On nonlinear analysis of phase retrieval and deep learning

Ph.D. Final Oral Exam

Dongmian Zou

April 13, 2017

Applied Mathematics & Statistics, and Scientific Computation
Norbert Wiener Center and CSCAMM
University of Maryland, College Park

Table of contents

1. Introduction
2. Phase retrievability implies bi-Lipschitz property
3. Global stable reconstruction
4. A framework of CNN for stability analysis
5. Lipschitz bounds for CNN

Introduction

Phase retrieval

In phase retrieval problems, we seek to recover the phase of a signal from the magnitude of linear measurements.

X-ray crystallography

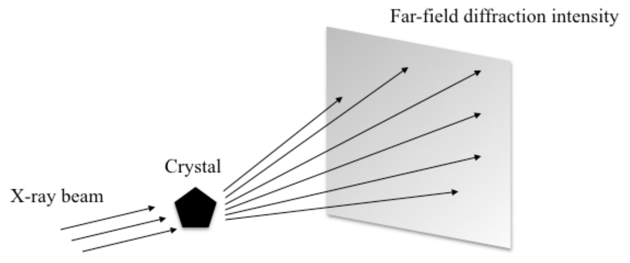


Figure 1: The experiment settings for X-ray crystallography. We desire to know the molecular structure of a crystall. On the far-field we observe the diffraction intensity, which is the magnitude of the Fourier transform of the crystal.

Phase retrieval

Quantum tomography

The quantum measurements are generally described by positive operator valued measures (POVM's). That is,

$$\mathsf{A} = \{A_1, \cdots, A_m\}$$

where each A_k, $k = 1, \cdots, m$, is Hermitian positive semidefinite. The measurement gives

$$\mathsf{A}(\rho) = (\mathrm{tr}\rho A_1, \cdots, \mathrm{tr}\rho A_m) \ .$$

If ρ is a pure state, that is, $\rho = |\psi\rangle\langle\psi|$, then the recovery of the state falls into the problem of generalized phase retrieval. Moreover, if we choose to use rank-one POVM's ($A_k = |f_k\rangle\langle f_k|$, $k = 1, \cdots, m$), then we have

$$\mathrm{tr}\rho A_k = |\langle\psi|f_k\rangle|^2 \ .$$

Phase retrieval

Speech recognition

In speech recognition, sampled speech signals are first transformed to the time-frequency domain via discrete windowed Fourier transform.

The speech signal is sampled as $\{x(t) : t = 0, 1, \cdots, T-1\}$.

The fast windowed Fourier transform is

$$X(k, w) = \sum_{t=0}^{M-1} g(t)x(t + kN)e^{-2\pi i\omega t/M}, \qquad k = 0, 1, \cdots, \frac{T-M}{N}.$$

In commonly used noise reduction method, we apply a nonlinear transform on $|X(k, w)|$ only and does not include the noisy phase. Also in some speech recognition applications, unwrapping the phase is computationally difficult. It is desired to do reconstruction without phase.

Scattering network

In a scattering network [Mallat], we perform linear measurements and take absolute values consecutively.

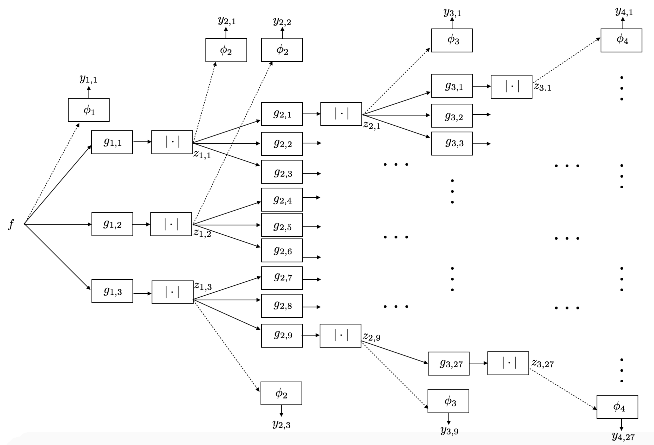


Figure 2: The scattering network. f is the input. y's are the outputs. For instance, $y_{4,1}$ on the top right corner reads $y_{4,1}(t) = \Big|\big||f * g_{1,1}| * g_{2,1}\big| * g_{3,1}\Big| * \phi_4(t)$.

Convolutional neural networks

The scattering network belongs to the large class of convolutional neural networks (CNN).

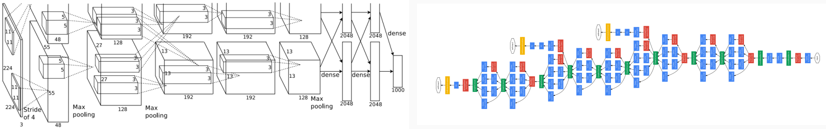


Figure 3: Examples of CNN. The left one is AlexNet used in [Krizhevsky et al.], the right one is GoogleNet used in [Szegedy et al.]

Common elements of a CNN:

- Convolution
    - sparse connection, equivariance to translation
- Activation
    - biological motivation, probabilistic explanation
- Pooling
    - reduce the complexity, invariant to small translation

Stability

We are interested in the stability in the phase retrieval problem and convolutional neural networks.

We look at the Lipschitz property for the respective maps.

Definition

Let (X, d_X) and (Y, d_Y) be two metric spaces where d_X and d_Y are the distance functions respectively. A continuous map $f : X \to Y$ is said to be Lipschitz continuous, or Lipschitz, if

$$\sup_{x_1, x_2 \in X} \frac{d_Y\left(f(x_1), f(x_2)\right)}{d_X\left(x_1, x_2\right)} < \infty \, .$$

In this case, we denote

$$\mathrm{Lip}(f) := \sup_{x_1, x_2 \in X} \frac{d_Y\left(f(x_1), f(x_2)\right)}{d_X\left(x_1, x_2\right)} \, .$$

Bi-Lipschitz

Definition

Let (X, d_X) and (Y, d_Y) be two metric spaces where d_X and d_Y are the distance functions respectively. A continuous map $f : X \to Y$ is said to be bi-Lipschitz, if there exist constants A and B, with $0 < A \leq B < \infty$, such that

$$A d_X(x_1, x_2) \leq d_Y(f(x_1), f(x_2)) \leq B d_X(x_1, x_2)$$

(1) If a function f is bi-Lipschitz, then it is injective.

(2) In general the injectivity of a Lipschitz function does not imply the bi-Lipschitz property.

Phase retrievability implies bi-Lipschitz property

Frame

Definition

Let $\mathcal{H}$ be a n-dimensional Hilbert space. $\mathcal{F} = \{f_1, f_2, \ldots, f_m\} \subset \mathcal{H}$ is a frame for $\mathcal{H}$ if there exist constants $A, B > 0$ such that

$$A \|x\|^2 \leq \sum_{k=1}^{m} |\langle x, f_k \rangle|^2 \leq B \|x\|^2$$

for any $x \in \mathcal{H}$.

The measurement maps

Magnitude measurement map:

$$\alpha : \mathcal{H} \to \mathbb{R}^m, \qquad \alpha(x) = (|\langle x, f_k \rangle|)_{1 \le k \le m}$$

Square measurement map:

$$\beta : \mathcal{H} \to \mathbb{R}^m, \qquad \beta(x) = \left(|\langle x, f_k \rangle|^2\right)_{1 \le k \le m}$$

There is an ambiguity of a universal phase

$$\alpha(x) = \alpha(e^{i\phi} x), \qquad \forall \phi \in [0, 2\pi)\ .$$

The measurement maps

Consider the equivalence relation $\sim$ such that

$$x \sim y \qquad \text{iff} \qquad \exists a, |a| = 1 \;\text{ s.t. }\; y = ax$$

Let $\hat{\mathcal{H}}$ denote the collection of the equivalence classes.

Magnitude measurement map:

$$\alpha : \hat{\mathcal{H}} \to \mathbb{R}^m, \qquad \alpha(\hat{x}) = (|\langle x, f_k \rangle|)_{1 \leq k \leq m}\,, \quad x \in \hat{x}$$

Square measurement map:

$$\beta : \hat{\mathcal{H}} \to \mathbb{R}^m, \qquad \beta(\hat{x}) = \left(|\langle x, f_k \rangle|^2\right)_{1 \leq k \leq m}\,, \quad x \in \hat{x}$$

Definition

$\mathcal{F}$ is phase retrievable if α (or equivalently β) is injective.

Distance function

A stability result usually quantifies how the “output” changes with a small change in the “input”.

To measure how much the change is, we need to define reasonable distance functions

Definition

Natural metrics: for $1 \leq p \leq \infty$ and $\hat{x}, \hat{y} \in \hat{\mathcal{H}}$

$$D_p(\hat{x}, \hat{y}) = \min_{|a|=1} \|x - ay\|_p \ .$$

Matrix-norm induced metrics: for $1 \leq p \leq \infty$ and $\hat{x}, \hat{y} \in \hat{\mathcal{H}}$

$$d_p(\hat{x}, \hat{y}) = \|xx^* - yy^*\|_p = \begin{cases} \left(\sum_{k=1}^{n} (\sigma_k)^p\right)^{1/p} & \textit{for} \quad 1 \leq p < \infty \\ \max_{1 \leq k \leq n} \ \sigma_k & \textit{for} \quad p = \infty \end{cases} ,$$

We are interested in D_2 corresponding to the Euclidean distance and d_1 corresponding to the nuclear norm.

The noisy measurement

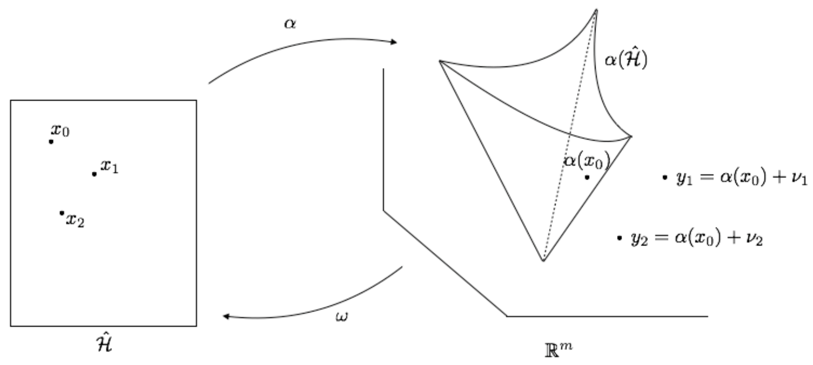

α
$\alpha(\hat{\mathcal{H}})$
x_0
x_1
x_2
$\alpha(x_0)$
$y_1 = \alpha(x_0) + \nu_1$
$y_2 = \alpha(x_0) + \nu_2$
ω
$\hat{\mathcal{H}}$
$\mathbb{R}^m$

Figure 4: Assume an additive noise model $y = \alpha(x) + \nu$. Suppose we have an inverse map $\omega : \mathbb{R}^m \to \hat{\mathcal{H}}$. Then for the reconstructed signal $x_1 = \omega(y_1)$ and $x_2 = \omega(y_2)$, $D_2(x_1, x_2) = D_2(\omega(y_1), \omega(y_2)) \leq \mathrm{Lip}(\omega) \cdot \|y_1 - y_2\| = \mathrm{Lip}(\omega) \cdot \|\nu_1 - \nu_2\|$.

