. (2016)A new classification method for MALDI imaging mass spectrometry data acquired on formalin-fixed paraffin-embedded tissue samples, *BBA-Proteins and Proteomics*

Behrmann,J. et al. (2017) Deep learning for tumor classification in imaging mass spectrometry, *Bioinformatics*

Bruna et al. (2014) Signal Recovery from Pooling Representations, *ICML*

Mahendran,A. ,Vedaldi,A. (2015) Understanding deep image representations by inverting them, *CVPR*

Raghu,M. et al. (2017) On the Expressive Power of Deep Neural Networks, *ICML*