AI Pictures at a Mathematical Exhibition: How Applied Harmonic Analysis meets Machine Learning

Radu Balan

Department of Mathematics and Norbert Wiener Center for Harmonic Analysis and Applications
University of Maryland, College Park, MD

June 28, 2023
University of Torino, Turin, Italy

Acknowledgments

This material is based upon work partially supported by the National Science Foundation under grant no. DMS-2108900 and Simons Foundation. "Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation."

Papers available online at:
https://www.math.umd.edu/ rvbalan/

Table of Contents:

Table of Contents

High-Level Overview

In this series of lectures, we discuss a few harmonic analysis techniques and problems applied to machine learning.

1. NN: Neural networks (NN) and their universal approximation property.
2. Lipschitz analysis: we provide rationals for studying Lipschitz properties of NNs, and then we perform a Lipschitz analysis of these networks. We focus on two aspects of this analysis: stochastic modelng of local vs. global analysis, and a scattering network inspired Lipschitz analysis of convolutive networks.
3. Invariance and Equivariance: We highlight the duality between invariance and covariance/equivariance, with focus on G-invariant representations.
4. Applications to data analysis and modeling: We present applications on a variety of problems: classification and regression on graphs; generative models for data sets; neural network based modeling of time-evolution of dynamical systems; discrete optimizatons.

Table of Contents

Neural Networks: Architectures and Properties

Neural networks were introduced a long time ago ...

1. 1925: Ising model – first Recurrent Neural Network (RNN)
2. 1940s: Hebbian learning for neuroplasticity – weights are learned dynamically
3. 1958: Rosenblatt introduced the perceptron, a 1-layer NN
4. 1965: Ivakhnenko and Lapa: Multi-Layer Perceptron (MLP)
5. 1967: Amari studied stochastic gradient descent (SGD) for training/learning
6. 1980: Fukushima introduced the convolutional neural network (CNN)
7. 1991-2: Schmidhuber introduced adversarial networks (precursors of GANs - 2014 by Goodfellow), generative models, and the transformers with linearized self-attention

Network Architectures

Deep Neural Networks

- Input layer: $x = (x_1, x_2, \cdots, x_n)^T$
- Output layer: $y = (y_1, y_2, \cdots, y_m)^T$
- Number of Layers: L

$$y = A_{L+1} \cdot \sigma(A_L \cdot \sigma(A_{L-1} \cdots \sigma(A_1 \cdot x + b_1) \cdots) + b_{L-1}) + b_L) + b_{L+1}$$

The scalar *activation function* $\sigma' : \mathbb{R} \to \mathbb{R}$ acts entrywise.

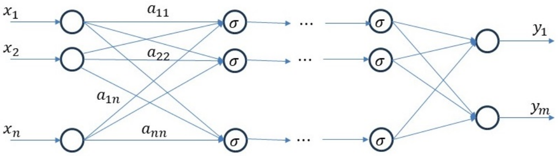


Figure: A general Feed-Forward Network, or a Deep Neural Network (DNN)

Network Architectures

Convolutive Neural Networks (CNN)

A Convolutive Neural Network is a Deep Neural Network with two additional features:

1. Linear operators A_k are convolutive operators, and implemented as convolutions
2. Activation functions are followed by downsampling and (optional) *pooling layers*: either max-pooling or sum-pooling.

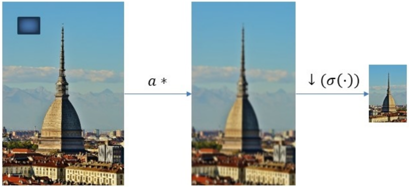


Figure: One layerr of a Convolutive Neural Network (picture curtesy of robygarba@pixabay)

Convolutive Neural Networks (CNN)

Alex Net

The AlexNet is 8 layer network, 5 convolutive layers plus 3 dense layers. Introduced by (Alex) Krizhevsky, Sutskever and Hinton in 2012 .

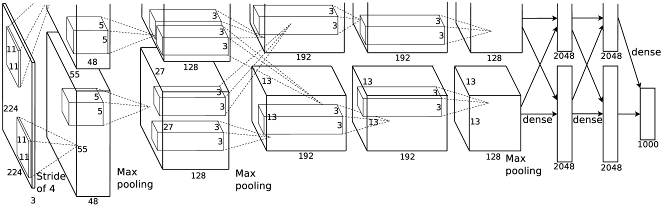


Figure: From Krizhevsky et all 2012 : AlexNet: 5 convolutive layers + 3 dense layers. Input size: 224x224x3 pixels. Output size: 1000.

Universal Approximation Properties of Neural Netwoks

Conventional wisdom says that neural networks can approximate arbitrary well any "reasonable" function $f : \mathbb{R}^n \to \mathbb{R}^m$.
Earliest results showed that even one hidden layer networks approximate target functions equally well. One hidden layer networks are called *perceptrons*. The input-output characterization of a perceptron $\Phi : \mathbb{R}^n \to \mathbb{R}$, is given by:

$$\Phi(x) = a^T \sigma(Wx + b) + b_0 \quad , \quad x \mapsto \Phi(x) = \sum_{k=1}^{p} a_k \sigma(\sum_{j=1}^{n} W_{k,j} x_j + b_k) + b_0.$$

Universal Approximation Properties of Neural Netwoks

Conventional wisdom says that neural networks can approximate arbitrary well any “reasonable” function $f : \mathbb{R}^n \to \mathbb{R}^m$.
Earliest results showed that even one hidden layer networks approximate target functions equally well. One hidden layer networks are called *perceptrons*. The input-output characterization of a perceptron $\Phi : \mathbb{R}^n \to \mathbb{R}$, is given by:

$$\Phi(x) = a^T \sigma(Wx + b) + b_0 \quad , \quad x \mapsto \Phi(x) = \sum_{k=1}^{p} a_k \sigma(\sum_{j=1}^{n} W_{k,j} x_j + b_k) + b_0.$$

Theorem (Cybenko 1989)

Assume $\sigma : \mathbb{R} \to \mathbb{R}$ is a bounded continuous function that satisfies $\lim_{t\to\infty} \sigma(t) = 1$ and $\lim_{t\to-\infty} \sigma(t) = 0$. Then the span of the set of functions $\{\sigma(w^T x + b) \ , \ w \in \mathbb{R}^n \ , \ b \in \mathbb{R}\}$ is dense in $C([0,1]^n)$.

Proof of Cybenko's Universal Approximation Theorem

Proof

The proof is by contradiction. Denote by $K=[0,1]^n$ the compact unit cube. Assume $V=span\{\sigma(w^Tx+b)\ \ ,\ \ w\in\mathbb{R}^n\ ,\ b\in\mathbb{R}\}$ is not dense in $C(K)$. Then its closure is a proper subspace of $C(K)$, and by Riesz representation theorem, there exists a signed, finite Borel measure μ over $[0,1]^n$ so that

$$\int_K \sigma(w^Tx+b)d\mu(x)=0\ \ ,\ \ \forall w\in\mathbb{R}^n\forall b\in\mathbb{R}.$$

We shall prove that $\sigma\in L^\infty(\mathbb{R})$ satisfying $\sigma(t)\overset{t\to\infty}{\longrightarrow}1$ and $\sigma(t)\overset{t\to-\infty}{\longrightarrow}0$ implies $\mu=0$. For $\lambda,b,\theta\in\mathbb{R}$ and $w\in\mathbb{R}^n$, let

$$\phi_\lambda(x)=\sigma(\lambda(w^Tx+b)+\theta)=\sigma((\lambda w)^Tx+(\lambda b+\theta))$$

Proof of Cybenko's Universal Approximation Theorem (cont'ed)

Notice:

$$\lim_{\lambda\to\infty}\phi_\lambda(x) = \begin{cases} 1 & if \quad w^Tx+b>0 \\ \sigma(\theta) & if \quad w^T+b=0 \\ 0 & if \quad w^Tx+b<0 \end{cases}$$

Let $\Pi_{w,b}=\{x,\ w^Tx+b=0\}$ denote a hyperplane, and $H_{w,b}=\{x\ ,\ w^T+b>0\}$ denote a half-space. Then by Lebesgue's dominated convergence theorem (even the simpler form, Lebesgue bounded convergence theorem),

$$0=\lim_{\lambda\to\infty}\int_K \phi_\lambda(x)d\mu(x)=\sigma(\theta)\mu(\Pi_{w,b})+\mu(H_{w,b})$$

Since σ takes at least two distinct values, we obtain $\mu(\Pi_{w,b})=0$ and $\mu(H_{w,b})=0$, for all $w\in\mathbb{R}^n$ and $b\in\mathbb{R}$.