The noisy measurement

Suppose we have an invertible measurement map.

We need a reconstruction map from $\mathbb{R}^m$ to $\hat{\mathcal{H}}$ that is Lipschitz.

We shall establish it in two steps.

1. For a phase retrievable map, show that it is bi-Lipschitz, that is, the inverse map is Lipschitz continuous.
2. The inverse map can be extended to $\mathbb{R}^m$.

Phase retrievability implies bi-Lipschitz property

Summary of results

Suppose the map is phase retrievable. The bi-Lipschitz property can be established in all the cases as follows:

	$\mathcal{H} = \mathbb{R}^n$	$\mathcal{H} = \mathbb{C}^n$
α	[Balan, Wang]	[Balan, Z.]
β	[Balan], [Bandeira et al.]	

It is relatively easy to find the upper Lipschitz bound in each case.

For α, the upper Lipschitz bound is given by the upper frame bound B.

For β, the upper Lipschitz bound is given by B^2.

The lower Lipschitz bound is more relevant to the stability of reconstruction.

Lipschitz bounds

Definition

1. The *global lower* and *upper Lipschitz bounds*, respectively:

$$A_0 = \inf_{x,y\in\hat{H}} \frac{\|\alpha(x)-\alpha(y)\|_2^2}{D_2(x,y)^2}\,, \qquad B_0 = \sup_{x,y\in\hat{H}} \frac{\|\alpha(x)-\alpha(y)\|_2^2}{D_2(x,y)^2}\,;$$

2. The *type I local lower* and *upper Lipschitz bounds* at $z \in \hat{H}$, respectively:

$$A(z) = \lim_{r\to 0} \inf_{\substack{x,y\in\hat{H}\\ D_2(x,z)<r\\ D_2(y,z)<r}} \frac{\|\alpha(x)-\alpha(y)\|_2^2}{D_2(x,y)^2}\,, \quad B(z) = \lim_{r\to 0} \sup_{\substack{x,y\in\hat{H}\\ D_2(x,z)<r\\ D_2(y,z)<r}} \frac{\|\alpha(x)-\alpha(y)\|_2^2}{D_2(x,y)^2}\,;$$

3. The *type II local lower* and *upper Lipschitz bounds* at $z \in \hat{H}$, respectively:

$$\tilde{A}(z) = \lim_{r\to 0} \inf_{\substack{x\in\hat{H}\\ D_2(x,z)<r}} \frac{\|\alpha(x)-\alpha(z)\|_2^2}{D_2(x,z)^2}\,, \quad \tilde{B}(z) = \lim_{r\to 0} \sup_{\substack{x\in\hat{H}\\ D_2(x,z)<r}} \frac{\|\alpha(x)-\alpha(z)\|_2^2}{D_2(x,z)^2}\,.$$

Lipschitz bounds

Definition

1. The *global lower* and *upper Lipschitz bounds*, respectively:

$$a_0 = \inf_{x,y\in\hat{H}} \frac{\|\beta(x)-\beta(y)\|_2^2}{d_1(x,y)^2}, \qquad b_0 = \sup_{x,y\in\hat{H}} \frac{\|\beta(x)-\beta(y)\|_2^2}{d_1(x,y)^2};$$

2. The *type I local lower* and *upper Lipschitz bounds* at $z\in\hat{H}$, respectively:

$$a(z) = \lim_{r\to 0} \inf_{\substack{x,y\in\hat{H}\\ d_1(x,z)<r\\ d_1(y,z)<r}} \frac{\|\beta(x)-\beta(y)\|_2^2}{d_1(x,y)^2}, \quad b(z) = \lim_{r\to 0} \sup_{\substack{x,y\in\hat{H}\\ d_1(x,z)<r\\ d_1(y,z)<r}} \frac{\|\beta(x)-\beta(y)\|_2^2}{d_1(x,y)^2};$$

3. The *type II local lower* and *upper Lipschitz bounds* at $z\in\hat{H}$, respectively:

$$\tilde{a}(z) = \lim_{r\to 0} \inf_{\substack{x\in\hat{H}\\ d_1(x,z)<r}} \frac{\|\beta(x)-\beta(z)\|_2^2}{d_1(x,z)^2}, \quad \tilde{b}(z) = \lim_{r\to 0} \sup_{\substack{x\in\hat{H}\\ d_1(x,z)<r}} \frac{\|\beta(x)-\beta(z)\|_2^2}{d_1(x,z)^2}.$$

Realification

- Consider the $\mathbb{R}$-linear map $\mathbf{j} : \mathbb{C}^n \to \mathbb{R}^{2n}$ defined by

$$\mathrm{j}(z) = \left[\begin{array}{c} \mathrm{real}(z) \\ \mathrm{imag}(z) \end{array} \right].$$

  Notation: $\xi = \mathrm{j}(x)$, $\eta = \mathrm{j}(y)$, $\zeta = \mathrm{j}(z)$, $\varphi = \mathrm{j}(f)$, $\delta = \mathrm{j}(d)$.

- For a frame set $\mathcal{F} = \{f_1, f_2, \cdots, f_m\}$, define the symmetric operator

$$\Phi_k = \varphi_k \varphi_k^T + J \varphi_k \varphi_k^T J^T, \quad k = 1, 2, \cdots, m.$$

  where

$$J = \left[\begin{array}{cc} 0 & -I \\ I & 0 \end{array} \right]$$

- Define $\mathcal{S} : \mathbb{R}^{2n} \to \mathrm{Sym}(\mathbb{R}^{2n})$ by

$$\mathcal{S}(\xi) = \begin{cases} 0 & , \text{ if } \quad \xi = 0 \\ \sum_{k:\Phi_k \xi \neq 0} \frac{1}{\langle \Phi_k \xi, \xi \rangle} \Phi_k \xi \xi^T \Phi_k & , \text{ if } \quad \xi \neq 0 \end{cases}.$$

Lower Lipschitz constant in case $\mathcal{H} = \mathbb{C}^n$

Theorem

Let $\mathcal{F} \subset \mathbb{C}^n$ be a phase retrievable frame for $\mathbb{C}^n$. Let A and B denote its optimal lower and upper frame bound, respectively. For any $z \in \mathbb{C}^n$, let $\zeta = \mathbf{j}(z)$ be its realification. Then

1. *For every $0 \neq z \in \mathbb{C}^n$, $A(z) = \lambda_{2n-1}(\mathcal{S}(\zeta))$;*
2. $A_0 = A(0) > 0$;
3. *For every $z \in \mathbb{C}^n$, $\tilde{A}(z) = \lambda_{2n-1}\left(\mathcal{S}(\zeta) + \sum_{k:\langle z, f_k\rangle = 0} \Phi_k\right)$;*
4. $\tilde{A}(0) = A$;
5. *For every $z \in \mathbb{C}^n$, $B(z) = \tilde{B}(z) = \lambda_1\left(\mathcal{S}(\zeta) + \sum_{k:\langle z, f_k\rangle = 0} \Phi_k\right)$;*
6. $B_0 = B(0) = \tilde{B}(0) = B$.

Now we prove Part 1 and 2.

(For Part 3 & 5, the calculation is similar to Part 1. Part 4 & 6 are easy cases.)

Proof of Part 1

Denote

$$p(x,y) := \frac{\|\alpha(x)-\alpha(y)\|^2}{D_2(x,y)^2}, \qquad x,y\in\mathbb{C}^n,\ \hat{x}\neq\hat{y}.$$

Then

$$\begin{aligned} p(x,y) &= P(\xi,\eta) \\ &:= \frac{\sum_{k=1}^{m}\langle\Phi_k\xi,\xi\rangle+\langle\Phi_k\eta,\eta\rangle-2\sqrt{\langle\Phi_k\xi,\xi\rangle\,\langle\Phi_k\eta,\eta\rangle}}{\|\xi\|^2+\|\eta\|^2-2\sqrt{\langle\xi,\eta\rangle^2+\langle\xi,J\eta\rangle^2}}\ . \end{aligned}$$

Fix $r>0$. Take $\xi,\eta\in\mathcal{B}(\zeta,r)$. Let $\mu=\frac{\xi+\eta}{2}$ and $\nu=\frac{\xi-\eta}{2}$.

Proof of Part 1

$$
\begin{aligned}
P(\xi,\eta) &= \left(\sum_{k=1}^{m} \langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle - \sqrt{(\langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle)^2 - 4\langle \Phi_k\mu,\nu\rangle^2}\right)\cdot \\
&\qquad \left(\|\mu\|^2 + \|\nu\|^2 - \sqrt{\|\mu\|^4 + \|\nu\|^4 - 2\|\mu\|^2\|\nu\|^2 + 4\langle \mu, J\nu\rangle^2}\right)^{-1} \\
&\geq \left(\sum_{k:\Phi_k\zeta\neq 0} \langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle - \sqrt{(\langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle)^2 - 4\langle \Phi_k\mu,\nu\rangle^2}\right)\cdot \\
&\qquad \left(\|\mu\|^2 + \|\nu\|^2 - \sqrt{\|\mu\|^4 + \|\nu\|^4 - 2\|\mu\|^2\|\nu\|^2}\right)^{-1} \\
&= \frac{1}{2\|\nu\|^2} \sum_{k:\Phi_k\zeta\neq 0} \langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle - \sqrt{(\langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle)^2 - 4\langle \Phi_k\mu,\nu\rangle^2} \\
&= \frac{1}{2\|\nu\|^2} \sum_{k:\Phi_k\zeta\neq 0} \langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle - \langle \Phi_k\mu,\mu\rangle \left(1 + \frac{\langle \Phi_k\nu,\nu\rangle}{\langle \Phi_k\mu,\mu\rangle} - 2\frac{\langle \Phi_k\mu,\nu\rangle^2}{\langle \Phi_k\mu,\mu\rangle^2}\right) + O(\|\nu\|^4) \\
&= \sum_{k:\Phi_k\zeta\neq 0} \frac{\langle \Phi_k\mu,\nu\rangle^2}{\langle \Phi_k\mu,\mu\rangle\,\|\nu\|^2} + O(\|\nu\|^2) \\
&= \frac{1}{\|\nu\|^2}\langle \mathcal{S}(\mu)\nu,\nu\rangle + O(\|\nu\|^2)\,.
\end{aligned}
$$

