INVARINCE AND INVERSE STABILITY UNDER ReLU

Anonymous authors
Paper under double-blind review

ABSTRACT

We flip the usual approach to study invariance and robustness of neural networks by considering the non-uniqueness and instability of the inverse mapping. We provide theoretical and numerical results on the inverse of ReLU-layers. First, we derive a necessary and sufficient condition on the existence of invariance that provides a geometric interpretation. Next, we move to robustness via analyzing local effects on the inverse. To conclude, we show how this reverse point of view not only provides insights into key effects, but also enables to view adversarial examples from different perspectives.

1 INTRODUCTION

Invariance and stability/robustness are two of the most important properties characterizing the behavior of a neural network. Due to growing requirements like robustness to adversarial examples (Szegedy et al., 2014) and the increasing use of deep learning in safety-critical applications, there has been a surge in interest in these properties. Invariance and stability are considered to be the key mechanisms in dealing with uninformative properties of the input (Achille & Soatto, 2017; Mallat, 2016) and are studied from the information theoretical perspective in form of the loss of information about the input (Tishby & Zaslavsky, 2015; Saxe et al., 2018).

Invariance and stability are also tightly linked to robustness against adversarial attacks (Cisse et al., 2017; Tsuzuku et al., 2018; Simon-Gabriel et al., 2018), generalization (Sokolic et al., 2017; Gouk et al., 2018) and even the training of Generative Adversarial Networks (Miyato et al., 2018). In general, stability is studied via two basic properties: 1) locally via a norm of the Jacobian (Sokolic et al., 2017; Simon-Gabriel et al., 2018), 2) globally via the Lipschitz constant (Cisse et al., 2017; Miyato et al., 2018; Tsuzuku et al., 2018). From a high-level perspective, both of these approaches study an upper bound on stability as the Lipschitz constant and a Jacobian norm quantifies the highest possible change under a perturbation with a given magnitude. We, unlike the approaches above, aim to broaden our understanding by analyzing the lowest possible change under a perturbation.

More formally, we study which perturbations $\Delta x$ do not (or only little) affect the outcome of a network $F$. Our analysis considers a given input data point $x$ and investigates the $\Delta x$’s, such that

$$F(x) = F(x + \Delta x) \quad \text{(invariant)} \quad \text{or} \quad \|F(x) - F(x + \Delta x)\| \leq \varepsilon \quad \text{(stable)},$$

where a small $\varepsilon > 0$ is given. While these properties can be crucial for many discriminative tasks (Mallat, 2016), the model could be flawed if perturbations that alter the semantics have only a minor impact on the features. This is a reverse perspective on adversarial examples (Szegedy et al., 2014), which commonly considers small input perturbations that lead to large changes and thus to arbitrary decisions of the network.

This flipped view and the study of smallest changes calls for a different approach: we study the instabilities of the inverse instead of the stabilities of the forward mapping. In particular, if $F$ is invariant to perturbations $\Delta x$, then $x$ and $x + \Delta x$ lie in the preimage of the output $z = F(x)$, i.e. $F$ is not uniquely invertible. Robustness towards large perturbations induces an instable inverse mapping as small changes in the output can be due to large changes in the input.

Based on the piecewise linear nature of ReLU networks (Montufar et al., 2014), we characterize the preimage of ReLU-activations as a single point, finite (bounded) or infinite (unbounded). Further, we study the stability of the linearization of rectifier networks via its singular values. To illustrate these locally changing properties and to demonstrate their tight connection, we visualize the behavior on a synthetic problem in Figure 1.
Figure 1: Left: Prediction of a small ReLU-network (one hidden layer with 3 neurons) trained to distinguish samples from two circles. Middle: Characterization of the preimage of first layer activations into unbounded (infinite), compact (finite) or unique (a single point). Right: Condition of the linearization of the first layer at each point.

As ReLU-layers are piecewise linear, the local behavior is constant on polytopes. Further, the regions with infinite/finite preimages correspond to regions with condition number of one or zero, while singleton preimages link to condition numbers larger than one. Thus, both properties are tightly connected and investigating one property alone yields an incomplete picture.

Our contributions are as follows:

- We derive conditions when the preimage of an output of a ReLU-layer has finite or infinite volume or is a single point. Based on these conditions, we derive an algorithm to check these conditions and exemplify its usability by applying it to investigate the preimages of a trained network. (See Section 2.)
- We study the stability of the inverse via analyzing the linearization at a point in input space, which is accurate within a polytope. We provide upper bounds on the smallest singular value of a linearization and prove how the removal of uncorrelated features could effect the stability of the inverse mapping. Based on these ideas, we experimentally demonstrate how singular values evolve over the different layers in rectifier networks. (See Section 3.)
- We introduce a reverse view on adversarial examples and connect it to invariance and robustness by leveraging our analysis of preimages. (see Section 5)

1.1 Related Work

While analyzing invariance and robustness properties is a major topic in theoretical treatments of deep networks (Mallat, 2016), studying it via the inverse is less common. Several works like Mahendran & Vedaldi (2015), Mahendran & Vedaldi (2016) or Dosovitskiy & Brox (2016) focus on reconstructing inputs from features of convolutional neural networks (CNNs) to visualize the information content of features. Instead, we investigate potential mechanisms affecting the invertibility.

Carlsson et al. (2017) gives a first geometrical view on the shape of preimages of outputs from ReLU layers, which is directly related to the question of injectivity of the mapping under ReLU. Shang et al. (2016) analyzes the reconstruction property of cReLU (concatenated ReLU); however, the more general situation of using the standard rectifier is not studied. A notable other line of work assumes random weights in order to derive guarantees for invertibility, see Gilbert et al. (2017) or Arora et al. (2015), whereas we focus on the preimage of ReLU-activations without assumptions on the weights.

Moreover, several reversible network structures were recently proposed (Gomez et al., 2017, Chang et al., 2018, Jacobsen et al., 2018). Most notably, in Jacobsen et al. (2018) a bijective network, up to its last layer, was trained successfully on ImageNet which does not exhibit any invariance. However, the network is very robust towards many directions in the input which is reflected in a strongly instable inverse. Hence, even carefully designed network show at least one of the two effects (invariance and robustness) studied in this work. Especially stability has seen growing interest due to adversarial examples (Szegedy et al., 2014), yet stability is mostly studied with respect to the forward mapping, see e.g. Cisse et al. (2017).
Two main resources for our view of rectifier networks as piecewise linear models are \cite{Montufar2014} and \cite{Raghu2017}. Closest to our approach is the work of \cite{Bruna2014} on global statements of injectivity and stability of a single layer including ReLU and pooling. The authors focus on global injectivity and stability bounds via combinatorial statements over all configurations attainable by ReLU and pooling. These conditions are valid on the entire input space, while the restriction to parts of the input space may be far from these worst-case conditions.