Proof of Cybenko's Universal Approximation Theorem (cont'ed)

Construct the linear functional $h \in Ł^{\infty}(\mathbb{R}) \mapsto F(h) = \int_K h(w^T x) d\mu(x)$. It follows that, for any interval $I \subset \mathbb{R}$ (either open, closed, bounded or not), $F(1_I) = 0$, where 1_I is the indicator function of I. Linear combinations of indicator functions are weak dense in $L^{\infty}(\mathbb{R})$. Hence $F(h) =$ over $Ł^{\infty}(\mathbb{R})$. In particular, for $h(t) = cos(2\pi t)$ and $h(t) = sin(2\pi t)$, and choosing $w = m \in \mathbb{Z}^n$, it follows

$$0 = \int_K cos(2\pi m^T x) + isin(2\pi m^T x) d\mu(x) = \int_K e^{2\pi i \langle m,x \rangle} d\mu(x) = \hat{\mu}(m).$$

Thus all Fourier coefficients of μ are 0, from where we conclude $\mu = 0$. Contradiction!
Hence $V = span\{\sigma(w^T x + b) \ , \ w \in \mathbb{R}^n \ , \ b \in \mathbb{R}\}$ is dense in $C(K)$.
Q.E.D.

Further Results

Remark

The compact set $[0,1]^n$ can be replaced by any compact set K: scale and translate to bring it inside $[0,1]^n$; then use Tietze extension theorem.

Remark

Recent results extend the density result to various other spaces, such as $C^k(K)$, $W^{k,p}(K)$, etc; they also extend to the case of certain unbounded σ, e.g., the ReLU function, $ReLU(x) = x1_{(0,\infty)}$.

Remark

Cybenko's proof (or several subsequent results) is not constructive. Recent results by other researchers (e.g., Petersen and Voigtlaender; Bolcskei, Grohs, Kutyniok and Petersen) provide explicit architectures (number of layers, number of hidden nodes) and even memory cost (i.e., quantized weights) that achieves a preset approximation accuracy.

Harmonic Analysis Perspective - The Ridgelet Transform

Candes' Results

Denote $\sigma_{a,u,b}(x) = \frac{1}{\sqrt{a}}\sigma(\frac{u^T x - b}{a})$ and let $d\mu(a,u,b) = \frac{da}{a^{n+1}} du db$ denote a normalized measure on $M = \mathbb{R}^+ \times S^{n-1} \times \mathbb{R}$.

Theorem (E. Candes, 1999)

Assume $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies the admissibility condition $\int_{-\infty}^{\infty} |\hat{\sigma}(\omega)|^2/|\omega|^n d\omega < \infty$. Then

1. *For any $f \in L^1(\mathbb{R}^n)$ so that $\hat{f} \in L^1(\mathbb{R}^n)$,*

$$f = c_\sigma \int_M \langle f, \sigma_{a,u,b}\rangle \sigma_{a,u,b} d\mu(a,u,b) \ , \ \|f\|_2^2 = c_\sigma \int_M |\langle f, \sigma_{a,u,b}\rangle|^2 d\mu(a,u,b)$$

   with absolute convergence of the integrals. The constant c_σ is proportional to the admissibility constant.

2. *The map $R : L^2(\mathbb{R}^n) \to L^2(M; d\mu)$, $f \mapsto R(f) = \langle f, \sigma_{a,u,b}\rangle$ is a multiple of an isometry.*

Frames of Ridglets

Theorem

3. *Assume further that: (i) $\hat{\sigma}$ has a 0 of order at least $n/2$ at origin; (ii) $\hat{\sigma}$ decays like $1/|\omega|^{2+\varepsilon}$ at $\pm\infty$; and (iii) For some $a_0 > 0$, $\inf_{1\leq|\omega|\leq a_0}\sum_{j\geq 0}|\hat{\sigma}(a_0^{-j}\omega)|^2|a_0^{-j}\omega|^{-(n-1)} > 0$. Let $j_0 = j_0(a_0,n) = \lfloor \log_{a_0}\left(\frac{\pi}{2\lceil \pi n/log(n)\rceil}\right)\rfloor - 1$ be a certain integer (defining the coarsest scale). Then there exists a $b_0^* > 0$ so that for every $b_0 < b_0^*$ the set of functions $\sigma_{j,u,k}(x) = a_0^{j/2}\sigma(a_0^j\langle u,x\rangle - kb_0)$ indexed by $\Gamma = \cup_{j\geq j_0}(\{j\}\times E_j\times\mathbb{Z})$ where E_j is an ε_j-net of the unit sphere S^{n-1} with $\varepsilon_j = \frac{1}{2}a_0^{j-j_0}$ defines a frame for $L^2([-1,1]^n)$. Specifically, this means that there are $0 < A \leq B < \infty$ so that for every $f\in L^2([-1,1]^n)$,*

$$A\|f\|_2^2 \leq \sum_{(j,u,k)\in\Gamma}|\langle f,\sigma_{j,u,k}\rangle|^2 \leq B\|f\|_2^2.$$

Table of Contents

"This material is based upon work supported by the National Science Foundation under Grant No. DMS-1413249. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation." The author has been partially supported by ARO under grant W911NF1610008 and LTS under grant H9823013D00560049.

More details included in:

- D. Zou, R. Balan, M. Singh, *On Lipschitz Bounds of General Convolutional Neural Networks*, IEEE Trans.on Info.Theory, vol. 66(3), 1738–1759 (2020) doi: 10.1109/TIT.2019.2961812.
- R. Balan, M. Singh, D. Zou, "Lipschitz Properties for Deep Convolutional Networks", arXiv:1701.05217 [cs.LG], Contemporary Mathematics 706, 129-151 (2018) http://dx.doi.org/10.1090/conm/706/14205.

Machine Learning

According to Wikipedia (attributed to Arthur Samuel 1959), "Machine Learning [...] gives computers the ability to learn without being explicitly programmed."
While it has been first coined in 1959, today's machine learning, as a field, evolved from and overlaps with a number of other fields: computational statistics, mathematical optimizations, theory of linear and nonlinear systems.

Machine Learning

According to Wikipedia (attributed to Arthur Samuel 1959), "Machine Learning [...] gives computers the ability to learn without being explicitly programmed."

While it has been first coined in 1959, today's machine learning, as a field, evolved from and overlaps with a number of other fields: computational statistics, mathematical optimizations, theory of linear and nonlinear systems.

Types of problems (tasks) in machine learning:

1. Supervised Learning: The machine (computer) is given pairs of inputs and desired outputs and is left to learn the general association rule.
2. Unsupervised Learning: The machine is given only input data, and is left to discover structures (patterns) in data.
3. Reinforcement Learning: The machine operates in a dynamic environment and had to adapt (learn) continuously as it navigates the problem space (e.g. autonomous vehicle).

Example 1: The AlexNet

The ImageNet Dataset

Dataset: ImageNet dataset. Currently: 14.2 mil.images; 21841 categories; image-net.org
Task: Classify an input image, i.e. place it into one category.

Figure: The "ostrich" category "Struthio Camelus" 1393 pictures. From image-net.org

Example 1: The AlexNet

The Supervised Machine Learning

The AlexNet is 8 layer network, 5 convolutive layers plus 3 dense layers. Introduced by (Alex) Krizhevsky, Sutskever and Hinton in 2012 [KSH12]. Trained on a subset of the ImageNet: Part of the ImageNet Large Scale Visual Recognition Challenge 2010-2012: 1000 object classes and 1,431,167 images.