Proof of Part 1

Examine μ and ν we have $\langle J\zeta, \nu\rangle = 0$ and

$$\left\|P_{J\mu}^{\perp}\nu\right\|^2 \geq \left(1 - \frac{r^2}{\|\mu\|^2}\right)\|\nu\|^2 \ .$$

Consequently

$$\begin{aligned} P(\xi,\eta) &= \frac{1}{\|\nu\|^2}\left\langle \mathcal{S}(\mu)P_{J\mu}^{\perp}\nu, P_{J\mu}^{\perp}\nu\right\rangle + O(\|\nu\|^2) \\ &\geq \frac{1}{\left\|P_{J\mu}^{\perp}\nu\right\|^2}\left\langle \mathcal{S}(\mu)P_{J\mu}^{\perp}\nu, P_{J\mu}^{\perp}\nu\right\rangle\left(1 - \frac{r^2}{\|\mu\|^2}\right) + O(r^2) \\ &\geq \left(1 - \frac{r^2}{\|\mu\|^2}\right)\lambda_{2n-1}\left(\mathcal{S}(\mu)\right) + O(r^2) \ . \end{aligned}$$

Take $r \to 0$, by the continuity of eigenvalues with respect to matrix entries we have that

$$A(z) \geq \lambda_{2n-1}(\mathcal{S}(\zeta)) \ .$$

Proof of Part 1

Take E_{2n-1} to be the unit-norm eigenvector correspondent to $\lambda_{2n-1}(\mathcal{S}(\zeta))$. For each $r > 0$, take $\xi = \zeta + \frac{r}{2}E_{2n-1}$ and $\eta = \zeta - \frac{r}{2}E_{2n-1}$. Then

$$p(x, y) = P(\xi, \eta) = \lambda_{2n-1}(\mathcal{S}(\zeta)) \ .$$

Hence

$$A(z) \leq \lambda_{2n-1}(\mathcal{S}(\zeta)) \ .$$

Combining both directions we have

$$A(z) = \lambda_{2n-1}(\mathcal{S}(\zeta)) \ .$$

Proof of Part 2

Proof by contradiction, compactness argument:

Assume $A_0 = 0$. Then for any $N \in \mathbb{N}$, $\exists\, x_N, y_N \in \mathbb{C}^n$ s.t.

$$p(x_N, y_N) = \frac{\|\alpha(x_N) - \alpha(y_N)\|^2}{D_2(x_N, y_N)^2} \leq \frac{1}{N}.$$

WLOG assume $1 = \|x_N\| \geq \|y_N\|, \forall N$.

By compactness of the closed ball $\mathcal{B}_1(0) = \{x \in H : \|x\| \leq 1\}$ in $\mathbb{C}^n$, there exist convergent subsequences of $\{x_N\}_{N\in\mathbb{N}}$ and $\{y_N\}_{N\in\mathbb{N}}$.

We write $\{x_N\}_{N\in\mathbb{N}} \to x_0 \in \mathbb{C}^N$ and $\{y_N\}_{N\in\mathbb{N}} \to y_0 \in \mathbb{C}^N$.

From Part 1, $A(x_0) > 0$. Also, $D_2(x_N, y_N) \leq 2$. So

$$\|\alpha(x_N) - \alpha(y_N)\| \to 0 \quad \text{and} \quad \|\alpha(x_0) - \alpha(y_0)\| = 0$$

By injectivity, $x_0 = y_0 \in \hat{\mathbb{C}}^n$. Hence $p(x_N, y_N) \geq A(x_0) - 1/N > 1/N$ for large N. Contradiction.

Q.E.D.

Global stable reconstruction

Extend the inverse map to $\mathbb{R}^m$

Recall:

We need a reconstruction map from $\mathbb{R}^m$ to $\hat{\mathcal{H}}$ that is Lipschitz.

We shall establish it in two steps.

1. For a phase retrievable map, show that it is bi-Lipschitz, that is, the inverse map is Lipschitz continuous.
2. The inverse map can be extended to $\mathbb{R}^m$.

As we shall see in the following slides, the Kirszbraun Theorem provides some conditions for an extend-able Lipschitz map.

Kirszbraun Theorem

Definition (The Kirszbraun Property (K))

Let X and Y be two metric spaces with metric d_X and d_Y respectively. (X, Y) is said to have Property (K) if for any pair of families of closed balls $\{B(x_i, r_i) : i \in I\}$, $\{B(y_i, r_i) : i \in I\}$, such that $d_Y(y_i, y_j) \leq d_X(x_i, x_j)$ for each $i, j \in I$, it holds that

$$\bigcap_{i \in I} B(x_i, r_i) \neq \emptyset \Rightarrow \bigcap_{i \in I} B(y_i, r_i) \neq \emptyset .$$

Theorem (Wells and Williams, Ch. 10)

Suppose X and Y are Hilbert spaces and d_X and d_Y are the metrics induced by the inner products in each space respectively. Then (X, Y) has Property (K).

Kirszbraun Theorem

Theorem (Kirszbraun Theorem)

Let X and Y be two metric spaces and (X, Y) has Property (K). Suppose U is a subset of X and $f : U \to Y$ is a Lipschitz map. Then there exists a Lipschitz map $F : X \to Y$ which extends f to X such that

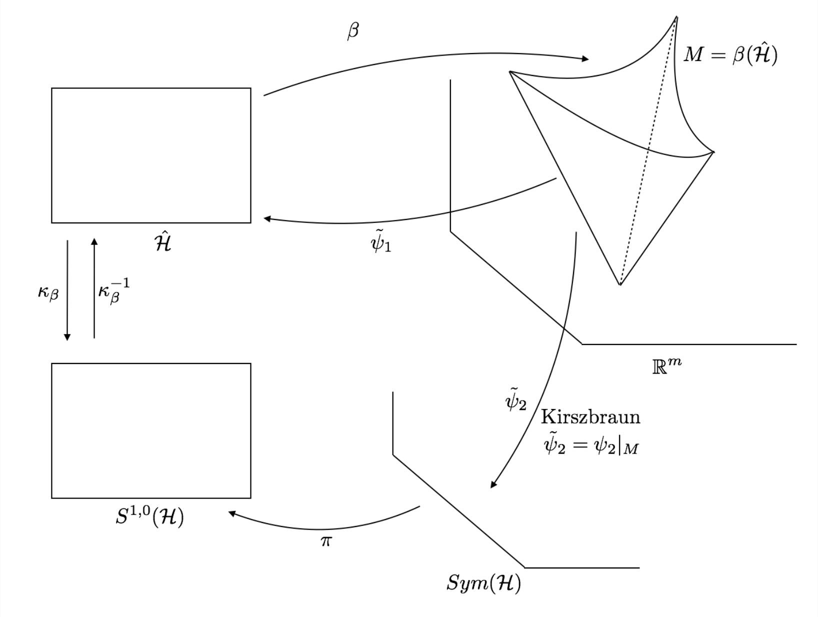
$$F|_U = f$$

and

$$\mathrm{Lip}(F) = \mathrm{Lip}(f)\ .$$

Unfortunately, $\left(\mathbb{R}^m, \hat{\mathcal{H}}\right)$ does not have Property (K).

General idea



The map $\pi : \mathrm{Sym}(\mathcal{H}) \to S^{1,0}(\mathcal{H})$

Lemma

Consider the spectral decomposition of any self-adjoint operator A in $\mathrm{Sym}(\mathcal{H})$, say $A = \sum_{k=1}^{d} \lambda_{m(k)} P_k$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the n eigenvalues including multiplicities, and P_1,...,P_d are the orthogonal projections associated to the d distinct eigenvalues. Additionally, $m(1) = 1$ and $m(k+1) = m(k) + r(k)$, where $r(k) = rank(P_k)$ is the multiplicity of eigenvalue $\lambda_{m(k)}$. Then the map

$$\pi : \mathrm{Sym}(\mathcal{H}) \to S^{1,0}(\mathcal{H}) \ \ , \ \ \pi(A) = (\lambda_1 - \lambda_2)P_1$$

satisfies the following two properties:

1. *for $1 \leq p \leq \infty$, π is Lipschitz continuous from $(\mathrm{Sym}(\mathcal{H}), \|\cdot\|_p)$ to $(S^{1,0}(\mathcal{H}), \|\cdot\|_p)$ with Lipschitz constant $\mathrm{Lip}(\pi) \leq 3 + 2^{1+\frac{1}{p}}$;*
2. *$\pi(A) = A$ for all $A \in S^{1,0}(\mathcal{H})$.*

Note: numerical experiments suggest a smaller $\mathrm{Lip}(\pi)$ is smaller. However, we can find an example with $\mathrm{Lip}(\pi) = 2$, so $\mathrm{Lip}(\pi) \geq 2$.

Proof

Let $A, B \in \mathrm{Sym}(\mathcal{H})$ where $A = \sum_{k=1}^{d} \lambda_{m(k)} P_k$ and $B = \sum_{k'=1}^{d'} \mu_{m(k')} Q_{k'}$. We now show that

$$\|\pi(A) - \pi(B)\|_p \leq (3 + 2^{1+\frac{1}{p}}) \|A - B\|_p \ .$$

WLOG, $0 < \lambda_1 - \lambda_2 \leq \mu_1 - \mu_2$ (other cases are easy).

$$(\lambda_1 - \lambda_2)P_1 - (\mu_1 - \mu_2)Q_1 = (\lambda_1 - \lambda_2)(P_1 - Q_1) + (\lambda_1 - \mu_1 - (\lambda_2 - \mu_2))Q_1 \ .$$

$$\|P_1\|_\infty = \|Q_1\|_\infty = 1 \quad \Rightarrow \quad \|P_1 - Q_1\|_\infty \leq 1 \quad \Rightarrow \quad \|P_1 - Q_1\|_p \leq 2^{\frac{1}{p}}$$

Weyl's inequality $\quad \Rightarrow \quad |\lambda_i - \mu_i| \leq \|A - B\|_\infty$ for each i.

$$\Rightarrow \quad |\lambda_1 - \mu_1| + |\lambda_2 - \mu_2| \leq 2\|A - B\|_\infty \leq 2\|A - B\|_p.$$

Let $g := \lambda_1 - \lambda_2$, $\delta := \|A - B\|_p$

$$\|\pi(A) - \pi(B)\|_p \leq g\|P_1 - Q_1\|_p + 2\delta \leq 2^{\frac{1}{p}} g + 2\delta \ .$$

If $0 < g \leq (2 + 2^{-\frac{1}{p}})\delta$, then $\|\pi(A) - \pi(B)\|_p \leq (2^{1+\frac{1}{p}} + 3)\delta$. Done.

Proof

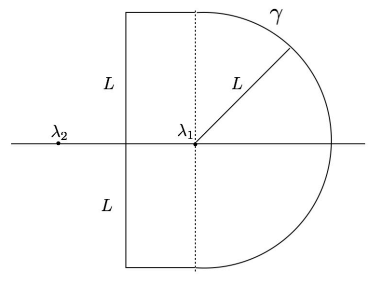
Case: $g > (2 + 2^{-\frac{1}{p}})\delta$.