Further works focus on applications like inverse problems with learned forward models \cite{Jensen1999, Lu1999} and parameter estimation problems \cite{Lahivaara2018}, which are often formulated as inverse problems and require the inversion of networks.

1.2 Notation

In this section, we briefly state our notation as a reference:

- Input: $x = x^0 \in \mathbb{R}^{d_0}$, sometimes shortened to $d := d_0$.
- Pre-activations: $z^l = A^l x^{l-1} + b^l \in \mathbb{R}^{d_l}$, with weight matrix $A^l \in \mathbb{R}^{d_l \times d_{l-1}}$ and bias $b^l \in \mathbb{R}^{d_l}$.
- Entire network: $F : \mathbb{R}^d \ni x \mapsto F(x) := z := z^L \in \mathbb{R}^{d_L}$, sometimes short $D := d_L$.

For matrices $A \in \mathbb{R}^{m \times n}$ and $I \subset [m] := \{1, \ldots, m\}$, $A|_I$ denotes the matrix consisting of the rows of $A$ whose index is in set $I$ – analogously for vectors. Also $A|_{y > 0}$ describes the restriction to the index set $\{i : y_i > 0\}$ for $y \in \mathbb{R}^m$, analogously for $\prec, =, \preceq, \succeq$. For vectors $y \in \mathbb{R}^m$, $y > 0$ is the elementwise relation, analogously for $\prec, =, \preceq, \succeq$. Furthermore, we define $\mathcal{N}(A)$ as the null space of a matrix $A$. The Euclidean inner product is denoted by $\langle \cdot, \cdot \rangle$.

For every matrix $A \in \mathbb{R}^{m \times n}$ with the rows $a_i$, $i \in [m]$, we associate the set $A = \{a_i\}_{i=1}^m$. Vice versa, we associate every finite set in $\mathbb{R}^n$ with a matrix (only possible up to permutation of the indices).

2 Preimages of ReLU Layer

2.1 Theoretical Analysis

In this section, we analyze different kinds of preimages of a ReLU-layer and investigate under which conditions the inverse image of a given point is a singleton (a set containing exactly one element) or has finite/infinite volume. These conditions will yield a simple algorithm able to distinguish between these different preimages, which is applied in Section 2.2.

For the analysis of preimages of a given output one can study single layers separately or multiple layers at once. However, since the concatenation of two injective functions is again injective while a non-injective function followed by an injective function is non-injective, studying single layers is crucial. We therefore develop a theory for the case of single layers in this section. Notice that in case of multiple layers one is also required to investigate the image space of the previous layer.

We will focus our study on the most common activation function, ReLU. One of its key features is the non-injectivity, caused by the constant mapping on the negative half space. It provides neural networks with an efficient way to deploy invariances. Basically all other common activation functions are injective, which would lead to a straightforward analysis of the preimages. However, injective activations like ELU \cite{Clevert2016} and Leaky ReLU \cite{Maas2013} only swap the invariance for robustness, which in turn leads to the problem of having instable inverses. This question of stability will be analyzed in more detail in Section 3.

We start by introducing one of our main tools – namely the omnidirectionality.

Definition 1 (Omnidirectionality)

i) $A \in \mathbb{R}^{m \times n}$ is called omnidirectional if every linear open halfspace in $\mathbb{R}^n$ contains a row of $A$, i.e. for every given $x \in \mathbb{R}^n \setminus \{0\}$ there exists an index $i \in [m]$, such that $\langle a_i, x \rangle > 0$. 

3
where we know \( A, b \) for \( p \in \mathbb{R}^2 \), as hyperplanes intersect in \( p \) and normal vectors are omnidirectional. Two in the middle: Intersection in \( p \), but vector-free halfspaces (hence, not omnidirectional). Right: hyperplanes do not intersect in a point, but normal vectors are omnidirectional.

Thus, if \( A \) is omnidirectional, for every direction of a hyperplane through the origin forming two halfspaces, there is a vector from the rows of \( A \) inside each open halfspace, hence the term omnidirectional (see Figure 2 for an illustration). Note that the hyperplanes are due to ReLU as it maps the open halfspace to positive values and the closed halfspace to zero. A straightforward way to construct an omnidirectional matrix is by taking a matrix whose rows form a spanning set \( F \) and use the vertical concatenation of \( F \) and \( -F \). This idea is related to cReLU [Shang et al., 2016].

The following Corollary gives several equivalent formulations of omnidirectionality, which will turn out to be useful for the proof of the subsequent Theorem 4 in this section. The short proofs of the statements are provided in Appendix A1.

**Corollary 2 (Equivalences of omnidirectionality)** The following statements are equivalent:

i) \( A \in \mathbb{R}^{m \times n} \) is omnidirectional.

ii) \( Ax \preceq 0 \) implies \( x = 0 \), where \( x \in \mathbb{R}^n \).

iii) There exists a unique \( x \in \mathbb{R}^n \), such that \( Ax \preceq 0 \).

iv) There exist no \( x \in \mathbb{R}^n \setminus \{0\} \), such that \( Ax \preceq 0 \).

More importantly, omnidirectionality is directly related to the ReLU-layer preimages and will provide us with a method to characterize their volume (see Theorem A4). To analyze such inverse images, we consider \( y = \text{ReLU}(Ax + b) \) for a given output \( y \in \mathbb{R}^m \) with \( A \in \mathbb{R}^{m \times n}, \ b \in \mathbb{R}^n \) and \( x \in \mathbb{R}^n \). If we know \( A, b \) and \( y \), we can write the equation as the following mixed linear system:

\[
\begin{align*}
A|_{y>0}x + b|_{y>0} &= y|_{y>0} \quad \text{(1)} \\
A|_{y=0}x + b|_{y=0} &= 0 \quad \text{(2)}
\end{align*}
\]

where \( A|_{y>0} \) denotes the restriction of the matrix \( A \) to the rows, which are specified by the index set \( \{i : y_i > 0\} \) (see Section A.2 for the used notation).