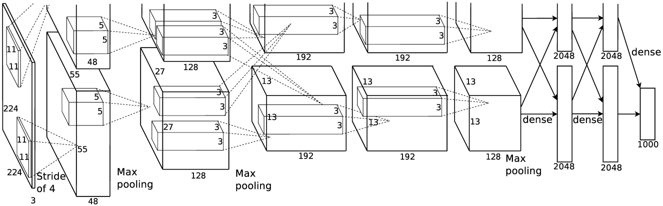


Figure: From Krizhevsky et all 2012: AlexNet: 5 convolutive layers + 3 dense layers. Input size: 224x224x3 pixels. Output size: 1000.

Example 1: The AlexNet

Adversarial Perturbations

The authors of [Szegedy'13] (Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus, 'Intriguing properties ...') found small variations of the input, almost imperceptible, that produced completely different classification decisions:

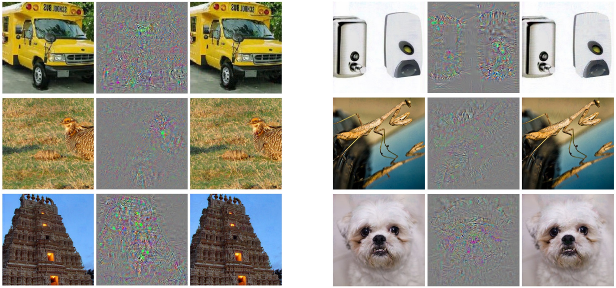


Figure: From Szegedy et all 2013: AlexNet: 6 different classes: original image, difference, and adversarial example – all classified as 'ostrich'

Example 1: The AlexNet

Lipschitz Analysis

Szegedy et all 2013 computed the Lipschitz constants of each layer.

Layer	Size	Sing.Val
Conv. 1	$3 \times 11 \times 11 \times 96$	20
Conv. 2	$96 \times 5 \times 5 \times 256$	10
Conv. 3	$256 \times 3 \times 3 \times 384$	7
Conv. 4	$384 \times 3 \times 3 \times 384$	7.3
Conv. 5	$384 \times 3 \times 3 \times 256$	11
Fully Conn.1	$9216(43264) \times 4096$	3.12
Fully Conn.2	4096×4096	4
Fully Conn.3	4096×1000	4

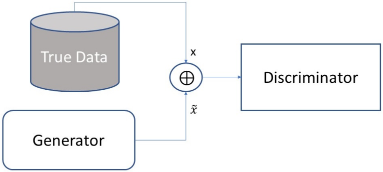
Overall Lipschitz constant:

$$Lip \leq 20 * 10 * 7 * 7.3 * 11 * 3.12 * 4 * 4 = 5,612,006$$

Example 2: Generative Adversarial Networks

The GAN Problem

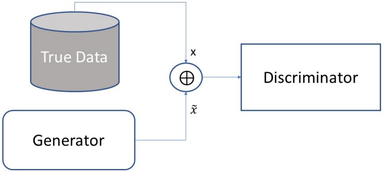
Two systems are involved: a *generator* network producing synthetic data; a *discriminator* network that has to decide if its input is synthetic data or real-world (true) data:



Example 2: Generative Adversarial Networks

The GAN Problem

Two systems are involved: a *generator* network producing synthetic data; a *discriminator* network that has to decide if its input is synthetic data or real-world (true) data:



Introduced by Goodfellow et al in 2014, GANs solve a minimax optimization problem:

$$\min_G \max_D \mathbb{E}_{x\sim P_r}\left[log(D(x))\right] + \mathbb{E}_{\tilde{x}\sim P_g}\left[log(1-D(\tilde{x}))\right]$$

where P_r is the distribution of true data, P_g is the generator distribution, and $D : x \mapsto D(x) \in [0,1]$ is the discriminator map (1 for likely true data; 0 for likely synthetic data).

Example 2: Generative Adversarial Networks

The Wasserstein Optimization Problem

In practice, the training algorithms do not behave well ("saddle point effect").
The Wasserstein GAN (Arjovsky et al 2017) replaces the Jensen-Shannon divergence by the Wasserstein-1 distance:

$$\min_{G} \max_{D \in Lip(1)} \mathbb{E}_{x \sim P_r}[D(x)] - \mathbb{E}_{\tilde{x} \sim P_g}[D(\tilde{x})]$$

where $Lip(1)$ denotes the set of Lipschitz functions with constant 1, enforced by weight clipping.

Example 2: Generative Adversarial Networks

The Wasserstein Optimization Problem

In practice, the training algorithms do not behave well ("saddle point effect").
The Wasserstein GAN (Arjovsky et al 2017) replaces the Jensen-Shannon divergence by the Wasserstein-1 distance:

$$\min_{G} \max_{D \in Lip(1)} \mathbb{E}_{x \sim P_r}[D(x)] - \mathbb{E}_{\tilde{x} \sim P_g}[D(\tilde{x})]$$

where $Lip(1)$ denotes the set of Lipschitz functions with constant 1, enforced by weight clipping.
Gulrajani et al in 2017 proposed to incorporate the Lip(1) condition into the optimization criterion using a soft Lagrange multiplier technique for minimization of:

$$L = \mathbb{E}_{\tilde{x} \sim P_g}[D(x)] - \mathbb{E}_{x \sim P_r}[D(x)] + \lambda \, \mathbb{E}_{\hat{x} \sim P_{\hat{x}}}\left[\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2 \right]$$

where $\hat{x}$ is sampled uniformly between $x \sim P_r$ and $\tilde{x} \sim P_g$.

Example 3: Uncertainty Propagation through DNN

This example is based on a recent project with Prof. Thomas Ernst, UMB, School of Medicine, Baltimore.
The standard way of quantifying uncertainty is through the Cramer-Rao Lower Bound (CRLB). Fisher Information Matrix $I(z)$ and $CRLB$:

$$I(z) = \mathbb{E}\left[(\nabla_z log(p(x;z))) (\nabla_z log(p(x;z)))^T \right] \quad , \quad CRLB = (I(z))^{-1}$$

Interpretation: Covariance of any *unbiased* estimator of z is lower bounded $CRLB$. For AWGN with variance σ^2,

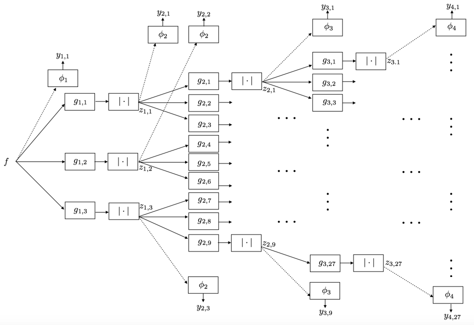
$$CRLB = \sigma^2 \left(J_F^T J_F \right)^{-1} \quad , \quad J_F = \left[\frac{\partial F_k}{\partial z_j} \right]_{(j,k) \in [n] \times [d]} \in \mathbb{R}^{n \times d}$$

where J_F denotes the Jacobian matrix of the forward model.
Goal: Determine $CRLB$ and use it to measure the confidence in the reconstructed image $\hat{z}$.
Challenge: The exact form of F is unknown! But we assume we know a left-inverse (the DNN) G_0. It turns out a good proxy is $CRLB = \sigma^2 J_{G_0} J_{G_0}^T$.

Example 4: The Scattering Network

Topology

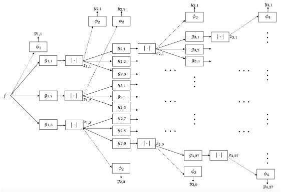
Example of Scattering Network; definition and properties: [Mallat'12]; this example from [B.,Singh,Zou'17]:



Input: f; Outputs: $y = (y_{l,k})$.