$\delta < g/2 \qquad \Rightarrow \qquad |\lambda_1 - \mu_1| < g/2$ and $|\lambda_2 - \mu_2| < g/2$

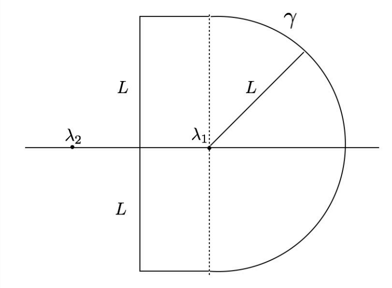
Use holomorphic functional calculus to rewrite:

$$P_1 = -\frac{1}{2\pi i}\oint_\gamma R_A dz, \quad Q_1 = -\frac{1}{2\pi i}\oint_\gamma R_B dz$$

where $R_A = (A - zI)^{-1}$, $R_B = (B - zI)^{-1}$, and $\gamma = \gamma(t)$ is the contour



Proof



$$\|P_1 - Q_1\|_p \leq \frac{1}{2\pi}\int_I \|(R_A - R_B)(\gamma(t))\|_p \, |\gamma'(t)| dt \, .$$

$$(R_A - R_B)(z) \; = \; R_A(z) - (I + R_A(z)(B - A))^{-1} R_A(z) \; = \; \sum_{n\geq 1} (-1)^n (R_A(z)(B - A))^n R_A(z) \, ,$$

$$\|R_A(z)(B - A)\|_\infty \; \leq \; \|R_A(z)\|_\infty \, \|B - A\|_p \; \leq \; \frac{\delta}{\mathrm{dist}(z, \sigma(A))} \; \leq \; \frac{2\delta}{g} < \frac{2}{2 + 2^{-\frac{1}{p}}} < 1 \, ,$$

Therefore

$$\|(R_A - R_B)(\gamma(t))\|_p \; \leq \; \sum_{n\geq 1} \|R_A(\gamma(t))\|_\infty^{n+1} \, \|A - B\|_p^n \; < \; \frac{\|A - B\|_p}{\mathrm{dist}^2(\gamma(t), \sigma(A))} \cdot (2^{1+\frac{1}{p}} + 1) \, .$$

Proof

Therefore,

$$\begin{aligned}\|P_1 - Q_1\|_p &\le (2^{\frac{1}{p}} + 2^{-1})\frac{\|A-B\|_p}{\pi}\int_I \frac{1}{\mathrm{dist}^2(\gamma(t),\sigma(A))}|\gamma'(t)|dt \\ &= (2^{\frac{1}{p}} + 2^{-1})\frac{\|A-B\|_p}{\pi}\cdot\frac{2\pi}{g} \\ &= (2^{1+\frac{1}{p}} + 1)\frac{\delta}{g}\ .\end{aligned}$$

Recall

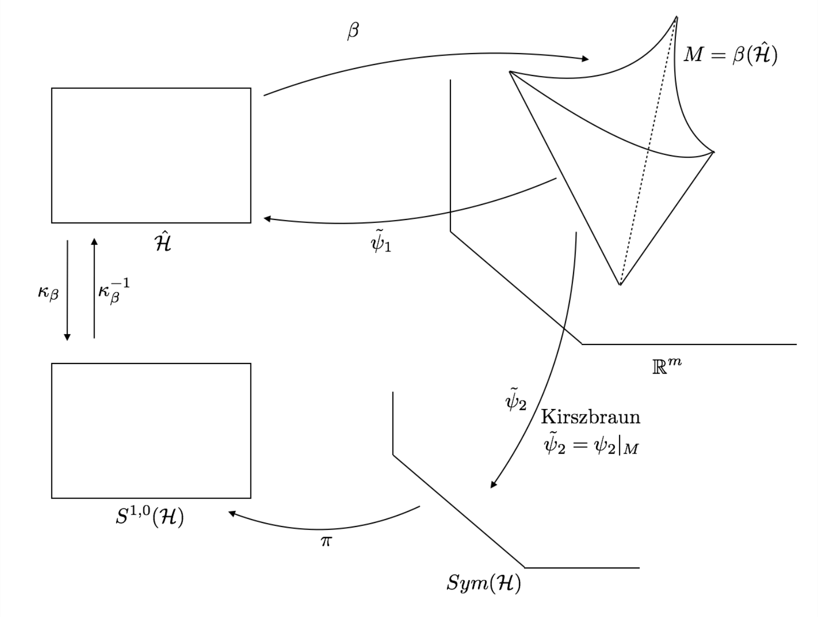
$$\|\pi(A) - \pi(B)\|_p \le g\,\|P_1 - Q_1\|_p + 2\delta\ .$$

We have

$$\|\pi(A) - \pi(B)\|_p \le (3 + 2^{1+\frac{1}{p}})\delta\ .$$

Q.E.D.

Global stable reconstruction



Global stable reconstruction

Need $\tilde{\psi}_2$, which is a composition of $\tilde{\psi}_1$ and an embedding κ_α (for α) or κ_β (for β).

$$\kappa_\alpha : \hat{\mathcal{H}} \to S^{1,0}(\mathcal{H}) \subset \mathrm{Sym}(\mathcal{H}) \ \ , \ \ \kappa_\alpha(x) = \begin{cases} \frac{1}{\|x\|} xx^* & \textit{if} \quad x \neq 0 \\ 0 & \textit{if} \quad x = 0 \end{cases} .$$

$$\kappa_\beta : \hat{\mathcal{H}} \to S^{1,0}(\mathcal{H}) \subset \mathrm{Sym}(\mathcal{H}) \ \ , \ \ \kappa_\beta(x) = xx^*.$$

κ_α is bi-Lipschitz; κ_β is an isometry.

According to the picture to combine all the maps, plus change of norms, we have a theorem.

Global stable reconstruction

Theorem

Let $\mathcal{F} = \{f_1, \ldots, f_m\}$ be a phase retrievable frame for the n dimensional Hilbert space $\mathcal{H}$ Let A_0 and a_0 denote the lower Lipschitz constant for α and β, respectively. Then

1. *there exists a Lipschitz continuous function $\omega : \mathbb{R}^m \to \hat{\mathcal{H}}$ s.t. $\omega(\alpha(x)) = x$ for all $x \in \hat{\mathcal{H}}$. For any $1 \leq p, q \leq \infty$, ω has a Lipschitz constant $\mathrm{Lip}(\omega)_{p,q}$ between $(\mathbb{R}^m, \|\cdot\|_p)$ and $(\hat{\mathcal{H}}, D_q)$ bounded by:*

$$\mathrm{Lip}(\omega)_{p,q} \leq \begin{cases} \frac{3\sqrt{2}+4}{\sqrt{A_0}} \cdot 2^{\frac{1}{q}-\frac{1}{2}} \cdot \max(1, m^{\frac{1}{2}-\frac{1}{p}}) & \textit{for } q \leq 2; \\ \frac{3\sqrt{2}+2^{\frac{3}{2}+\frac{1}{q}}}{\sqrt{A_0}} \cdot n^{\frac{1}{2}-\frac{1}{q}} \cdot \max(1, m^{\frac{1}{2}-\frac{1}{p}}) & \textit{for } q > 2. \end{cases}$$

2. *there exists a Lipschitz continuous function $\psi : \mathbb{R}^m \to \hat{\mathcal{H}}$ s.t. $\psi(\beta(x)) = x$ for all $x \in \hat{\mathcal{H}}$. For any $1 \leq p, q \leq \infty$, ψ has a Lipschitz constant $\mathrm{Lip}(\psi)_{p,q}$ between $(\mathbb{R}^m, \|\cdot\|_p)$ and $(\hat{\mathcal{H}}, d_q)$ bounded by:*

$$\mathrm{Lip}(\psi)_{p,q} \leq \begin{cases} \frac{3+2\sqrt{2}}{\sqrt{a_0}} \cdot 2^{\frac{1}{q}-\frac{1}{2}} \cdot \max(1, m^{\frac{1}{2}-\frac{1}{p}}) & \textit{for } q \leq 2; \\ \frac{3+2^{1+\frac{1}{q}}}{\sqrt{a_0}} \max(1, m^{\frac{1}{2}-\frac{1}{p}}) & \textit{for } q > 2. \end{cases}$$

A framework of CNN for stability analysis

Why stability analysis

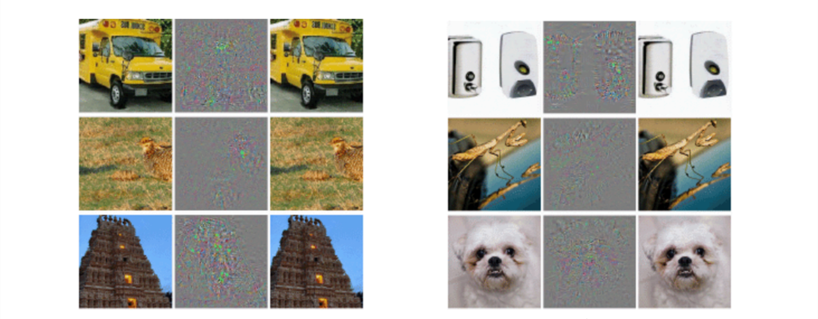


Figure 5: The adversarial examples given in [Szegedy et al.]. In each group (row) of pictures, the picture on the left is correctly labeled by AlexNet, the picture on the right is labeled wrong as "ostrich", and the picture in the middle show their difference.

[Szegedy et al.] compute the frame bound for each layer of AlexNet, and conclude that the upper bound is large in each layer.

Why stability analysis

Scattering networks are mathematically shown to be stable, but its filters are not subject to learning.

Case study:

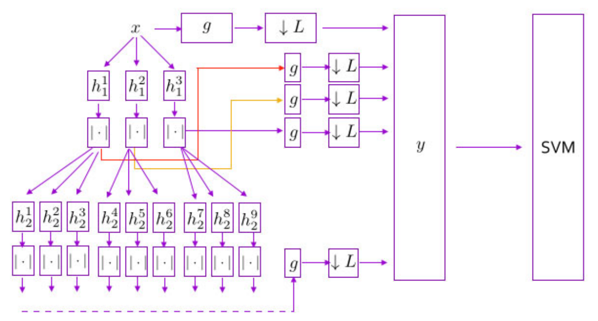


Figure 6: The structure of the scattering network for our case study. x is the input signal; h_k^j's are the convolutional filters taken to be the dilation of a Morlet wavelet with trained scales; g is the pooling function followed by a downsampling factor L; the feature y goes through a linear SVM to generate the classification result.