**Remark 3** It is possible to enrich the mixed system to include conditions/priors on \( x \) (e.g. \( x \in \mathbb{R}^n_{\geq 0} \)).

The inequality system in equation (2) links its set of solutions and therefore the volume of the preimages of the ReLU-layer with the omnidirectionality of \( A \) and \( b \). Defining \( \overline{A} := AO^T \), where \( O \in \mathbb{R}^{k \times n} \) denotes an orthonormal basis of \( \mathcal{N}(A|_{y>0}) \) with \( k := \dim(\mathcal{N}(A|_{y>0})) \) and \( \overline{b} := b|_{y\leq 0} + A|_{y\leq 0}(P_V(A|_{y>0})^T x) \), where \( P_V \) denotes the orthogonal projection into the closed space \( V \), leads to the following main theorem of this section, which is proven in Appendix A1.

**Theorem 4 (Preimages of ReLU-layers)** Let \( \overline{A}, \overline{b} \) and \( k = \dim(\mathcal{N}(A|_{y>0})) \) be as above. The preimage of a point \( y \) under a ReLU-layer is

i) for \( k = 0 \) a singleton.

ii) for \( k > 0 \) a singleton, if and only if there exists an index set \( I \) for the rows of \( \overline{A} \) and \( \overline{b} \), such that \( (\overline{A}|_I, \overline{b}|_I) \) is omnidirectional for some point \( p \in \mathbb{R}^k \).
for \( k > 0 \) a compact polytope with finite volume, if and only if \( \bar{A} \) is omnidirectional.

Thus, omnidirectionality allows in theory to distinguish whether the inverse image of a ReLU-layer is a singleton, a compact polytope or has infinite volume. However, obtaining a method to check if a given matrix is omnidirectional is crucial for later numerical investigations. For this reason, we will go back to the geometrical perspective of omnidirectionality (see Figure 2). This will also help us to get a better intuition on the frequency of occurrence of the different preimages. The following Theorem gives another geometrical interpretation of omnidirectionality, whose short proof is given in Appendix A1.

**Theorem 5 (Convex hull)** A matrix \( A \in \mathbb{R}^{m \times n} \) is omnidirectional if and only if \( 0 \in \text{Conv}(A)^{\circ} \), where \( \text{Conv}(A)^{\circ} \) is the interior of the convex hull spanned by the rows of \( A \) (see Definition 10 in Appendix A1).

Therefore, the matrix must contain a simplex in order to be omnidirectional, as the convex hull of the matrix \( A \in \mathbb{R}^{m \times n} \) has to have an interior. Hence, we have the following:

**Corollary 6** If \( A \in \mathbb{R}^{m \times n} \) is omnidirectional, then \( m > n \).

By considering the geometric perspective, a tuple \((A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m)\) is omnidirectional for a point \( p \in \mathbb{R}^n \), if and only if the \( m \) hyperplanes generated by the rows of \( A \) with bias \( b \) intersect at \( p \) and their normal vectors (rows of \( A \)) form an omnidirectional set. We can use Corollary 6 to conclude that singleton preimages of ReLU-layers are very unlikely to happen in practice (if we do not design for it), since a necessary condition is that \( n + 1 \) hyperplanes have to intersect in one point in \( \mathbb{R}^n \). Therefore we conclude, that singleton preimages of ReLU layers in practice only and exclusively occurs, if the mixed linear system already has sufficient linear equalities.

**Algorithm to check uniqueness:** The above results can be used to derive an algorithm to check whether a preimage of a given output is finite, infinite or just a singleton. A singleton inverse image is obtained as long as \( \text{rank}(A|_{y \succ 0}) = n \) holds true, which can be easily computed. To distinguish preimages with finite and infinite volumes, it is enough to check if \( \bar{A} \) is omnidirectional (see Theorem 4iii), which can be done numerically by using the definition of the convex hull, Theorem 5 and Corollary 6. This leads to a linear programming problem, which is presented in Appendix A3 and was also used to create Figure 1.

2.2 Numerical Analysis

In this section, we demonstrate for a simple model that the preimage of a layer can be a singleton, infinite or finite depending on the given point. For this purpose, we trained a MLP with two hidden ReLU layers of size 3500 and 784 on MNIST [LeCun & Cortes, 2010]. We chose the layer size of 3500, because the likelihood of having roughly 784 (input dimension of MNIST) positive outputs was high for this setting. In Figure 3 we plotted the number of samples in the test set that have infinite (red curve) or finite (blue curve) preimages over the number of positive outputs. It can be assumed that all samples which have more or equal to 784 (the input dimension) positive outputs have a singleton preimage and are therefore finite. In the dark gray region between 723 and 784, both effects occurred, which can be seen by the overlap of the red and blue curve.

To determine whether a preimage for less than 784 positive outputs was compact we used Theorem 5iii and the algorithm described in Appendix A3.

3 Stability

3.1 Theoretical Analysis

In this section we analyze the robustness of rectifier MLPs against large perturbations via studying the stability of the inverse mapping. Concretely, we study the effect of ReLU on the singular values of the linearization of network \( F \). While the linearization of a network \( F \) at some point \( x \) only provides a first impression on its global stability properties, the linearization of ReLU networks is exact in some neighborhood due to its piecewise-linear nature [Raghu et al., 2017]. In particular, the
Let $\sigma_l$ be the singular values of $D_lA$. Then for all admissible index sets $I$, the smallest non-zero singular value is upper bounded by $\min \{ \sigma_1 : \sigma_1 > 0 \} \leq \tilde{\sigma}_k$, where $k = N - |I|$ and $\tilde{\sigma}_1 \geq \ldots \geq \tilde{\sigma}_N > 0$ are the non-zero singular values of $A$.

Furthermore, the largest singular value is upper bounded by $\max \{ \sigma_1 : \sigma_1 > 0 \} \leq \tilde{\sigma}_1$. 

Lemma 8 analyzes the best case scenario with respect to the highest value of the smallest singular value. While this would yield a more stable inverse mapping, one needs to keep in mind that $N(A_P)$ grows by the corresponding elimination of rows via $D_I$. Moreover, reaching this bound is very
unlikely as it requires the singular vectors to perfectly align with the directions that collapse due to $D_i$. Thus, we now turn to study effects which could happen locally for some input polytopes $P$.