Example 4: Scattering Network

Lipschitz Analysis

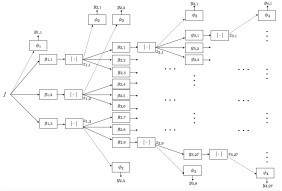


Remarks:

- Outputs from each layer

Example 4: Scattering Network

Lipschitz Analysis

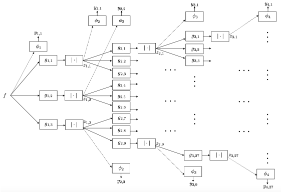


Remarks:

- Outputs from each layer
- Tree-like topology

Example 4: Scattering Network

Lipschitz Analysis

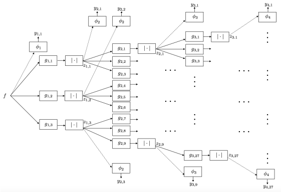


Remarks:

- Outputs from each layer
- Tree-like topology
- Backpropagation/Chain rule: Lipschitz bound 40.

Example 4: Scattering Network

Lipschitz Analysis



Remarks:

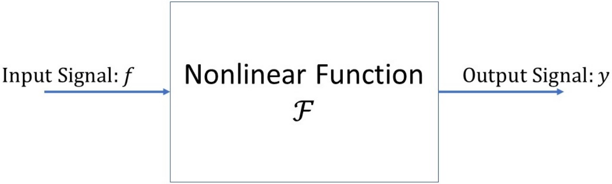
- Outputs from each layer
- Tree-like topology
- Backpropagation/Chain rule: Lipschitz bound 40.
- Mallat's result predicts $Lip = 1$.

Problem Formulation

Nonlinear Maps

Consider a nonlinear function between two metric spaces,

$$\mathcal{F} : (X, d_X) \rightarrow (Y, d_Y).$$



Problem Formulation

Lipschitz analysis of nonlinear systems

$$\mathcal{F} : (X, d_X) \to (Y, d_Y)$$

$\mathcal{F}$ is called *Lipschitz* with constant C if for any $f, \tilde{f} \in X$,

$$d_Y(\mathcal{F}(f), \mathcal{F}(\tilde{f})) \le C \; d_X(f, \tilde{f})$$

The optimal (i.e. smallest) Lipschitz constant is denoted $Lip(\mathcal{F})$. The square C^2 is called Lipschitz bound (similar to the Bessel bound).

$\mathcal{F}$ is called *bi-Lipschitz* with constants $C_1, C_2 > 0$ if for any $f, \tilde{f} \in X$,

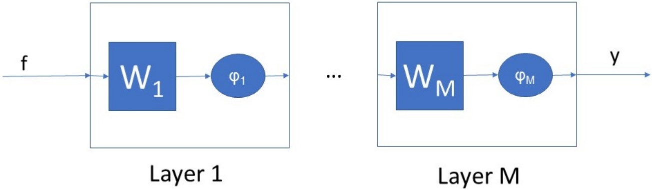
$$C_1 \; d_X(f, \tilde{f}) \le d_Y(\mathcal{F}(f), \mathcal{F}(\tilde{f})) \le C_2 \; d_X(f, \tilde{f})$$

The square C_1^2, C_2^2 are called *Lipschitz bounds* (similar to frame bounds).

Problem Formulation

Motivating Examples

Consider the typical neural network as a feature extractor component in a classification system:



$$g = \mathcal{F}(f) = \mathcal{F}_M(\ldots\mathcal{F}_1(f; W_1, \varphi_1); \ldots; W_M, \varphi_M)$$

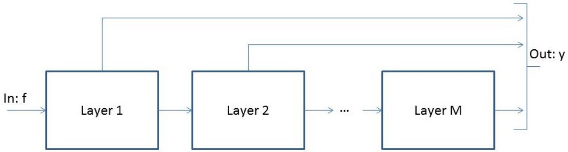
$$\mathcal{F}_m(f; W_m, \varphi_m) = \varphi_m(W_m f)$$

W_m is a linear operator (matrix); φ_m is a Lip(1) scalar nonlinearity (e.g. Rectified Linear Unit).

Problem Formulation

Problem 1

Given a deep network:



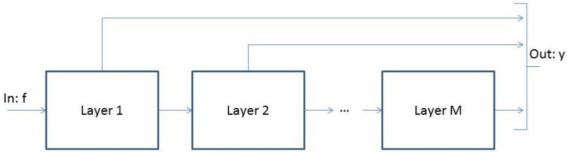
Estimate the Lipschitz constant, or bound:

$$Lip = \sup_{f \neq \tilde{f} \in L^2} \frac{\|y - \tilde{y}\|_2}{\|f - \tilde{f}\|_2} \, , \quad Bound = \sup_{f \neq \tilde{f} \in L^2} \frac{\|y - \tilde{y}\|_2^2}{\|f - \tilde{f}\|_2^2}.$$

Problem Formulation

Problem 1

Given a deep network:



Estimate the Lipschitz constant, or bound:

$$Lip = \sup_{f \neq \tilde{f} \in L^2} \frac{\|y - \tilde{y}\|_2}{\|f - \tilde{f}\|_2} \quad , \quad Bound = \sup_{f \neq \tilde{f} \in L^2} \frac{\|y - \tilde{y}\|_2^2}{\|f - \tilde{f}\|_2^2}.$$

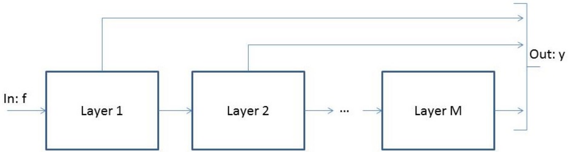
Methods (Approaches):

1. Standard Method: Backpropagation, or chain-rule
2. New Method: Storage function based approach (dissipative systems)
3. Numerical Method: Simulations

Problem Formulation

Problem 2

Given a deep network:



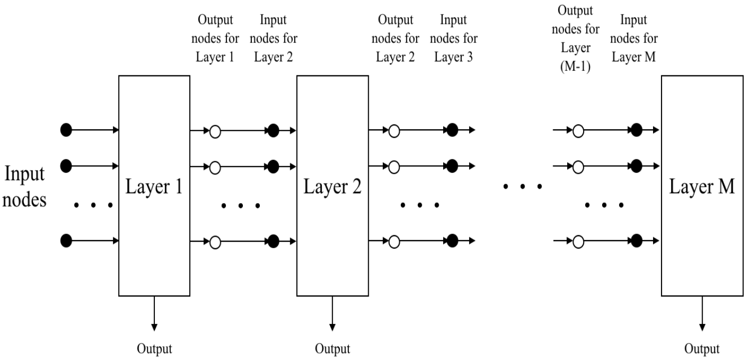
Estimate the stability of the output to specific variations of the input:

1. Invariance to deformations: $\tilde{f}(x) = f(x - \tau(x))$, for some smooth τ.
2. Covariance to such deformations $\tilde{f}(x) = f(x - \tau(x))$, for smooth τ and bandlimited signals f;
3. Tail bounds when f has a known statistical distribution (e.g. normal with known spectral power)

ConvNet

Topology

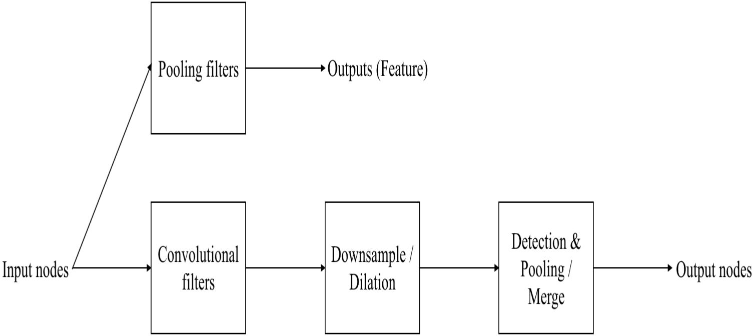
A deep convolution network is composed of multiple layers:



ConvNet

One Layer

Each layer is composed of two or three sublayers: convolution, downsampling, detection/pooling/merge.