Why stability analysis

The optimization problem for learning is

$$\min_{\boldsymbol{\lambda};w,b} \quad \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N} l(y_n, a_n; w, b) \ ,$$

where

$$l(y, a; w, b) = \max(0, 1 - a(b + \langle w, y \rangle)) \ ,$$

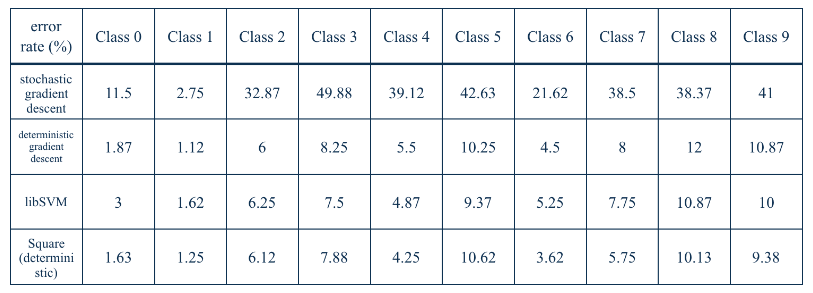
and y is the vector composed of the following vectors:

$$\begin{aligned} y_0 &= \ x * g \ ; \\ y_1^j &= \ \left| x * h_1^j \right| * g \ , 1 \le j \le 3 \ ; \\ y_2^j &= \ \left| \left| x * h_1^{\lceil j/3 \rceil} \right| * h_2^j \right| * g \ , 1 \le j \le 9 \ . \end{aligned}$$

where

$$h_k^j(t_1, t_2) = \psi_{\lambda_{k,1}^j} \otimes \psi_{\lambda_{k,2}^j}(t_1, t_2) = \lambda_{k,1}^j \lambda_{k,2}^j \psi(\lambda_{k,1}^j t_1)\psi(\lambda_{k,2}^j t_2) \ .$$

Why stablility analysis



error rate (%)	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
stochastic gradient descent	11.5	2.75	32.87	49.88	39.12	42.63	21.62	38.5	38.37	41
deterministic gradient descent	1.87	1.12	6	8.25	5.5	10.25	4.5	8	12	10.87
libSVM	3	1.62	6.25	7.5	4.87	9.37	5.25	7.75	10.87	10
Square (deterministic)	1.63	1.25	6.12	7.88	4.25	10.62	3.62	5.75	10.13	9.38

Figure 7: The classification results for MNIST. The error rate shows the percentage of data correctly labeled. The first row shows the results using the stochastic gradient descent method, the second row shows the results using the deterministic gradient descent method, the third row shows the results using libSVM, the fourth row shows the results where $|\cdot|$ is replaced by $|\cdot|^2$.

Want: a general framework for studying stability properties.

General structure of a CNN

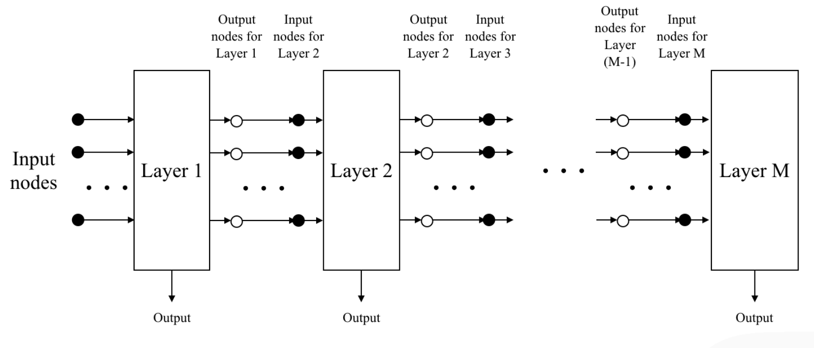
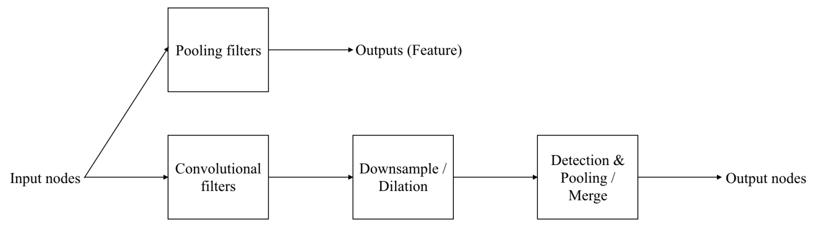


Figure 8: The detail of an M-layer ConvNet. The signals at output nodes are identical as at input nodes in the next layer. There may or may not be output in each layer.

One layer



1. The input nodes are signals from the output nodes in the previous layer (it is the input of the whole network for the first layer).
2. The convolutional filters are the filters that perform convolution with the signal from the input nodes.
3. The detection / merge operations are nonlinear operations applied pointwise to the output of the convolutional filters.
4. The pooling filters lower the dimensionality to generate the outputs.
5. The output nodes are signals that are passed to the next layer. The signal on the output nodes is identical to that on the input nodes of the next layer.

Model

Signals are taken from $L^2(\mathbb{R}^d)$.

Filters are taken from $\mathcal{B} = \{f \in \mathcal{S}'(\mathbb{R}^d), \left\|\hat{f}\right\|_\infty < \infty\}$.

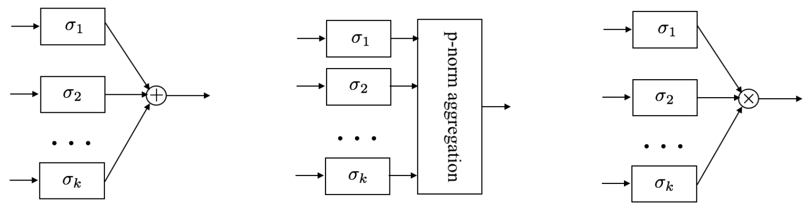


Figure 9: The three types of merge. Type I is taking sum of the inputs, Type II is taking p-norm aggregation, Type III is taking pointwise product.

Pooling

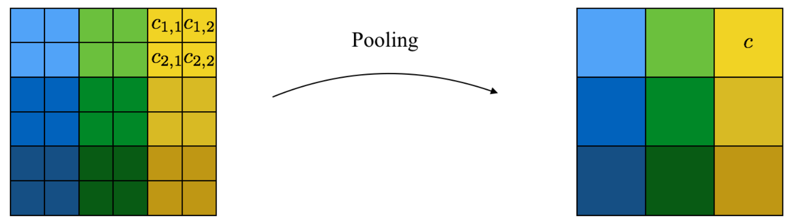


Figure 10: A toy example that shows how pooling works. The left image is subdivided into nine regions. The pooling operation outputs one value for each region. We take the top right corner for example. In the case of max pooling, we have $c = \max\{c_{1,1}, c_{1,2}, c_{2,1}, c_{2,2}\}$; in the case of average pooling, we have $c = (c_{1,1} + c_{1,2} + c_{2,1} + c_{2,2})/4$.

Pooling

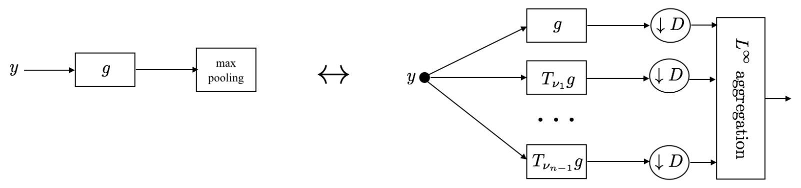


Figure 11: Max pooling modeled as Type II aggregation using L^∞ norm.

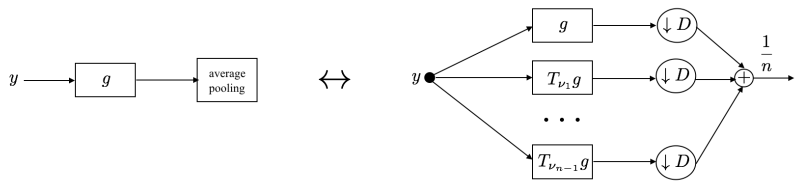


Figure 12: Average pooling modeled as Type I aggregation.

Details of the m-th layer

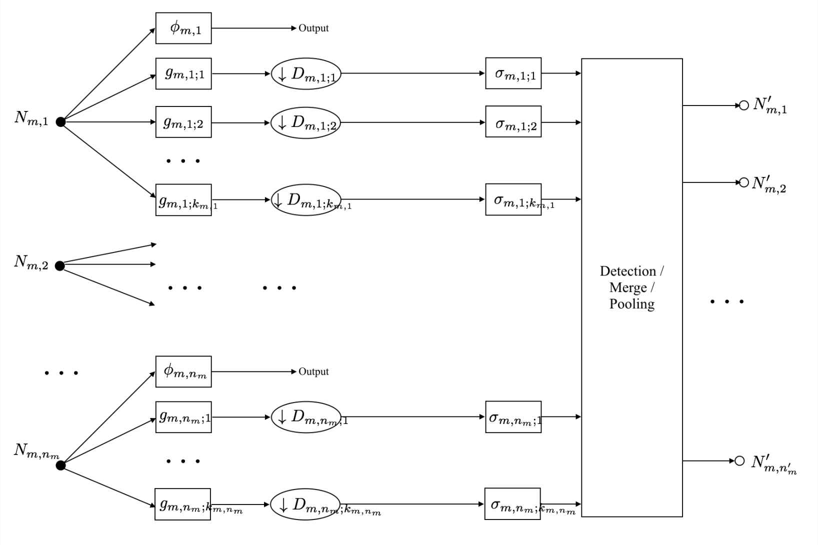


Figure 13: The detail of one layer. N's are the input nodes, N''s are the output nodes. ϕ's and g's are the filters, D's are the dilation factors. σ's are the nonlinearities.