An example of a drastic effect through the application of ReLU is depicted in Figure 3. Since one vector is only weakly correlated to the removed vector and the situation is overdetermined, removing this feature for some inputs $x$ in the blue area leaves over the strongly correlated features. While the two singular values of the 3-vectors-system were close to one, the singular vectors after the removal of ReLU are badly ill-conditioned. As many modern deep networks increase the dimension in the first layers, redundant situations as in Figure 3 are common, which are inherently vulnerable to such phenomena. For example, Rodriguez et al. (2017) proposes a regularizer to avoid such strongly correlated features. The following lemma formalizes the situation exemplified before:

**Lemma 9 (Removal of weakly correlated rows)** Let $A \in \mathbb{R}^{m \times n}$ 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

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

with $M = m - |I|$ and constant $c > 0$. Then for the singular values $\sigma_i \neq 0$ of $D_I A$ it holds

$$0 < \sigma_k = \min \{ \sigma_i : \sigma_i \neq 0 \} \leq c.$$

Note that $I$ has to be admissible when considering the effect of ReLU.

**Effect under multiple layers:** For the effect of ReLU applied to multiple layers, we are particularly interested in following questions:

- Can the application of another layer have a pre-conditioning effect yielding a stable inverse?
- What happens when we only compose orthogonal matrices which have stable inverses?

Note that a way to enforce an approximate orthogonality constraint was proposed for CNNs in Cisse et al. (2017), however only for the filters of the convolution. For both situations the answer is similar: the nonlinear nature of ReLU induces locally different effects. Thus, if we choose a pre-conditioner $A^l$ for a specific matrix $A^{l-1}_P$, it might not stabilize the matrix product for matrices $A^{l-1}_P$ corresponding to different input polytopes $P^*$. For the case of composing only orthogonal matrices, consider a network up to layer $l - 1$, where the linearization $A^{l-1}_P$ has orthogonal columns (assume the network gets larger, thus $A^{l-1}_P$ has more rows than columns). Then, the application of ReLU in form of $A^l D_l A^{l-1}_P$ removes the orthogonality property of the rows of $A^{l-1}_P$, if setting entries in the rows from $I^l$ to zero results in non-orthogonal columns (likely when considering dense matrices). Hence, $D_l A^{l-1}_P$ is not orthogonal for some $I^l$. In this case, the matrix product $A^l D_l A^{l-1}_P$ is not orthogonal, which results in decaying singular values.

This is why, even when especially designing the network by e.g. orthogonal matrices, stability issues with respect to the inverse arise. To conclude this section, we remark that the presented results are rather of a *qualitative* nature showcasing effects of ReLU on the singular values. Yet, the analysis does not require any assumptions and is thus valid for any MLP (including CNNs without pooling). To give an idea of *quantitative* effects we study numerical examples in the subsequent subsection.

### 3.2 Numerical Analysis

In this section, we show how the previously discussed theoretical stability properties can be examined for a given network. In particular, we conduct experiments on CIFAR10 (Krizhevsky & Hinton, 2009) using two baseline CNNs, see [A4] for details on architectures and training setup. Our CNNs use only strides instead of pooling and use no residual connections and normalization layers. Thus,
the architectures fit to the theoretical study as the strided discrete convolution can be written as a matrix-vector multiplication.

**Singular values over multiple layers:** Experimentally most interesting is the development of singular values over multiple layers as several effects are potentially at interplay. Figure 6 shows how all singular values evolve in convolutional layers (layers 1-6, after application of ReLU). While the shape of the curve is similar for layer 1-5, it can be seen that the largest singular value grows, while the small singular values decrease significantly. Note that this growth of the largest singular values is in line with observations for adversarial examples, see Szegedy et al. (2014). While many defense strategies like Cisse et al. (2017) or Jia (2017) focus on the largest singular value, the behavior of the smaller singular values is often overlooked. Additionally, we provide in Appendix A5 a numerical analysis of the condition from Lemma 9 to gain an understanding of possible effects of ReLU on the singular values. Furthermore, we add results for a thinner CNN (ThinCIFAR) and for the MLP from Section 2.2 in Appendix A6.

**Relationship between stability and invariance:** While invariance is characterized by zero singular values, the condition number only takes non-zero singular values into account, see e.g. layer 6 from Figure 6. This tight relationship is further investigated in Figure 5 which compares the output size, the condition number and the number of non-zero singular values vs. the layers for WideCIFAR and ThinCIFAR. In combination with lower output dimension, ReLU has a different effect for ThinCIFAR. The number of singular values decreases in layer 5, which cuts off the smallest singular values, resulting in a lower condition number. Yet, there are more invariance directions within the corresponding linear region. For a visual comparison see Figure 1.

**Computational costs and scaling analysis:** First, we remark that the linearization of a network \( F \) for an input point \( x^0 \) can be computed via backpropagation. Based on this linearization the computation of the full SVD scales cubically. Especially, early CNN-layers have high dimensional outputs which may cause memory issues when computing the entire SVD. We thus choose a small CNN trained on CIFAR10 as these inputs are only of size \( 32 \times 32 \times 3 \). To scale this analysis up to e.g. ImageNet with VGG-networks, a restriction to a window of the input image is necessary to reduce the complexity of the full SVD especially for early layers. See Jacobson et al. (2018), where the singular values restricted to input windows were used to estimate the stability of the entire i-RevNet trained on ImageNet.
4 Scope

Characterization of preimages over multiple layers: Theorem 4 yields a characterization (singleton, finite, infinite) of the preimage of a point \( y \) under a single ReLU-layer. When considering the preimage of \( y \) under multiple layers \( l \) to \( l - i \), two difficulties arise: 1) If the preimage of \( y \) under layer \( l \) is not a singleton, one needs to compute the intersection of the image of layer \( l - 1 \) and the preimage of \( y \) under layer \( l - 1 \). 2) For all points in the intersection, the conditions of Theorem 4 need to be checked, which requires the solution of a linear program, see Algorithm 1 in Appendix A3. Hence, our analysis is currently restricted to a layer-by-layer approach. However, this layer-wise and local study could enable to pin down the specific layers where information, which is expressed in non-singleton preimages.