ConvNet: Sublayers

Linear Filters: Convolution and Pooling-to-Output Sublayer

$$f^{(2)} = g * f^{(1)} \quad , \quad g * f^{(1)}(x) = \int g(x-\xi) f^{(1)}(\xi) d\xi$$

where $g \in \mathcal{B} = \{g \in \mathcal{S}' \ , \ \hat{g} \in L^\infty(\mathbb{R}^d)\}$.

$(\mathcal{B}, *)$ is a Banach algebra with norm $\|g\|_{\mathcal{B}} = \|\hat{g}\|_\infty$.
Notation: g for regular convolution filters, and Φ for pooling-to-output filters.

ConvNet: Sublayers

Downsampling Sublayer

$$f^{(2)}(x) = f^{(1)}(Dx)$$

For $f^{(1)} \in L^2(\mathbb{R}^d)$ and $D = D_0 \cdot I$, $f^{(2)} \in L^2(\mathbb{R}^d)$ and

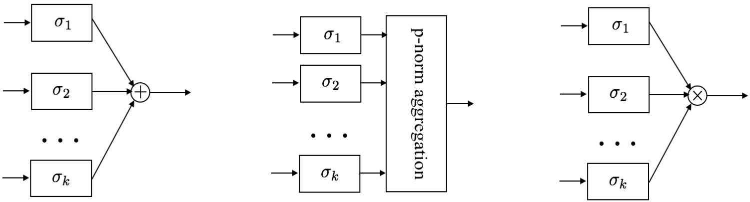
$$\|f^{(2)}\|_2^2 = \int_{\mathbb{R}^d} |f^{(2)}(x)|^2 dx = \frac{1}{|det(D)|} \int_{\mathbb{R}^d} |f^{(1)}(x)|^2 dx = \frac{1}{D_0^d} \|f^{(1)}\|_2^2$$

ConvNet: Sublayers

Detection and Pooling Sublayer

We consider three types of detection/pooling/merge sublayers:

- Type I, τ_1: Componentwise Addition: $z = \sum_{j=1}^{k} \sigma_j(y_j)$
- Type II, τ_2: p-norm aggregation: $z = \left(\sum_{j=1}^{k} |\sigma_j(y_j)|^p\right)^{1/p}$
- Type III, τ_3: Componentwise Multiplication: $z = \prod_{j=1}^{k} \sigma_j(y_j)$

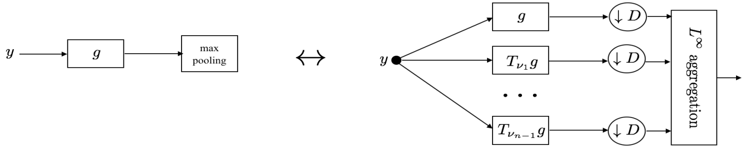


Assumptions: (1) σ_j are scalar Lipschitz functions with $Lip(\sigma_j) \leq 1$; (2) If σ_j is connected to a multiplication block then $\|\sigma_j\|_\infty \leq 1$.

ConvNet: Sublayers

MaxPooling and AveragePooling

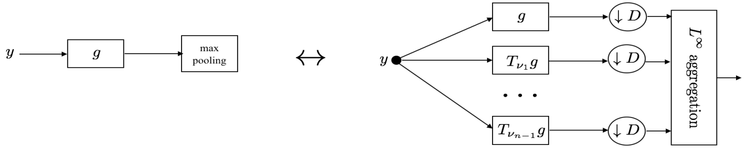
MaxPooling can be implemented as follows:



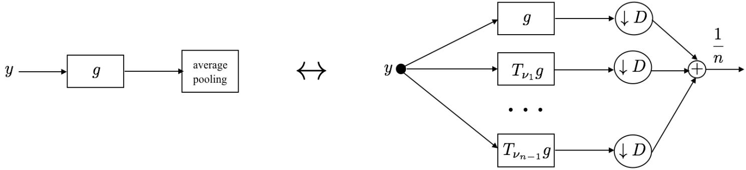
ConvNet: Sublayers

MaxPooling and AveragePooling

MaxPooling can be implemented as follows:

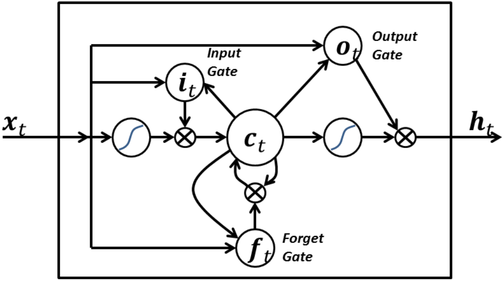


AveragePooling can be implemented as follows:



ConvNet: Sublayers

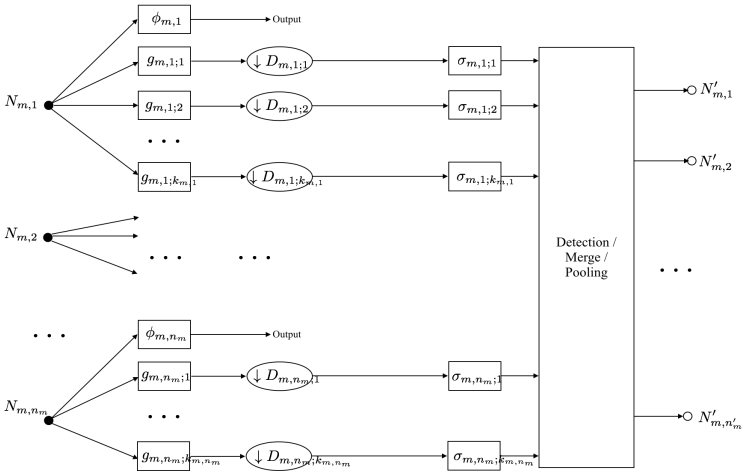
Long Short-Term Memory



Long Short-Term Memory (LSTM) networks
[Hochreiter,Schmidhuber.'97],[Greff et.al.'15].
By BiObserver - Own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=43992484

ConvNet: Layer m

Components of the m^{th} layer



ConvNet: Layer m

Topology coding of the m^{th} layer

n_m denotes the number of input nodes in the m-th layer:
$\mathcal{I}_m = \{N_{m,1}, N_{m,2}, \cdots, N_{m,n_m}\}$.
Filters:

1. pooling filter: $\phi_{m,n}$ for node n, in layer m;
2. convolution filter: $g_{m,n,k}$ for input node n to output node k, in layer m;

For node n: $G_{m,n} = \{g_{m,n;1}, \cdots g_{m,n;k_{m,n}}\}$.
The set of all convolution filters in layer m: $G_m = \cup_{n=1}^{n_m} G_{m,n}$.

ConvNet: Layer m

Topology coding of the m^{th} layer

n_m denotes the number of input nodes in the m-th layer:
$\mathcal{I}_m = \{N_{m,1}, N_{m,2}, \cdots, N_{m,n_m}\}$.
Filters:

1. pooling filter: $\phi_{m,n}$ for node n, in layer m;
2. convolution filter: $g_{m,n,k}$ for input node n to output node k, in layer m;

For node n: $G_{m,n} = \{g_{m,n;1}, \cdots g_{m,n;k_{m,n}}\}$.
The set of all convolution filters in layer m: $G_m = \cup_{n=1}^{n_m} G_{m,n}$.
$\mathcal{O}_m = \{N'_{m,1}, N'_{m,2}, \cdots, N'_{m,n'_m}\}$ the set of output nodes of the m-th layer.
Note that $n'_m = n_{m+1}$ and there is a one-one correspondence between $\mathcal{O}_m$ and $\mathcal{I}_{m+1}$.
The output nodes automatically partitions G_m into n'_m disjoint subsets $G_m = \cup_{n'=1}^{n'_m} G'_{m,n'}$, where $G'_{m,n'}$ is the set of filters merged into $N'_{m,n'}$.