Details of the m-th layer

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

$$l_{m,n;k} = \begin{cases} K & \text{, if } g_{m,n;k} \in \tau_1 \cup \tau_3 \\ K^{\max\{0,2/p-1\}} & \text{, if } g_{m,n;k} \in \tau_2 \end{cases}$$

We define the 1st type Bessel bound for the node $N_{m,n}$ to be

$$B^{(1)}_{m,n} = \left\| \left|\hat{\phi}_{m,n}\right|^2 + \sum_{k=1}^{k_{m,n}} l_{m,n;k} D^{-d}_{m,n;k} \left|\hat{g}_{m,n;k}\right|^2 \right\|_\infty ,$$

the 2nd type Bessel bound to be

$$B^{(2)}_{m,n} = \left\| \sum_{k=1}^{k_{m,n}} l_{m,n;k} D^{-d}_{m,n;k} \left|\hat{g}_{m,n;k}\right|^2 \right\|_\infty ,$$

and the generating bound to be

$$B^{(3)}_{m,n} = \left\| \hat{\phi}_{m,n} \right\|^2_\infty .$$

Details of the m-th layer

Then we define the 1st type Bessel bound for the m-th layer to be

$$B_m^{(1)} = \max_{1 \leq n \leq n_m} B_{m,n}^{(1)} ,$$

the 2nd type Bessel bound to be

$$B_m^{(2)} = \max_{1 \leq n \leq n_m} B_{m,n}^{(2)} ,$$

and the generating bound to be

$$B_m^{(3)} = \max_{1 \leq n \leq n_m} B_{m,n}^{(3)} .$$

Lipschitz bound

For any input signal f and $\tilde{f}$. Let f_N be the output for f from the node N, and $\tilde{f}_N$ be the output for $\tilde{f}$ from the node N. Let V be the collection of all nodes. We say L is a Lipschitz bound for the CNN if

$$\sum_{N \in V} \left\| f_N - \tilde{f}_N \right\|_2^2 \leq L \left\| f - \tilde{f} \right\|_2^2 .$$

Define the map $\Phi : L^2(\mathbb{R}^d) \to [L^2(\mathbb{R}^d)]^{|V|}$ by

$$\Phi(f) = (f_N)_{N \in V} .$$

Then a norm $|||\cdot|||$ defined on $[L^2(\mathbb{R}^d)]^{|V|}$ by

$$\left|\left|\left| (f_N)_{N \in V} \right|\right|\right| = \left(\sum_{N \in V} \|f_N\|_2^2 \right)^{1/2}$$

is well defined and $\sqrt{L}$ is a Lipschitz constant in the sense that

$$\left|\left|\left| \Phi(f) - \Phi(\tilde{f}) \right|\right|\right| \leq \sqrt{L} \left\| f - \tilde{f} \right\|_2 .$$

Lipschitz bounds for CNN

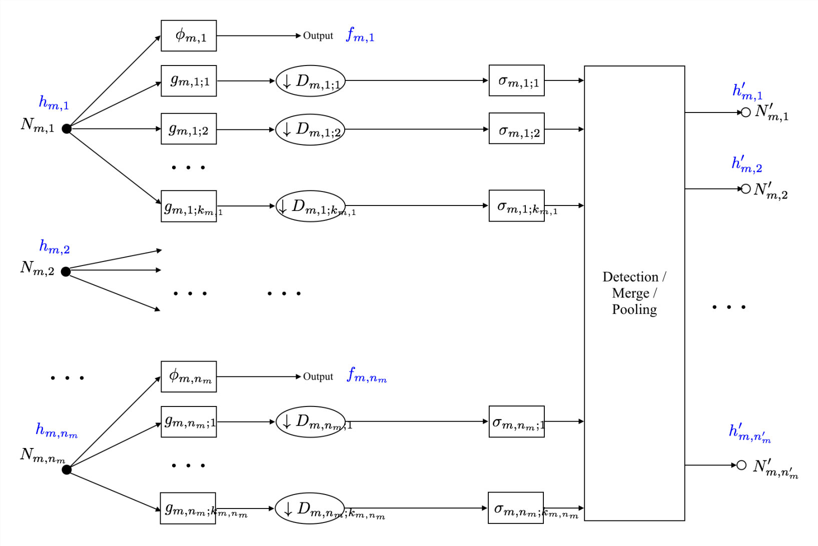
Lipschitz bound

Theorem

Consider a CNN with M layers and in the m-th layer it has 1st type Bessel bound $B_m^{(1)}$, 2nd type Bessel bound $B_m^{(2)}$ and generating bound $B_m^{(3)}$. Then the CNN implies a nonlinear map that is Lipschitz continuous, and its Lipschitz bound is given by the optimal value of the following linear program:

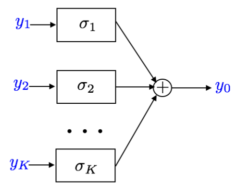
$$
\begin{aligned}
\max \quad & \sum_{m=1}^{M} z_m \\
s.t. \quad & y_0 = 1 \\
& y_m + z_m \leq B_m^{(1)} y_{m-1}, \quad 1 \leq m \leq M-1 \\
& y_m \leq B_m^{(2)} y_{m-1}, \quad 1 \leq m \leq M-1 \\
& z_m \leq B_m^{(3)} y_{m-1}, \quad 1 \leq m \leq M \\
& y_m, z_m \geq 0, \quad \textit{for all} \quad m
\end{aligned}
$$

Proof


$h_{m,1}$
$N_{m,1}$
$\phi_{m,1}$
Output $f_{m,1}$
$g_{m,1;1}$
$\downarrow D_{m,1;1}$
$\sigma_{m,1;1}$
$g_{m,1;2}$
$\downarrow D_{m,1;2}$
$\sigma_{m,1;2}$
$g_{m,1;k_{m,1}}$
$\downarrow D_{m,1;k_{m,1}}$
$\sigma_{m,1;k_{m,1}}$
$h_{m,2}$
$N_{m,2}$
ϕ_{m,n_m}
Output f_{m,n_m}
h_{m,n_m}
N_{m,n_m}
$g_{m,n_m;1}$
$\downarrow D_{m,n_m;1}$
$\sigma_{m,n_m;1}$
$g_{m,n_m;k_{m,n_m}}$
$\downarrow D_{m,n_m;k_{m,n_m}}$
$\sigma_{m,n_m;k_{m,n_m}}$
Detection / Merge / Pooling
$h'_{m,1}$
$N'_{m,1}$
$h'_{m,2}$
$N'_{m,2}$
h'_{m,n'_m}
N'_{m,n'_m}

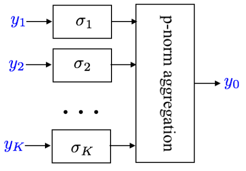
Proof

Study three types of merging:



$$\begin{aligned}
\|y_0 - \tilde{y}_0\|_2^2 &= \left\| \sum_{k=1}^{K} \sigma_k(y_k) - \sigma_k(\tilde{y}_k) \right\|_2^2 \\
&\leq K \sum_{k=1}^{K} \|\sigma_k(y_k) - \sigma_k(\tilde{y}_k)\|_2^2 \\
&\leq K \sum_{k=1}^{K} \|y_k - \tilde{y}_k\|_2^2 \ .
\end{aligned}$$

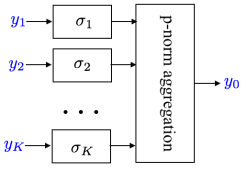
Proof



If $p \leq 2$,

$$\begin{aligned}
\|y_0 - \tilde{y}_0\|_2^2 &= \left\| \left(\sum_{k=1}^{K} |\sigma_k(y_k)|^p \right)^{1/p} - \left(\sum_{k=1}^{K} |\sigma_k(\tilde{y}_k)|^p \right)^{1/p} \right\|_2^2 \\
&\leq \left\| \left(\sum_{k=1}^{K} |\sigma_k(y_k) - \sigma_k(\tilde{y}_k)|^p \right)^{1/p} \right\|_2^2 \\
&\leq K^{2/p-1} \cdot \left\| \left(\sum_{k=1}^{K} |\sigma_k(y_k) - \sigma_k(\tilde{y}_k)|^2 \right)^{1/2} \right\|_2^2 \\
&= K^{2/p-1} \cdot \sum_{k=1}^{K} \|\sigma_k(y_k) - \sigma_k(\tilde{y}_k)\|_2^2 \leq K^{2/p-1} \cdot \sum_{k=1}^{K} \|y_k - \tilde{y}_k\|_2^2 \ ;
\end{aligned}$$

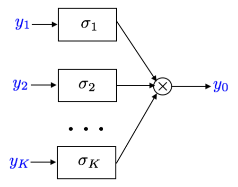
Proof



if $p > 2$,

$$
\begin{aligned}
\|y_0 - \tilde{y}_0\|_2^2 &= \left\| \left(\sum_{k=1}^{K} |\sigma_k(y_k)|^p \right)^{1/p} - \left(\sum_{k=1}^{K} |\sigma_k(\tilde{y}_k)|^p \right)^{1/p} \right\|_2^2 \\
&\leq \left\| \left(\sum_{k=1}^{K} |\sigma_k(y_k) - \sigma_k(\tilde{y}_k)|^p \right)^{1/p} \right\|_2^2 \\
&\leq \left\| \left(\sum_{k=1}^{K} |\sigma_k(y_k) - \sigma_k(\tilde{y}_k)|^2 \right)^{1/2} \right\|_2^2 \\
&= \sum_{k=1}^{K} \|\sigma_k(y_k) - \sigma_k(\tilde{y}_k)\|_2^2 \leq \sum_{k=1}^{K} \|y_k - \tilde{y}_k\|_2^2 \ .
\end{aligned}
$$

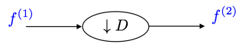
Proof



$$
\begin{aligned}
\|y_0 - \tilde{y}_0\|_2 \;&=\; \left\| \prod_{k=1}^{K} \sigma_k(y_k) - \prod_{k=1}^{K} \sigma_k(\tilde{y}_k) \right\|_2 \\
= \left\| \prod_{k=1}^{K} \sigma_k(y_k) + \sum_{J=1}^{K-1} \Big[- \prod_{k=1}^{J} \sigma_k(y_k) \prod_{k=J+1}^{K} \sigma_k(\tilde{y}_k) + \prod_{k=1}^{J} \sigma_k(y_k) \prod_{k=J+1}^{K} \sigma_k(\tilde{y}_k) \Big] - \prod_{k=1}^{K} \sigma_k(\tilde{y}_k) \right\|_2 \\
= \left\| \prod_{k=1}^{K-1} \sigma_k(y_k) \cdot (\sigma_K(y_K) - \sigma_K(\tilde{y}_K)) + \sum_{J=2}^{K-1} \prod_{k=1}^{J-1} \sigma_k(y_k) \cdot (\sigma_J(y_J) - \sigma_J(\tilde{y}_J)) \cdot \prod_{k=J+1}^{K} \sigma_k(\tilde{y}_k) + (\sigma_1(y_1) - \sigma_1(\tilde{y}_1)) \cdot \prod_{k=2}^{K} \sigma_k(\tilde{y}_k) \right\|_2 \\
\le \prod_{k=1}^{K-1} \|\sigma_k(y_k)\|_\infty \cdot \|\sigma_K(y_K) - \sigma_K(\tilde{y}_K)\|_2 + \sum_{J=2}^{K-1} \prod_{k=1}^{J-1} \|\sigma_k(y_k)\|_\infty \cdot \prod_{k=J+1}^{K} \|\sigma_k(\tilde{y}_k)\|_\infty \cdot \left\|\sigma_J(y_J) - \sigma_J(\tilde{y}_J)\right\|_2 + \\
\prod_{k=2}^{K} \|\sigma_k(\tilde{y}_k)\|_\infty \cdot \|\sigma_1(y_1) - \sigma_1(\tilde{y}_1)\|_2 \\
\le \sum_{k=1}^{K} \|\sigma_k(y_k) - \sigma_k(\tilde{y}_k)\|_2 \;\le\; \sum_{k=1}^{K} \|y_k - \tilde{y}_k\|_2 \;,
\end{aligned}
$$

and thus $\|y_0 - \tilde{y}_0\|_2^2 \le K \sum_{k=1}^{K} \|y_k - \tilde{y}_k\|_2^2$.