Preimages for convolutional layers: In general, convolutional layers are affine transforms \( Ax + b \), but have a sparse and shared structure compared to dense matrices used in MLPs. Thus, the trivial case \( \text{rank}(A|_{y>0}) = n \) (Algorithm 1, Appendix A3) needs to be explicitly checked. For MLPs it was assumed in section 2.2 that \( \text{rank}(A|_{y>0}) = |\{i : y_i > 0\}| \) as dense matrix rows are almost surely linear independent in practice.

Inverse stability for convolutional networks: For inverse stability, we consider the linearization \( A_P = A^L D_{L-1} \cdots D_1 A^1 \) for an input polytope \( P \). In convolutional networks, each \( A^l \) implements a multi-channel discrete convolution. While singular values of each \( A^l \) can be efficiently computed by leveraging the convolutional structure, see Sedghi et al. (2018), the shared structure is not preserved in the matrix chain \( A_P \) due to the application of ReLU (expressed via \( D_P \)). Thus, a tighter analysis that leverages the convolutional structure in \( A^l \), compared to our general assumption that \( A^l \) can be any linear mapping, is not straightforward with current tools but would certainly lead to further insights.

Extension of inverse stability across polytopes: In our stability analysis, we employ a piecewise-linear viewpoint which allows to characterize stability via the singular values of the linearization which is exact within an input polytope. However, when considering an \( \varepsilon \)-ball \( B_{\varepsilon}(y) \) around a point \( y = A_P x + b \), to model e.g. reconstruction from noisy activations \( y \), further questions arise: 1) Can all points in \( B_{\varepsilon}(y) \) be reached by an \( x^\ast \) from the polytope \( P \)? 2) Are points from another polytope \( P' \) mapping to points in \( B_{\varepsilon}(y) \)? In this case, the inverse stability needs to be augmented by nonlinear considerations to model movements between piecewise-linear regions.

5 Practical Implications

While the focus of this work was an in-depth analysis of potential effects on crucial properties like invariance and robustness due to ReLU, we envision several practical implications of our approach:

Network design and regularization: As both the concept of omnidirectionality and removal of rows due to ReLU showed, there is a breadth of potential effects. In terms of network design, controlling such effects could be desirable. In particular, a change to injective activation functions (tanh, leakyReLU, ELU etc.) remove the discussed preimages, but immediately transfer to an instable inverse due to saturation. Furthermore, Lemma 9 draws a connection to regularizing correlation between feature maps as introduced in Rodríguez et al. (2017). Hence, both omnidirectionality and correlation between rows can be thought of as geometrical properties which could partially be controlled by regularization or architecture design. Furthermore, the analysis also shows the difficulty of controlling these properties in vanilla architectures. However, by incorporating additional structure like dimension splitting in reversible networks (Jacobsen et al. 2018) or invertible residual connections (Behrmann et al. 2018), the preimage is by design a singleton.

Connection to information loss: Our analysis is tightly related to mutual information \( I(x^i; x) \) loss, which has gained growing interest due to the information bottleneck (Tishby & Zaslavsky 2015; Sax et al. 2018). In particular, invariance in layer \( l \) may induce \( I(x^i; x) \leq I(x^{l-1}; x) \) due to the data processing inequality (Cover & Thomas 2006). Similarly, an instable inverse can induce an information loss as activations \( x^i \) are quantized due to finite precision on hardware.

Implications for adversarial examples: Despite being crucial for many discriminative tasks to contract the space along uninformative directions (Mallat 2016), invariance and robustness may induce severe vulnerabilities for adversarial examples (Szegedy et al. 2014). For instance, a model
Figure 7: Invariances of the first layer (100 ReLU neurons) of a vanilla multilayer perceptron (MLP). Despite the semantically very different examples, the features are identical as the original image "3" and the two perturbed variants "6" and "4" are in the same preimage. Further details in Appendix A7.

would be flawed if perturbations that alter the semantics only have a minor impact on the features of the network. Classically, adversarial examples are viewed as small perturbations which induce large changes in the network outputs (Goodfellow et al., 2014). Yet, reversing this perspective leads to another failure case: if large changes in the input alter its semantics for a given task, but the networks output is robust or even invariant to such changes, the model might just be as flawed from this reverse point of view.

This change in perspective leads to a natural way of addressing invariance and robustness via invertibility: If $F$ is invariant to perturbations $\Delta x$, then $x$ and $x + \Delta x$ lie in the preimage of the output $z = F(x)$ i.e. $F$ is not uniquely invertible. Robustness towards large perturbations induces an instable inverse mapping as small changes in the output can be due to large changes in the input.

Finally Figure 7 demonstrates such a striking failure, where perturbations alter the semantics drastically, yet the activations even after the first layer are identical. To find these examples, we leveraged the developed theory about preimages and a linear programming formulation, see Appendix A7.

6 CONCLUSION AND OUTLOOK

We presented the inverse as an approach to tackle the invariance and robustness properties of ReLU networks. Particularly, we studied two main effects: 1) conditions under which the preimage of a ReLU layer is a point, finite or infinite and 2) how ReLU can effect the inverse stability of the linearization. By deriving approaches to numerically examine these effects, we highlighted the broad range of possible effects. Moreover, controlling such properties may be desirable as our experiment on adversarial examples showed.

Besides the open questions on how to control the structure of preimages and inverse stability via architecture design or regularization, we envision several theoretical directions based on our work. Especially, incorporate nonlinear effects like moving between linear regions of rectifier networks could lift the analysis closer to practice. Furthermore, studying similarities of omnidirectionality as a geometrical property and singular values could further strengthen the link between these two crucial properties.

REFERENCES


François Chollet et al. Keras. https://github.com/keras-team/keras 2015.


A1 APPENDIX FOR SECTION[2]

Proof (Corollary[2] Equivalences of omnidirectionality) We show the equivalences by proving \(i) \Rightarrow ii) \Rightarrow iii) \Rightarrow iv) \Rightarrow i)\).