ConvNet: Layer m

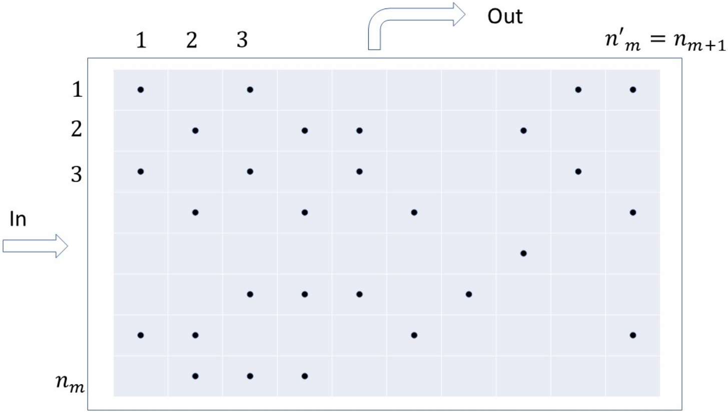
Topology coding of the m^{th} layer

For each filter $g_{m,n;k}$, we define an associated *multiplier* $l_{m,n;k}$ in the following way: suppose $g_{m,n;k} \in G'_{m,k}$, let $K = \left|G'_{m,k}\right|$ denote the cardinality of $G'_{m,k}$. Then

$$l_{m,n;k} = \begin{cases} K & \text{, if } g_{m,n;k} \in \mathcal{T}_1 \cup \mathcal{T}_3 \\ K^{\max\{0,2/p-1\}} & \text{, if } g_{m,n;k} \in \mathcal{T}_2 \end{cases} \tag{3.1}$$

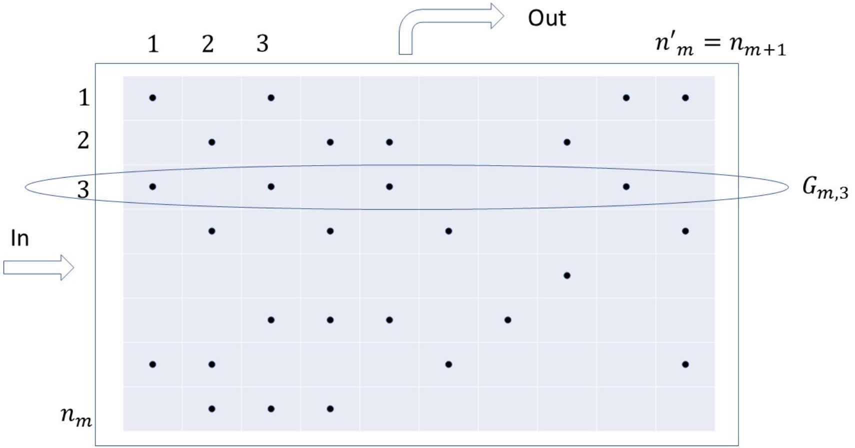
ConvNet: Layer m

Topology coding of the m^{th} layer



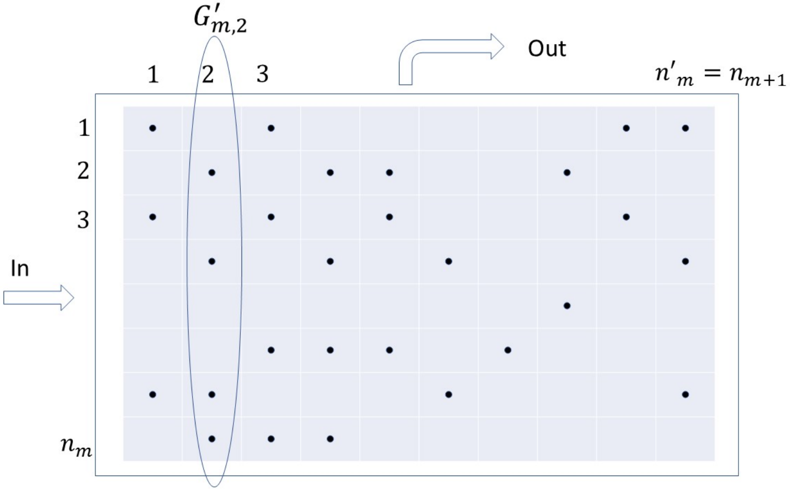
ConvNet: Layer m

Topology coding of the m^{th} layer



ConvNet: Layer m

Topology coding of the m^{th} layer



Semi-discrete Bessel Systems

A countable set of functions $\{g_n\ ,\ n\geq 1\}\subset L^2(S)$ (where S is a LCA group) is called a *semi-discrete Bessel system* in $L^2(S)$ if there is a constant (called a *Bessel bound*) $B\geq 0$ such that, for every $f\in L^2(S)$,

$$\sum_{n\geq 1}\|f*g_n\|_2^2\leq B\|f\|_2^2 \quad , \quad f*g_n(x)=\int_S f(x-y)g_n(y)dy.$$

The Lipschitz constant of a linear operator equals its operator norm. For nonlinear maps, the Lipschitz bound (square of its Lipschitz constant) is a replacement for the Bessel bound (or, the upper frame bound).

Lemma

Assume $\{g_n\ ,\ n\geq 1\}$ is a semi-discrete Bessel system in $L^2(\mathbb{R}^d)$. Then its optimal Bessel bound is given by

$$B=\sup_{\omega\in\mathbb{R}^n}\sum_{n\geq 1}|\widehat{g_n}(\omega)|^2=:\|\sum_{n\geq 1}|\widehat{g_n}|^2\|_\infty.$$

Layer Analysis

Bessel Bounds

In each layer m and for each *input* node n we define three types of Bessel bounds (one for each type of the detection/pooling/merge sublayer):

- 1st type Bessel bound:

$$B_{m,n}^{(1)} = \| \left|\hat{\phi}_{m,n}\right|^2 + \sum_{g_{m,n;k} \in G_{m,n}} l_{m,n;k} D_{m,n;k}^{-d} \left|\hat{g}_{m,n;k}\right|^2 \|_\infty \tag{3.2}$$

- 2nd type Bessel bound:

$$B_{m,n}^{(2)} = \| \sum_{g_{m,n;k} \in G_{m,n}} l_{m,n;k} D_{m,n;k}^{-d} \left|\hat{g}_{m,n;k}\right|^2 \|_\infty \tag{3.3}$$

- 3rd type (or generating) bound:

$$B_{m,n}^{(3)} = \|\hat{\phi}_{m,n}\|_\infty^2 . \tag{3.4}$$

Layer Analysis

Bessel Bounds

Next we define the layer m Bessel bounds:

$$1^{\text{st}} \text{ type Bessel bound} \quad B_m^{(1)} = \max_{1 \leq n \leq n_m} B_{m,n}^{(1)} \tag{3.5}$$

$$2^{\text{nd}} \text{ type Bessel bound} \quad B_m^{(2)} = \max_{1 \leq n \leq n_m} B_{m,n}^{(2)} \tag{3.6}$$

$$3^{\text{rd}} \text{ type (generating) Bessel bound} \quad B_m^{(3)} = \max_{1 \leq n \leq n_m} B_{m,n}^{(3)}. \tag{3.7}$$

Remark. These bounds characterize Bessel bounds of the associated semi-discrete Bessel systems.