Proof



$$\begin{aligned}
\left\| f^{(2)} - \tilde{f}^{(2)} \right\|_2^2 &= \int \left| f^{(1)}(Dx) - \tilde{f}^{(1)}(Dx) \right|^2 dx \\
&= \frac{1}{D^d} \int \left| f^{(1)}(x) - \tilde{f}^{(1)}(x) \right|^2 dx \\
&= \frac{1}{D^d} \left\| f^{(1)} - \tilde{f}^{(1)} \right\|_2^2 .
\end{aligned}$$

Combine the above and compare with the definition of $B^{(1)}$, we have

$$\sum_{1}^{n'_m} \left\| h'_{m,n} - \tilde{h}'_{m,n} \right\|_2^2 + \sum_{n=1}^{n_m} \| f_{m,n} - f'_{m,n} \|_2^2 \;\leq\; B_m^{(1)} \left\| h_{m,n} - \tilde{h}_{m,n} \right\|_2^2 .$$

Also

$$\sum_{n=1}^{n_{m+1}} \left\| h_{m+1,n} - \tilde{h}_{m+1,n} \right\|_2^2 = \sum_{n=1}^{n'_m} \left\| h'_{m,n} - \tilde{h}'_{m,n} \right\|_2^2 .$$

Therefore,

$$\sum_{n=1}^{n_{m+1}} \left\| h_{m+1,n} - \tilde{h}_{m+1,n} \right\|_2^2 + \sum_{n=1}^{n_m} \left\| f_{m,n} - \tilde{f}_{m,n} \right\|_2^2 \;\leq\; B_m^{(1)} \sum_{n=1}^{n_m} \left\| h_{m,n} - \tilde{h}_{m,n} \right\|_2^2 .$$

Proof

Similarly,

$$\sum_{n=1}^{n_m} \left\| h_{m+1,n} - \tilde{h}_{m+1,n} \right\|_2^2 \leq B_m^{(2)} \sum_{n=1}^{n_m} \left\| h_{m,n} - \tilde{h}_{m,n} \right\|_2^2 \, ,$$

and

$$\sum_{n=1}^{n_m} \left\| f_{m,n} - \tilde{f}_{m,n} \right\|_2^2 \leq B_m^{(3)} \sum_{n=1}^{n_m} \left\| h_{m,n} - \tilde{h}_{m,n} \right\|_2^2 \, .$$

Now to determine a Lipschitz bound, we just need to solve the linear program

$$\begin{aligned} \max \quad & \sum_{m=1}^{M} z_m \\ \text{s.t.} \quad & y_0 = 1 \\ & y_m + z_m \leq B_m^{(1)} y_{m-1}, \quad 1 \leq m \leq M-1 \\ & y_m \leq B_m^{(2)} y_{m-1}, \quad 1 \leq m \leq M-1 \\ & z_m \leq B_m^{(3)} y_{m-1}, \quad 1 \leq m \leq M \\ & y_m, z_m \geq 0, \quad \text{for all} \quad m \end{aligned}$$

Q.E.D.

An easy estimate

Corollary

Consider a CNN with M layers and in the m-th layer it has 1st type Bessel bound $B_m^{(1)}$. Then the CNN induces a nonlinear map that is Lipschitz continuous, and its Lipschitz bound is given by

$$\prod_{m=1}^{M} \max\{1, B_m^{(1)}\} .$$

Moreover, if $B_m^{(1)} = B_m^{(3)}$ for each m, then this bound coincide with the output of the linear program.

Proof

Note that if $\{y_m\}_{m=0}^{M-1}$ and $\{z_m\}_{m=0}^{M-1}$ are the maximums of the linear program, then

$$z_m \leq B_m^{(1)} y_{m-1} - y_m, \qquad 1 \leq m \leq M-1,$$

and

$$z_M \leq B_M^{(1)} y_{M-1}.$$

Take the sum over all m's (denote $y_M = 0$),

$$\sum_{m=1}^{M} z_m \;\leq\; \sum_{m=1}^{M} B_m^{(1)} y_{m-1} - y_m \;=\; \sum_{m=0}^{M-1} B_{m+1}^{(1)} y_m - \sum_{m=1}^{M-1} y_m \;=\; B_1^{(1)} + \sum_{m=1}^{M-1} (B_{m+1}^{(1)} - 1) y_m \;.$$

Also, $y_m \leq B_m^{(2)} y_{m-1}$ implies $y_m \leq B_m^{(1)} y_{m-1}$, so

$$\begin{aligned} \sum_{m=1}^{M} z_m \;&\leq\; B_1^{(1)} + \sum_{m=1}^{M-1} (\max\{1, B_{m+1}^{(1)}\} - 1) \cdot \prod_{m'=1}^{m} \max\{1, B_{m'}^{(1)}\} \\ &= \prod_{m=1}^{M} \max\{1, B_m^{(1)}\} \;. \end{aligned}$$

Q.E.D.

Implication for stationary processes

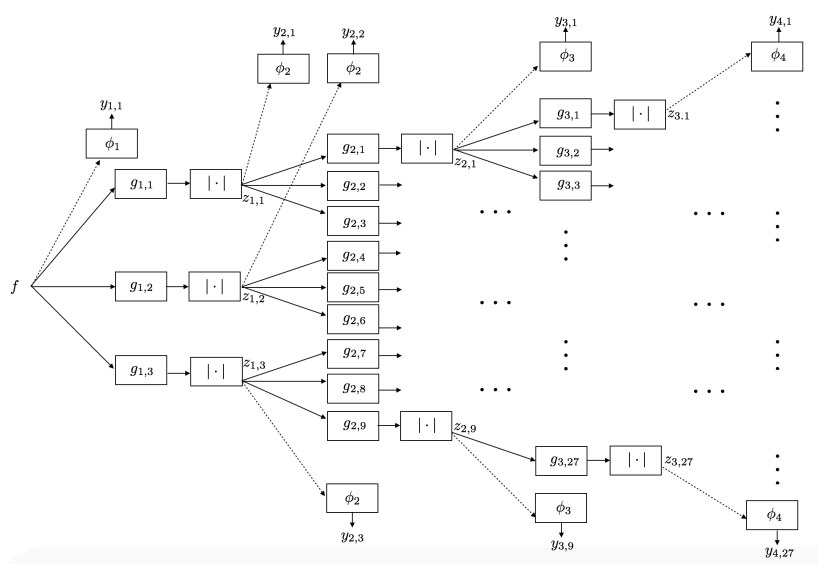
Theorem

Assume there is no dilation in CNN. Let X and Y be Strict-sense-stationary processes with finite second-order moments. Then

$$\mathbb{E}\left(\left|\left|\left|\Phi(X) - \Phi(Y)\right|\right|\right|^2\right) \leq L \cdot \mathbb{E}\left(|X - Y|^2\right) \ ,$$

where L is the Lipschitz constant associated with the CNN. In particular, $|||\Phi(X)|||^2 \leq L \cdot \mathbb{E}\left(|X|^2\right)$.

Example: a four-layer scattering network



Example: a four-layer scattering network

We consider the 1D case and the wavelet given by the Haar wavelets

$$\phi(t) = \begin{cases} 1, & \text{if } 0 \leq t < 1 \\ 0, & \text{otherwise} \end{cases} \qquad \text{and} \qquad \psi(t) = \begin{cases} 1, & \text{if } 0 \leq t < 1/2 \\ -1, & \text{if } 1/2 \leq t < 1 \\ 0, & \text{otherwise} \end{cases}.$$

we have a four-layer convolutional network for which the filters are given by $g_{1,l_1}, l_1 \in \{1, 2, 3\}$, $g_{2,l_2}, l_2 \in \{1, \cdots, 9\}$ and $g_{3,l_3}, l_3 \in \{1, \cdots, 27\}$, where

$$g_{m,l} = \begin{cases} \psi, & \text{if} \quad \mathrm{mod}\ (l, 3) = 1; \\ \psi_{2^{-1}}, & \text{if} \quad \mathrm{mod}\ (l, 3) = 2; \\ \psi_{2^{-2}}, & \text{if} \quad \mathrm{mod}\ (l, 3) = 0. \end{cases}$$

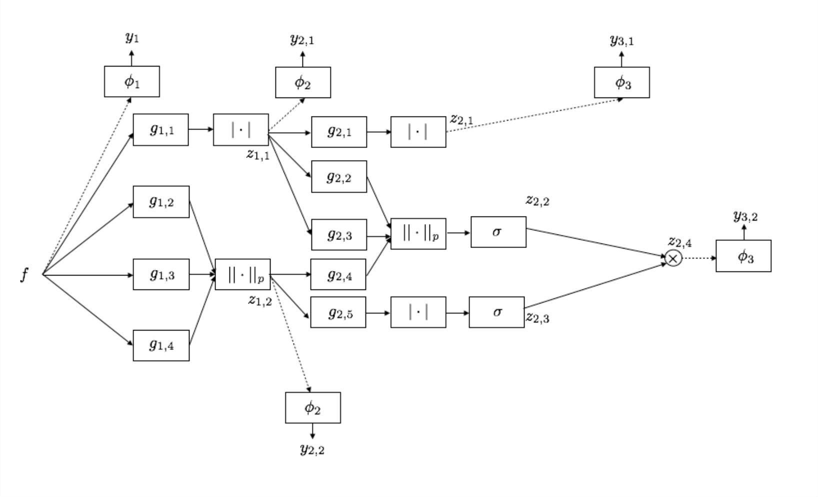
Also, for the output generation, $\phi_1 = \phi_2 = \phi_3 = \phi_4 = 2^{-3}\phi(2^{-3}\cdot)$.

Example: a four-layer scattering network

We use three approaches to estimate the Lipschitz constant.

1. Backpropagation: backtrack from outputs and estimate toward inputs
$$|||\Phi(f) - \Phi(\tilde{f})|||^2 = \sum_{m,l_m} \|y_{m,l_m} - \tilde{y}_{m,l_m}\|_2^2 \leq 40 \left\| f - \tilde{f} \right\|_2^2 .$$
2. Use the theorem: $B_1 = B_2 = B_3 = B_4 = 1$, we have
$$|||\Phi(f) - \Phi(\tilde{f})|||^2 \leq \left\| f - \tilde{f} \right\|_2^2 .$$
3. A lower bound is derived by considering only the output $y_{1,1}$ from the input layer. Obviously
$$|||\Phi(f) - \Phi(\tilde{f})|||^2 \geq \left\| (f - \tilde{f}) * \phi_1 \right\|_1^2 .$$
Thus
$$\sup_{f \neq \tilde{f}} \frac{|||\Phi(f) - \Phi(\tilde{f})|||^2}{\left\| f - \tilde{f} \right\|_2^2} \geq \sup_{f \neq \tilde{f}} \frac{\left\| (f - \tilde{f}) * \phi_1 \right\|_1^2}{\left\| f - \tilde{f} \right\|_2^2} = \left\| \hat{\phi}_1 \right\|_\infty^2 = 1 .$$
Therefore, 1 is the exact Lipschitz bound (and Lipschitz constant) in our example.