Let \(A \in \mathbb{R}^{m \times n}\) be omnidirectional, i.e. for every \(x \neq 0\), it holds that \(Ax \not\leq 0\). This is equivalent to

\[Ax \leq 0 \Rightarrow x = 0,\]

which is \(ii)\). The implications from \(ii)\) to \(iii)\) and from \(iii)\) to \(iv)\) are obvious. From \(iv)\), we have that

\[\forall x \neq 0 \exists i \in \{1, \ldots, m\} : (x, a_i) > 0,\]

which is equivalent to \(i)\), the omnidirectionality of \(A\). Altogether, this shows the equivalence of all four points. \(\square\)

Definition 10 (Convex hull) For \(A \in \mathbb{R}^{m \times n}\), the convex hull is defined as

\[\text{Conv}(A) = \left\{ \sum_{i=1}^{m} \alpha_i a_i : \forall i \in \mathbb{R}_{\geq 0} \land \sum_i \alpha_i = 1 \right\},\]

where \(a_i \in \mathbb{R}^n\) are the rows of \(A\).

Theorem 11 (Stiemke’s theorem, see[1963]) Let \(A \in \mathbb{R}^{m \times n}\) be a matrix, then the following two expressions are equivalent.

- \(\exists \not\exists) y : Ay \not\geq 0\)
- \(\exists x \geq 0 : A^T x = 0\)

Here \(z \not\geq 0\) means that \(0 \neq z \leq 0\).

Theorem 12 (Singleton solutions of inequality systems) Let \(A \in \mathbb{R}^{m \times n}\), \(b \in \mathbb{R}^m\) and \(x \in \mathbb{R}^n\).

Furthermore, let the inequality system

\[Ax + b \leq 0,\]

written as \((A, b)\), have a solution \(x_0\). Then this solution is unique if and only if there exists an index set, \(I\), for the rows s.t. \((A|_I, b|_I)\) is omnidirectional for \(x_0\).

Proof (Theorem[12] Singleton solutions of inequality systems)

\(\leq\)

Let \((A|_I, b|_I)\) be omnidirectional for \(x_0\). Then it holds that \(A|_Ix + b|_I = A|_I(x - x_0) \leq 0\). Due to the omnidirectionality of \(A|_I, x_0\) is the unique solution of the inequality system \(A|_Ix + b|_I \leq 0\). The existence of a solution for the whole system \(Ax + b \leq 0\) is guaranteed by assumption and therefore \(x_0\) is the unique solution of \(Ax + b \leq 0\).

\(\Rightarrow\)

Here we will prove

\[\exists! I : (A|_I, b|_I) \text{ omnidirectional for some } p \Rightarrow \text{ solution non-unique}\].

We will start by doing the following logical transformations:

\[\exists! I : (A|_I, b|_I) \text{ omnidirectional for some } p\]
\[\iff \exists(I, p) : (A|_I \text{ omnidirectional} \land b|_I = -A|_Ip)\]
\[\iff \forall(I, p) : -A|_I \text{ omnidirectional} \land b|_I = -A|_Ip\]
\[\iff \forall(I, p) : (A|_I \text{ not omnidirectional} \lor b|_I \neq -A|_Ip).\]

Now we define the vector \(c_0 := Ax_0 + b \leq 0\) and the set \(I\) as the index set given via \(c_0 = 0\).

This means that \(A|_I\) is not omnidirectional, because otherwise \(A|_Ix_0 + b|_I = 0\) due to the definition of \(I\), which would lead to the contradiction that \((A|_I, b|_I)\) is omnidirectional for \(x_0\). But this means \(\exists x' \neq 0 : A|_Ix' \leq 0\) as a result of Corollary[2]. Since \(A|_Ix_0 + b|_I < 0\), we also have
\[ \forall x \exists \epsilon > 0 : \left| A \cdot (x_0 + \epsilon x) + b \right| < 0. \]
This holds in particular for \( x' \), so we define accordingly
\[ x^* := \epsilon x' \neq 0. \]
Therefore, we have \( A \cdot (x_0 + x^*) + b \epsilon < 0 \) as well as
\[ A \cdot (x_0 + x^*) + b \epsilon \leq 0. \]

Altogether it holds that \( A(x_0 + x^*) + b \leq 0 \) with \( x^* \neq 0 \), which means that \( x_0 \) is a non-unique solution for the inequality system \( Ax + b \leq 0 \).

**Proof (Theorem 4, Preimages of ReLU-layers)** We consider the ReLU-layer
\[ y = \text{ReLU}(Ax + b), \]
given its output \( y \in \mathbb{R}^m \) with \( A \in \mathbb{R}^{m \times n} \), \( b \in \mathbb{R}^m \) and \( x \in \mathbb{R}^n \). Clearly, this equation can also be written as the mixed linear system
\[ \begin{align*}
A_{|y > 0}x + b_{|y > 0} &= y_{|y > 0}, \\
A_{|y = 0}x + b_{|y = 0} &\leq 0.
\end{align*} \]

This allows us to consider the two cases
\[ \mathcal{N}(A_{|y > 0}) = \{0\} \text{ and } \mathcal{N}(A_{|y > 0}) \neq \{0\}. \]

In the first case, we have a linear system which allows us to calculate \( x \) uniquely, i.e. we can do retrieval. This leads us to the second case, the interesting one. In this case we can only recover \( x \) uniquely if and only if the system of inequalities “pins down” \( P_{\mathcal{N}(A_{|y > 0})}x \), where \( P_V \) is the orthogonal projection into the closed space \( V \). Formally this requires
\[ A_{|y \leq 0}(P_{\mathcal{N}(A_{|y > 0})} + x + P_{\mathcal{N}(A_{|y > 0})}x) + b_{|y \leq 0} \leq 0 \]
to have a unique solution for \( x \in \mathbb{R}^n \) and \( P_{\mathcal{N}(A_{|y > 0})}x \) fixed (given via the equality system). By defining \( \bar{b} := b_{|y \leq 0} + A_{|y \leq 0}(P_{\mathcal{N}(A_{|y > 0})}x) \) we have
\[ A_{|y \leq 0}(P_{\mathcal{N}(A_{|y > 0})}x) + \bar{b} \leq 0. \]

If \( O \in \mathbb{R}^{k \times n} \) now denotes an orthonormal basis of \( \mathcal{N}(A_{|y > 0}) \), where \( k := \dim(\mathcal{N}(A_{|y > 0})) \), we can write
\[ \bar{\mathbf{A}}\mathbf{p} + \bar{b} \leq 0, \]
where \( \bar{\mathbf{A}} := AO^T \) and \( \mathbf{p} := Ox \) is a general element in \( \mathbb{R}^k \). It now follows from Theorem 12 that the inequality system \( (\bar{\mathbf{A}}, \bar{b}) \) has a unique solution if and only if \( (\bar{\mathbf{A}}, \bar{b}) \) has a subset of rows that are omnidirectional for some point \( p \).