Lipschitz Analysis

First Result

Theorem (1. BSZ'17)

Consider a Convolutional Neural Network $\mathcal{F}$ with M layers as described before, with non-expansive Lipschitz activation functions, $Lip(\varphi_{m,n,n'}) \leq 1$. Additionally, those $\varphi_{m,n,n'}$ that aggregate into a multiplicative block satisfy $\|\varphi_{m,n,n'}\|_\infty \leq 1$. Let the m-th layer 1st type Bessel bound be

$$B_m^{(1)} = \max_{1 \leq n \leq n_m} \| \left|\hat{\phi}_{m,n}\right|^2 + \sum_{k=1}^{k_{m,n}} l_{m,n;k} D_{m,n;k}^{-d} \left|\hat{g}_{m,n;k}\right|^2 \|_\infty .$$

Then the Lipschitz bound of the entire CNN is upper bounded by $\prod_{m=1}^{M} max(1, B_m^{(1)})$. Specifically, for any $f, \tilde{f} \in L^2(\mathbb{R}^d)$:

$$\|\mathcal{F}(f) - \mathcal{F}(\tilde{f})\|_2^2 \leq \left(\prod_{m=1}^{M} max(1, B_m^{(1)}) \right) \|f - \tilde{f}\|_2^2,$$

Lipschitz Analysis

Second Result

Theorem (2. BSZ'20)

Consider a Convolutional Neural Network with M layers as described before, where all scalar nonlinearities satisfy the same conditions as in the previous result. For layer m, let $B_m^{(1)}$, $B_m^{(2)}$, and $B_m^{(3)}$ denote the three Bessel bounds defined earlier. Denote by L the optimal solution of the following linear program:

$$\begin{aligned}
\Gamma = \max_{y_1,\dots,y_M,z_1,\dots,z_M\geq 0} \quad & \sum_{m=1}^{M} z_m \\
s.t. \quad & y_0 = 1 \\
& y_m + z_m \leq B_m^{(1)} y_{m-1}, \quad 1 \leq m \leq M \\
& y_m \leq B_m^{(2)} y_{m-1}, \quad 1 \leq m \leq M \\
& z_m \leq B_m^{(3)} y_{m-1}, \quad 1 \leq m \leq M
\end{aligned} \tag{3.8}$$

Lipschitz Analysis

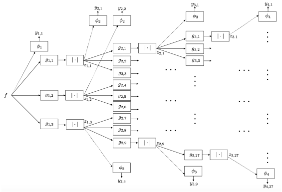
Second Result - cont'd

Theorem (2. BSZ'20)

Then the Lipschitz bound satisfies $Lip(\mathcal{F})^2 \leq \Gamma$. *Specifically, for any* $f, \tilde{f} \in L^2(\mathbb{R}^d)$:

$$\|\mathcal{F}(f) - \mathcal{F}(\tilde{f})\|_2^2 \leq \Gamma \|f - \tilde{f}\|_2^2,$$

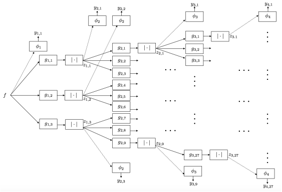
Example 1: Scattering Network



The Lipschitz constant:

- Backpropagation/Chain rule: Lipschitz bound 40 (hence $Lip \leq 6.3$).

Example 1: Scattering Network

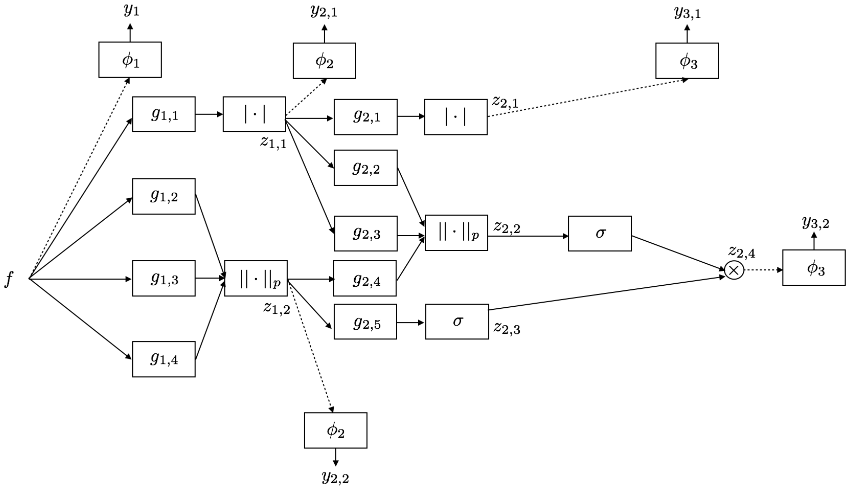


The Lipschitz constant:

- Backpropagation/Chain rule: Lipschitz bound 40 (hence $Lip \leq 6.3$).
- Using our main theorem, $Lip \leq 1$, but Mallat's result: $Lip = 1$.

Filters have been choosen as in a dyadic wavelet decomposition. Thus $B_m^{(1)} = B_m^{(2)} = B_m^{(3)} = 1$, $1 \leq m \leq 4$.

Example 2: A General Convolutive Neural Network



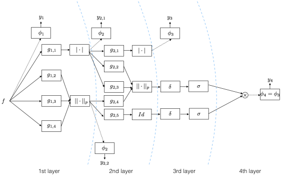
Example 2: A General Convolutive Neural Network

Set $p = 2$ and:

$$F(\omega) = \exp(\frac{4\omega^2 + 4\omega + 1}{4\omega^2 + 4\omega})\chi_{(-1,-1/2)}(\omega) + \chi_{(-1/2,1/2)}(\omega) + \exp(\frac{4\omega^2 - 4\omega + 1}{4\omega^2 - 4\omega})\chi_{(1/2,1)}(\omega).$$

$$\begin{aligned}
\hat{\phi}_1(\omega) &= F(\omega) \\
\hat{g}_{1,j}(\omega) &= F(\omega + 2j - 1/2) + F(\omega - 2j + 1/2) \ , \ j = 1, 2, 3, 4 \\
\hat{\phi}_2(\omega) &= \exp(\frac{4\omega^2 + 12\omega + 9}{4\omega^2 + 12\omega + 8})\chi_{(-2,-3/2)}(\omega) + \\
&\quad \chi_{(-3/2,3/2)}(\omega) + \exp(\frac{4\omega^2 - 12\omega + 9}{4\omega^2 - 12\omega + 8})\chi_{(3/2,2)}(\omega) \\
\hat{g}_{2,j}(\omega) &= F(\omega + 2j) + F(\omega - 2j) \ , \ j = 1, 2, 3 \\
\hat{g}_{2,4}(\omega) &= F(\omega + 2) + F(\omega - 2) \\
\hat{g}_{2,5}(\omega) &= F(\omega + 5) + F(\omega - 5) \\
\hat{\phi}_3(\omega) &= \exp(\frac{4\omega^2 + 20\omega + 25}{4\omega^2 + 20\omega + 24})\chi_{(-3,-5/2)}(\omega) + \\
&\quad \chi_{(-5/2,5/2)}(\omega) + \exp(\frac{4\omega^2 - 20\omega + 25}{4\omega^2 - 20\omega + 25})\chi_{(5/2,3)}(\omega).
\end{aligned}$$

Example 2: A General Convolutive Neural Network



Bessel Bounds: $B_m^{(1)} = 2e^{-1/3} = 1.43$, $B_m^{(2)} = B_m^{(3)} = 1$.

The Lipschitz bound:

- Using backpropagation/chain-rule: $Lip^2 \leq 5$.
- Using Theorem 1: $Lip^2 \leq 2.9430$.
- Using Theorem 2 (linear program): $Lip^2 \leq 2.2992$.

Example 3: Lipschitz constant based objective functions

Nonlinear Discriminant Analysis

In Linear Discriminant Analysis (LDA), the objective is to maximize the "separation" between two classes, while controlling the variances within class.

A similar nonlinear *discriminant* can be defined:

$$S = \frac{\|\mathbb{E}[\mathcal{F}(f)|f \in C_1] - \mathbb{E}[\mathcal{F}(f)|f \in C_2]\|^2}{\|Cov(\mathcal{F}(f)|f \in C_1)\|_F + \|Cov(\mathcal{F}(f)|f \in C_2)\|_F}.$$

Example 3: Lipschitz constant based objective functions

Nonlinear Discriminant Analysis

In Linear Discriminant Analysis (LDA), the objective is to maximize the "separation" between two classes, while controlling the variances within class.