Example: a general CNN



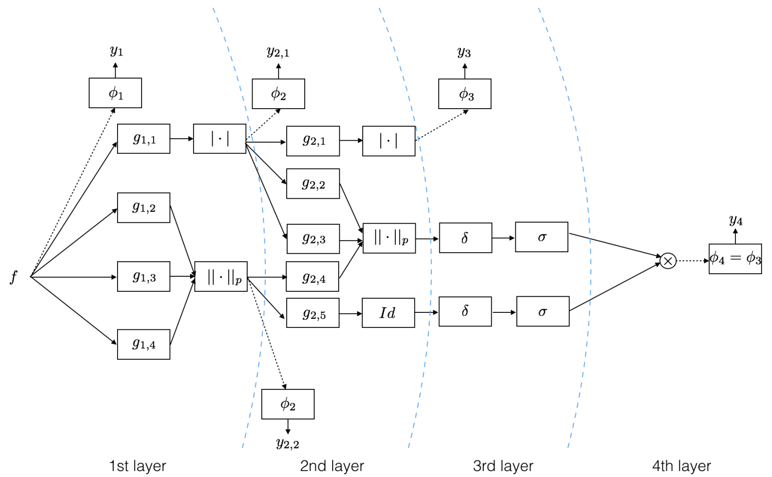
Example: a general CNN

First approach: backpropagation from the outputs

$$
\begin{aligned}
\left|\left|\left|\Phi(f) - \Phi(\tilde{f})\right|\right|\right|^2 &= \sum_{m,l} \|y_{m,l} - \tilde{y}_{m,l}\|_2^2 \\
&\leq \left\|f - \tilde{f}\right\|_2^2 \Bigg(\|\phi_1\|_1^2 + \|g_{1,1}\|_1^2 \|\phi_2\|_1^2 + \\
&\qquad (\|g_{1,2}\|_1 + \|g_{1,3}\|_1 + \|g_{1,4}\|_1)^2 \|\phi_2\|_1^2 + \\
&\qquad \|g_{1,1}\|_1^2 \|g_{2,1}\|_1^2 \|\phi_3\|_1^2 + \Big(\|g_{1,1}\|_1 (\|g_{2,2}\|_1 + \|g_{2,3}\|_1) + \\
&\qquad (\|g_{1,2}\|_1 + \|g_{1,3}\|_1 + \|g_{1,4}\|_1)(\|g_{2,4}\|_1 + \|g_{2,5}\|_1)\Big)^2 \|\phi_3\|_1^2 \Bigg) .
\end{aligned}
$$

Example: a general CNN

Second approach:



Example: a general CNN

Second approach:

$$\begin{aligned}
\tilde{B}_1 &= \left\| |\hat{g}_{1,1}|^2 + |\hat{g}_{1,2}|^2 + |\hat{g}_{1,3}|^2 + |\hat{g}_{1,4}|^2 + \left|\hat{\phi}_1\right|^2 \right\|_\infty ; \\
\tilde{B}_2 &= \max\left\{1, \left\| |\hat{g}_{2,1}|^2 + |\hat{g}_{2,2}|^2 + |\hat{g}_{2,3}|^2 + \left|\hat{\phi}_2\right|^2 \right\|_\infty, \left\| |\hat{g}_{2,4}|^2 + |\hat{g}_{2,5}|^2 + \left|\hat{\phi}_2\right|^2 \right\|_\infty \right\} ; \\
\tilde{B}_3 &= \max\left\{2, \left\|\hat{\phi}_3\right\|_\infty^2 \right\} ; \\
\tilde{B}_4 &= \max\left\{1, \left\|\hat{\phi}_3\right\|_\infty^2 \right\} .
\end{aligned}$$

Then the Lipschitz constant is given by $(\tilde{B}_1\tilde{B}_2\tilde{B}_3\tilde{B}_4)^{1/2}$, that is,

$$\left|\left|\left|\Phi(f) - \Phi(\tilde{f})\right|\right|\right|^2 \leq (\tilde{B}_1\tilde{B}_2\tilde{B}_3\tilde{B}_4) \left\|f - \tilde{f}\right\|_2^2 . \qquad (1)$$

Example: a general CNN

Define $F(\omega) = \exp(4\omega^2/(4\omega^2-1)) \cdot \chi_{(-1/2,0)}(\omega)$, and $G(\omega) = F(-\omega)$. The filters are defined in the Fourier domain to be

$$
\begin{aligned}
\hat{\phi}_1(\omega) &= F(\omega+1) + \chi_{(-1,1)}(\omega) + G(\omega-1) \\
\hat{g}_{1,1}(\omega) &= F(\omega+3) + \chi_{(-3,-2)}(\omega) + G(\omega+2) + F(\omega-2) + \chi_{(2,3)}(\omega) + G(\omega-3) \\
\hat{g}_{1,2}(\omega) &= F(\omega+5) + \chi_{(-5,-4)}(\omega) + G(\omega+4) + F(\omega-4) + \chi_{(4,5)}(\omega) + G(\omega-5) \\
\hat{g}_{1,3}(\omega) &= F(\omega+7) + \chi_{(-7,-6)}(\omega) + G(\omega+6) + F(\omega-6) + \chi_{(6,7)}(\omega) + G(\omega-7) \\
\hat{g}_{1,4}(\omega) &= F(\omega+9) + \chi_{(-9,-8)}(\omega) + G(\omega+8) + F(\omega-8) + \chi_{(8,9)}(\omega) + G(\omega-9) \\
\hat{\phi}_2(\omega) &= F(\omega+2) + \chi_{(-2,2)}(\omega) + G(\omega-2) \\
\hat{g}_{2,1}(\omega) &= F(\omega+4) + \chi_{(-4,-3)}(\omega) + G(\omega+3) + F(\omega-3) + \chi_{(3,4)}(\omega) + G(\omega-4) \\
\hat{g}_{2,2}(\omega) &= F(\omega+6) + \chi_{(-6,-5)}(\omega) + G(\omega+5) + F(\omega-5) + \chi_{(5,6)}(\omega) + G(\omega-6) \\
\hat{g}_{2,3}(\omega) &= F(\omega+8) + \chi_{(-8,-7)}(\omega) + G(\omega+7) + F(\omega-7) + \chi_{(7,8)}(\omega) + G(\omega-8) \\
\hat{g}_{2,4}(\omega) &= F(\omega+5) + \chi_{(-5,-3)}(\omega) + G(\omega+3) + F(\omega-3) + \chi_{(3,5)}(\omega) + G(\omega-5) \\
\hat{g}_{2,5}(\omega) &= F(\omega+8) + \chi_{(-8,-6)}(\omega) + G(\omega+6) + F(\omega-6) + \chi_{(6,8)}(\omega) + G(\omega-8) \\
\hat{\phi}_3(\omega) &= F(\omega+9) + \chi_{(-9,9)}(\omega) + G(\omega-9)
\end{aligned}
$$

Example: a general CNN

Define the filters as in the previous slide. Then we have the Lipschitz constant in the first approach is

$$\Gamma_1 = 31.1$$

and in the second approach is

$$\Gamma_2 = \sqrt{2}$$

Numerical experiment suggests that the constant is about

$$\Gamma_3 = 1.1937$$

Example: a general CNN

Using the same network, define

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

With that done, we define the filters in the Fourier domain to be

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

Example: a general CNN

For this setting,

$$B_m^{(1)} = 2\exp(-1/3)$$

$$B_m^{(2)} = B_m^{(3)} = 1$$

for each m.

The linear program outputs the optimal Lipschitz bound

$$L_l = 2.2992 \qquad \Gamma_l = \sqrt{L_l} = 1.5163$$

the Lipschitz bound given by the corollary is

$$L_c = 8[\exp(-1/3)]^3 = 2.9430 \qquad \Gamma_c = \sqrt{L_c} = 1.7155$$

Thank you!

References I

- R. Balan, *Reconstruction of signals from magnitudes of redundant representations: The complex case*, Foundations of Computational Mathematics, 16 (2016), pp. 677–721.
- R. Balan, P. Casazza, and D. Edidin, *On signal reconstruction without phase*, Applied and Computational Harmonic Analysis, 20 (2006), pp. 345–356.
- R. Balan, M. Singh, and D. Zou, *Lipschitz properties for deep convolutional networks*, arXiv preprint arXiv:1701.05217, (2017).
- R. Balan and Y. Wang, *Invertibility and robustness of phaseless reconstruction*, Applied and Computational Harmonic Analysis, 38 (2015), pp. 469–488.
- R. Balan and D. Zou, *On lipschitz inversion of nonlinear redundant representations*, Contemporary Mathematics, 650 (2015), pp. 15–22.

References II

- ———, *On lipschitz analysis and lipschitz synthesis for the phase retrieval problem*, Linear Algebra and its Applications, 496 (2016), pp. 152–181.
- A. S. Bandeira, J. Cahill, D. G. Mixon, and A. A. Nelson, *Saving phase: Injectivity and stability for phase retrieval*, Applied and Computational Harmonic Analysis, 37 (2014), pp. 106–125.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, in Advances in neural information processing systems, 2012, pp. 1097–1105.
- S. Mallat, *Group invariant scattering*, Communications on Pure and Applied Mathematics, 65 (2012), pp. 1331–1398.
- C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going deeper with convolutions*, in CVPR 2015, 2015.

References III

- C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, *Intriguing properties of neural networks*, CoRR, abs/1312.6199 (2013).
- M. Tygert, J. Bruna, S. Chintala, Y. LeCun, S. Piantino, and A. Szlam, *A mathematical motivation for complex-valued convolutional networks*, Neural computation, (2016).
- J. H. Wells and L. R. Williams, *Embeddings and extensions in analysis*, vol. 84, Springer Science & Business Media, 2012.
- T. Wiatowski and H. Bölcskei, *Deep convolutional neural networks based on semi-discrete frames*, in Proc. of IEEE International Symposium on Information Theory (ISIT), June 2015, pp. 1212–1216.
- ———, *A mathematical theory of deep convolutional neural networks for feature extraction*, IEEE Transactions on Information Theory, (2015).
- L. Zwald and G. Blanchard, *On the convergence of eigenspaces in kernel principal component analysis*, in NIPS, 2005, pp. 1649–1656.