**Proof (Theorem 5, Convex hull)** Since \( \mathcal{N}(A) = \{0\} \) follows from both sides of the equivalence, the following sequence of equivalencies holds.
\[ 0 \in \text{Conv}(A)^o \iff \exists x > 0 : A^T x = 0 \quad \text{Theorem 13} \]
\[ \exists y : Ay \leq 0. \]
Together with \( \mathcal{N}(A) = \{0\} \), which means that \( \exists y \neq 0 : Ay = 0 \), leads altogether to
\[ \exists y \neq 0 : Ay \leq 0. \]

**A2 PROOFS FOR SECTION 5**

**Proof (Lemma 5, Global upper bound for largest and smallest singular value)** The upper bound on the largest singular value is trivial, as ReLU is contractive or in other terms \( \|D_I Ax\|_2 \leq \|Ax\|_2 \) for all \( I \) and \( x \in \mathbb{R}^n \).

To prove the upper bound for the smallest singular value, we assume
\[ \sigma_M := \min \{\sigma_l : \sigma_l > 0\} > \tilde{\sigma}_k \quad \text{(4)} \]
and aim to produce a contradiction. Consider all singular vectors \( \tilde{v}_k \) with \( k \geq k^* \) from matrix \( A \). It holds for all \( \tilde{v}_k \)
\[ \tilde{\sigma}_k \geq \tilde{\sigma}_k^* = \|A\tilde{v}_k^*\|_2 \geq \|D_I A\tilde{v}_k^*\|_2, \quad \text{(5)} \]
as \( D_I \) is a projection matrix and thus only contracting. As
\[ \sigma_M = \min_{\|x\|_2=1} \|D_I Ax\|_2, \]

\[ x \in \mathcal{N}(D_I A)^o \]
all $\tilde{v}_{k^*} \not\in \mathcal{N}(D_I A)^\perp$. Otherwise, a $\tilde{v}_{k^*}$ would be a minimizer by estimation which would violate the assumption.

Due to $\mathcal{N}(D_I A)^\perp \oplus \mathcal{N}(D_I A) = \mathbb{R}^n$, it holds $\tilde{v}_{k^*} \in \mathcal{N}(D_I A)$. As $\tilde{v}_{k^*}$ are orthogonal, dim($\text{span}(v_{k^*})$) = $|I| + 1$ (note: $k^* = k, ..., N$ and $k = N - |I|$, thus there are $|I| + 1$ singular vectors $v_{k^*}$ in total). Furthermore, $\tilde{v}_{k^*}$ were not in $\mathcal{N}(A)$ by definition (corresponding singular values were strictly positive).

Hence, the nullspace of $D_I$ must have dim($\mathcal{N}(D_I)$) $\geq |I| + 1$. But $D_I$ is the identity matrix except $|I|$ zeros on the diagonal, thus dim($\mathcal{N}(D_I)$) = $|I|$, which yields a contradiction. □

Proof (Lemma 9, Removal of weakly correlated rows) Consider $v = \frac{a_k}{\|a_k\|_2}$. Then,

$$(D_I A v)_k = 0,$$

since $k \in I$ ($k$-th row of $D_I$ is zero). Furthermore, for all $j \neq k$ it holds by condition

$$(D_I A v)_j = \frac{(a_k, a_j)}{\|a_k\|_2} \leq \frac{|(a_k, a_j)|}{\|a_k\|_2} \leq \frac{c}{\sqrt{M}}.$$

Hence,

$$\|D_I A v\|_2 = \sqrt{\sum_{j \notin I} \left(\frac{(a_k, a_j)}{\|a_k\|_2}\right)^2} \leq \sqrt{M \left(\frac{c}{\sqrt{M}}\right)^2} = c.$$

As $a_k \in \mathcal{N}(D_I A)^\perp$, $v \in \mathcal{N}(D_I A)^\perp$ as well. Thus,

$$\sigma_K = \min_{\|x\|_2 \leq \|D_I A\|_2} \|D_I A\|_2 \leq \|D_I A v\|_2 \leq c.$$

□
A3 Appendix for Section 2.2

In this section, we formulate the algorithm to determine whether the preimage of $y$ given by

$$y = \text{ReLU}(Ax + b)$$

is finite.
This requires to check whether $\overline{A}$ (see Theorem 4) is omnidirectional, which is equivalent to

$$0 \in \text{Conv}(\overline{A})^\circ,$$

see Theorem 5. Since it is reasonable to assume that $0$ will not lie on the boundary of the convex hull, we can formulate this as a linear programming problem. The side-conditions incorporate the definition of convex hulls (Definition 10, Appendix A1). The objective function is chosen arbitrary, as we are only interested in a solution.

**Algorithm 1** Finite preimage

**Input:** $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $y \in \mathbb{R}^m$

```markdown
if $\text{rank}(A|_y > 0) = n$ then
    return True {Preimage is a singleton}
end if
```

$$O \leftarrow \text{orthonormal basis of } \mathcal{N}(A|_y > 0), \ (\in \mathbb{R}^{k \times n})$$

$$\overline{A} \leftarrow A|_y = 0 O^T, \ (\in \mathbb{R}^{k \times k})$$

```markdown
if $k \leq k$ then
    return False \ {see Corollary [6]}
end if
```

$$c \leftarrow (1; \ldots; 1) \ \{\text{arbitrary objective}\}$$

```markdown
return Does a solution for the linear program

$$\begin{align*}
\max c^T x \\
\text{subject to} \\
\overline{A}^T x &= 0 \\
(1; \ldots; 1)^T x &= 1 \\
x &\in [0, 1]^k
\end{align*}$$
```

A4 Architectures for Numerical Studies

Training details for MLP on MNIST:

- Training using Adam optimizer (Kingma & Ba 2015)
- Epochs: 25
- Batch size: 1000

Training details for WideCIFAR and ThinCIFAR:

- Training setup from Keras (Chollet et al. 2015) examples: cifar10_cnn
- No data augmentation
- RMSprop optimizer
- Epochs: 100
- Batch size: 32
Table 1: Architecture of MLP trained on MNIST