A similar nonlinear *discriminant* can be defined:

$$S = \frac{\|\mathbb{E}[\mathcal{F}(f)|f \in C_1] - \mathbb{E}[\mathcal{F}(f)|f \in C_2]\|^2}{\|Cov(\mathcal{F}(f)|f \in C_1)\|_F + \|Cov(\mathcal{F}(f)|f \in C_2)\|_F}.$$

Replace the statistics $\|Cov\|_F$ by Lipschitz bounds:

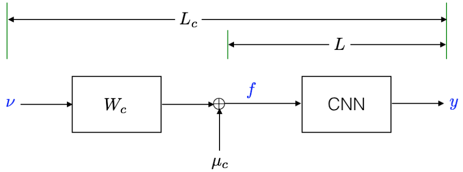
Lipschitz bound based separation:

$$\tilde{S} = \frac{\|\mathbb{E}[\mathcal{F}(f)|f \in C_1] - \mathbb{E}[\mathcal{F}(f)|f \in C_2]\|^2}{Lip_1^2 + Lip_2^2}.$$

Example 3: Lipschitz constant based objective functions

Nonlinear Discriminant Analysis

The Lipschitz bounds Lip_1^2, Lip_2^2 are computed using Gaussian generative models for the two classes: $(\mu_c, W_c W_c^T)$, where W_c represents the whitening filter for class $c \in \{1, 2\}$.



Example 3: Lipschitz constant based objective functions

Numerical Results

Dataset: MNIST database; input images: 28×28 pixels. Two classes: "3" and "8"

Classifier: 3 layer and 4 layer random CNN, followed by a trained SVM.

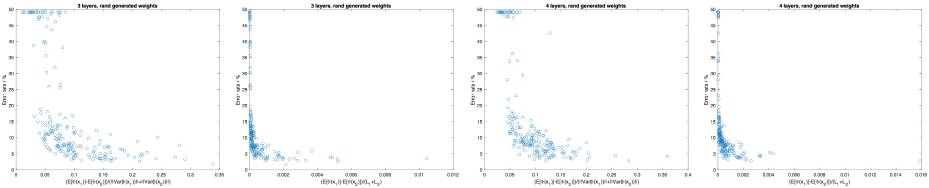


Figure: Results for uniformly distributed random weights

Conclusion: The error rate decreases as the Lipschitz bound separation increases. The discriminant spread is wider.

Example 3: Lipschitz constant based objective functions

Numerical Results

Dataset: MNIST database; input images: 28×28 pixels. Two classes: "3" and "8"

Classifier: 3 layer and 4 layer random CNN, followed by a trained SVM.

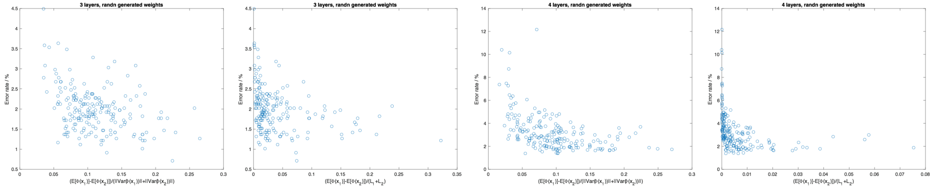


Figure: Results for normaly distributed random weights

Local Analysis

Consider a deep network $\mathcal{F} : (X, \|\cdot\|_2) \to (Y, \|\cdot\|_2)$ between Euclidean finite-dimensional linear spaces with M layers, where the i^{th} layer is characterized by the input-output nonlinear Lipschitz map $\mathcal{F}_i$. Denote by $J_{\mathcal{F}}$, $J_{\mathcal{F}_i}$ the Jacobian matrices of these maps. Then by an application of the Fundamental Theorem of Calculus (plus Lebesgue's differentiation theorem), the optimal Lipschitz constant is

$$Lip(\mathcal{F}) = \sup_{x \in X} \|J_{\mathcal{F}}(x)\|_{Op} = \sup_{x \in X} \|J_{\mathcal{F}_M} \cdots J_{\mathcal{F}_1}(x)\|_{Op}$$

where the Op norm is the largest singular value of the corresponding Jacobian.

In the case of type I or II network (i.e., no multiplicative aggregation), the nonlinear are homogeneous of degree 1, and in each layer the Jacobian factors as a product of 3 matrices:

$$J_{\mathcal{F}}(x) = P_M(x) D_M(x) A_M P_{M-1}(x) D_{M-1}(x) A_{M-1} \cdots P_1(x) D_1(x) A_1,$$

Local Analysis (2)

$$J_{\mathcal{F}}(x) = P_M(x)D_M(x)A_M P_{M-1}(x)D_{M-1}(x)A_{M-1}\cdots P_1(x)D_1(x)A_1,$$

where: A_i is the matrix associated to linear operators (filters), D_i is the diagonal matrix associated to derivative of activation functions (it is a binary matrix composed of 0's and 1's in the case of ReLU activation), and P_i is the matrix associated to the composition of downsampling and pooling sublayers. In the case of sum-pooling, P_i is independent of input x; in the case of max-filter, it has a weak dependency on x. In both cases it is sparse, with binary entries.

Local Analysis (2)

$$J_{\mathcal{F}}(x) = P_M(x)D_M(x)A_M P_{M-1}(x)D_{M-1}(x)A_{M-1}\cdots P_1(x)D_1(x)A_1,$$

where: A_i is the matrix associated to linear operators (filters), D_i is the diagonal matrix associated to derivative of activation functions (it is a binary matrix composed of 0's and 1's in the case of ReLU activation), and P_i is the matrix associated to the composition of downsampling and pooling sublayers. In the case of sum-pooling, P_i is independent of input x; in the case of max-filter, it has a weak dependency on x. In both cases it is sparse, with binary entries.

Results for Alex Net using method:	Lip const
Analytical estimate: based on Theorem 1	2.51×10^3
Empirical bound: quotient from pairs of samples	7.32×10^{-3}
Numerical estimate: maximize the "sandwich" formula	1.44

Local Analysis: Domains of linearity

It is not suprising that the analytic estimate 2.51×10^3 is bigger than the numerical estimate 1.44. The suprising conclusion is the difference between the numerical estimate, 1.44, and the empirical bound 7.32^{-3}.

The "sandwich" formula provides additional information: The upper bound is achieved locally for the principal right-singular vector v at the specific input x where the maximum is achieved. We performed the following numerical expriment: we computed the ratio $R(t) = \frac{1}{t}\|\mathcal{F}(x+tv) - \mathcal{F}(x)\|_2$:

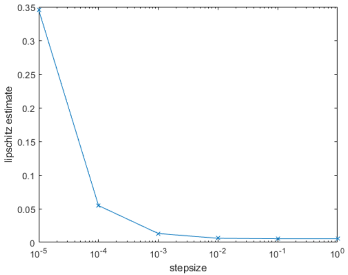


Figure: The ratio $R(t) = ||\mathcal{F}(x + t \cdot v) - \mathcal{F}(x)||/t$ for different t.

Lipschitz Analysis: Stochastic Model

The numerical study of the Alex Net showed that the optimal Lipschitz constant is somewhat theoretical and is achieved by very small perturbations. Notice for two inputs x_1 and x_2:

$$\mathcal{F}(x_1)-\mathcal{F}(x_2) = \int_0^1 J_{\mathcal{F}_M} J_{\mathcal{F}_{M-1}} \cdots J_{\mathcal{F}_1}((1-t)x_1+tx_2)(x_2-x_1)dt = J_* \cdot (x_2-x_1)$$

where the *effective Jacobian* J_* is estimated by

$$J_* \approx (\mathbb{E}[P_M])(\mathbb{E}[D_M])A_M \cdots (\mathbb{E}[P_1])(\mathbb{E}[D_1])A_1$$

where we assume:

1. (ergodicity) x_1 and x_2 are sufficientlly distinct so that the network passes through all linearity domains during the convex combination $x_1 \to (1-t)x_2 + tx_2 \to x_2$, and
2. (independence) the behavior of activation maps and pooling sublayers are independent from layer to layer.

Table of Contents

Invariance and Equivariance

Table of Contents

Graph Deep Learning