<table>
<thead>
<tr>
<th>Index</th>
<th>Type</th>
<th>kernel size</th>
<th>stride</th>
<th># feature maps</th>
<th># output units</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Input layer</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>11</td>
<td>Dense layer (softmax)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2: Architecture of MLP trained on MNIST

<table>
<thead>
<tr>
<th>Index</th>
<th>Type</th>
<th>kernel size</th>
<th>stride</th>
<th># feature maps</th>
<th># output units</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Input layer</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3500</td>
</tr>
<tr>
<td>2</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>784</td>
</tr>
<tr>
<td>3</td>
<td>Dense layer (softmax)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 3: Architecture of WideCIFAR

<table>
<thead>
<tr>
<th>Index</th>
<th>Type</th>
<th>kernel size</th>
<th>stride</th>
<th># feature maps</th>
<th># output units</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Input layer</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>Convolutional layer</td>
<td>(3,3)</td>
<td>(1,1)</td>
<td>32</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Convolutional layer</td>
<td>(3,3)</td>
<td>(2,2)</td>
<td>64</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>Convolutional layer</td>
<td>(3,3)</td>
<td>(1,1)</td>
<td>64</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>Convolutional layer</td>
<td>(3,3)</td>
<td>(1,1)</td>
<td>32</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>Convolutional layer</td>
<td>(3,3)</td>
<td>(1,1)</td>
<td>32</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>Convolutional layer</td>
<td>(3,3)</td>
<td>(2,2)</td>
<td>64</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>512</td>
</tr>
<tr>
<td>8</td>
<td>Dense layer (softmax)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 4: Architecture of ThinCIFAR

<table>
<thead>
<tr>
<th>Index</th>
<th>Type</th>
<th>kernel size</th>
<th>stride</th>
<th># feature maps</th>
<th># output units</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Input layer</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>Convolutional layer</td>
<td>(3,3)</td>
<td>(1,1)</td>
<td>32</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Convolutional layer</td>
<td>(3,3)</td>
<td>(2,2)</td>
<td>32</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>Convolutional layer</td>
<td>(3,3)</td>
<td>(1,1)</td>
<td>16</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>Convolutional layer</td>
<td>(3,3)</td>
<td>(1,1)</td>
<td>16</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>Convolutional layer</td>
<td>(3,3)</td>
<td>(1,1)</td>
<td>16</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>Convolutional layer</td>
<td>(3,3)</td>
<td>(2,2)</td>
<td>32</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>Dense layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>512</td>
</tr>
<tr>
<td>8</td>
<td>Dense layer (softmax)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10</td>
</tr>
</tbody>
</table>

A5 Effect of ReLU Numerical Analysis of Lemma 9

In order to better understand the bound on the smallest singular value after ReLU, given by Lemma 9, we numerically proceed as follows:
1. We choose $c \in [a, b]$, where $a, b$ are suitable interval endpoints.

2. Given $c$, we compute for every $a_k$ with $k \in I$ the value of $c \frac{\|a_k\|_2}{\sqrt{M}}$ ($M$ is the number of remaining rows, in the example $M = 5120$).

3. For every $a_k$ we count the number of $a_i$ satisfying

$$|\langle a_i, a_k \rangle| \leq c \frac{\|a_k\|_2}{\sqrt{M}}.$$

4. We take the $a_k$ with the maximal number of $a_i$ satisfying the condition. (Note, that this ignores the requirement $a_k \in \mathcal{N}(D_1 A)^\perp$.)

5. If we have an $a_k$, where all $a_i$ satisfy the condition, the corresponding constant $c$ gives the upper bound on the smallest singular value after ReLU.

Figure 9 shows the number of $a_i$ satisfying the correlation condition given different choices of $c$. The red line is reached for $c \approx 6$. However, even the largest singular value after ReLU is smaller than 2.5 (shown in Figure 8). Thus, the bound given by Lemma 9 is far off. This can be explained by the
fact, that this situation is quite redundant ($M = 5120$) and there are rows $a_i$ still correlated to the removed rows $a_k$.

However, in the further Experiments on ThinCIFAR, we observe (see Figure A6) a stronger effect of ReLU in layer 2, which can be explained by having a less redundant scenario with fewer remaining rows.
A6 Further Experiments

Figure 10: Left: Effect of ReLU on the singular values for ThinCifar. The curves show the effect in layer 2 (layer 3 and 4 in legend, because ReLU is counted as an extra activation layer) and layer 5 (layer 9 and 10). Right: Decay of singular values over the layers ThinCifar. Here, each layer includes the convolution and ReLU-activation. Reported number are taken from median over 50 samples. Best viewed in color.

Figure 11: Left: Effect of ReLU on the singular values for the MLP on MNIST. The curves show the effect in layer 1 (layer 1 and 2 in legend, because ReLU is counted as an extra activation layer) and layer 2 (layer 3 and 4). Right: Decay of singular values over the layers of MLP on MNIST. Here, each layer includes the fully-connected layer and ReLU-activation. Reported number are taken from median over 50 samples.

A7 Invariance Experiment using an MLP on MNIST

This section briefly describes how the results in Figure 7 from the introduction were obtained (copied in Figure 11 for readability). After training the network from 1 (in Appendix A4), we searched the MNIST test set for input images with yield the fewest positive activations in the first layer, in the figure the digits ”3” and ”4”. After selecting the example input $x^*$, we selected another input $c$ belonging to a different class (e.g. a ”6” and ”4” in the first example).
Figure 12: Invariances of the first layer (100 ReLU neurons) of a vanilla MLP. (Exact architecture in Appendix [A4 Table 1])

Afterwards, we solved following linear programming problem to find a perturbed $x$:

$$\begin{cases}
\max \langle c, x \rangle \\
\text{subject to}
\end{cases}
\begin{align*}
A|_{y^*\succ0} x + b|_{y^*\succ0} &= y^*|_{y^*\succ0} \\
A|_{y^*\prec0} x + b|_{y^*\prec0} &\preceq 0 \\
x &\in [0, 1]^k
\end{align*}
$$

where the features of the first layer are computed via

$$y^* = \text{ReLU}(Ax^* + b).$$

Hence, we searched within the preimage of the features $y^*$ of the first layer for examples $x$ which resemble images $c$ from another class. By doing this we observe, that the preimages of the MLP may have large volume. In these cases, the network is invariant to some semantics changes which shows how the study of preimages can reveal previously unknown properties.