Harmonic Analysis of Deep Convolutional Neural Networks

Helmut Bőlcskei

ETH zürich

Department of Information Technology and Electrical Engineering

October 2017

joint work with Thomas Wiatowski and Philipp Grohs

ImageNet











ImageNet

ski

rock

coffee

plant

ImageNet

ski

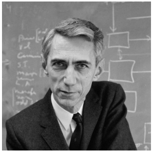
rock

coffee

plant

CNNs win the ImageNet 2015 challenge [*He et al., 2015*]

Feature extraction and classification

input: $f =$  

$\downarrow$

$\boxed{\text{non-linear feature extraction}}$

$\downarrow$

feature vector $\Phi(f)$

$\downarrow$

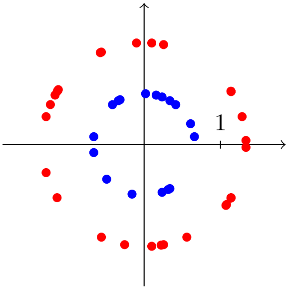
$\boxed{\text{linear classifier}}$

$\downarrow$

output: $\begin{cases} \langle w, \Phi(f) \rangle > 0, & \Rightarrow \text{Shannon} \\ \langle w, \Phi(f) \rangle < 0, & \Rightarrow \text{von Neumann} \end{cases}$

Why non-linear feature extractors?

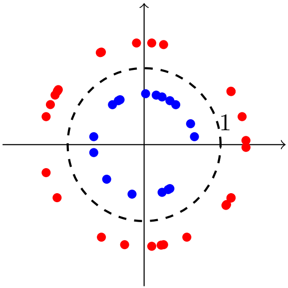
Task: Separate two categories of data through a **linear** classifier



- (red) : $\langle w, f \rangle > 0$
- (blue) : $\langle w, f \rangle < 0$

Why non-linear feature extractors?

Task: Separate two categories of data through a **linear** classifier



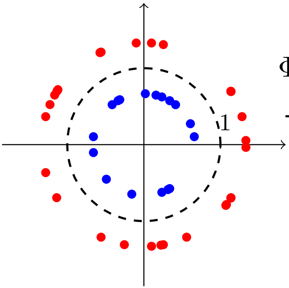
● : $\langle w, f \rangle > 0$

● : $\langle w, f \rangle < 0$

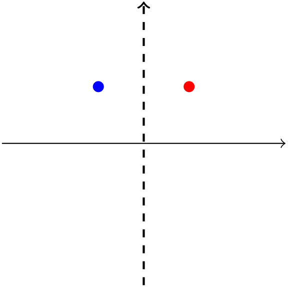
not possible!

Why non-linear feature extractors?

Task: Separate two categories of data through a **linear** classifier



$$\Phi(f) = \begin{bmatrix} \|f\| \\ 1 \end{bmatrix}$$



● : $\langle w, f \rangle > 0$

● : $\langle w, f \rangle < 0$

not possible!

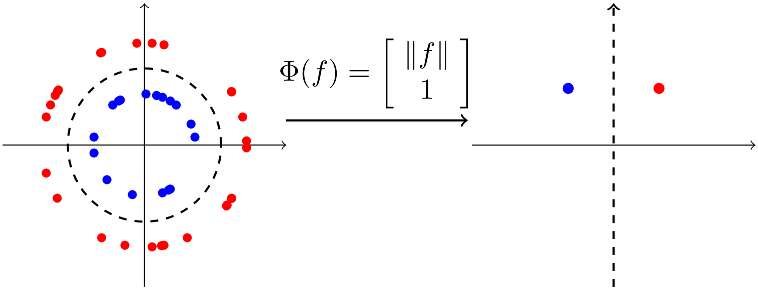
● : $\langle w, \Phi(f) \rangle > 0$

● : $\langle w, \Phi(f) \rangle < 0$

possible with $w = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

Why non-linear feature extractors?

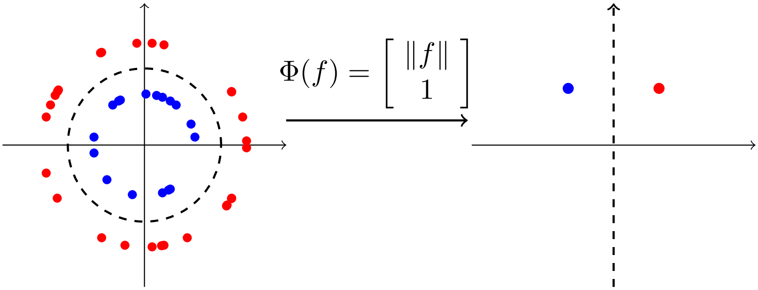
Task: Separate two categories of data through a **linear** classifier



$\Rightarrow \Phi$ is **invariant** to angular component of the data

Why non-linear feature extractors?

Task: Separate two categories of data through a **linear** classifier



$\Rightarrow \Phi$ is **invariant** to angular component of the data

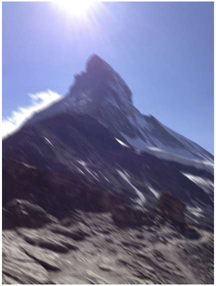
$\Rightarrow$ **Linear separability** in feature space!

Translation invariance

Handwritten digits from the MNIST database [*LeCun & Cortes, 1998*]

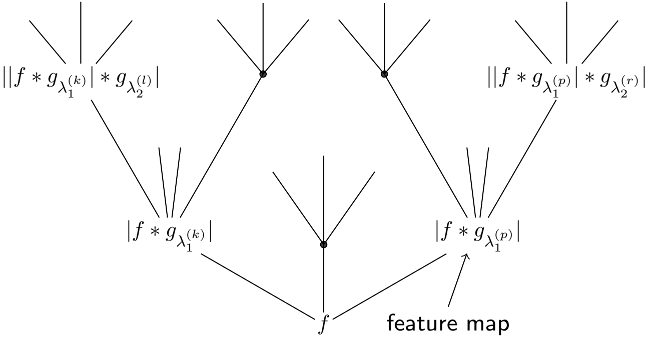
Feature vector should be invariant to spatial location
$\Rightarrow$ translation invariance

Deformation insensitivity

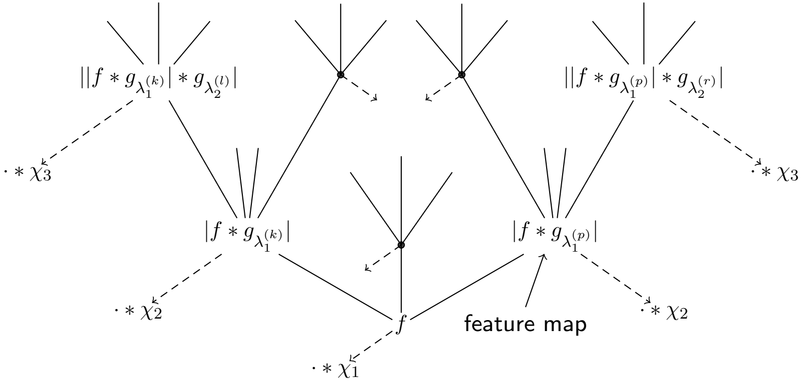


Feature vector should be independent of cameras (of different resolutions), and insensitive to small acquisition jitters

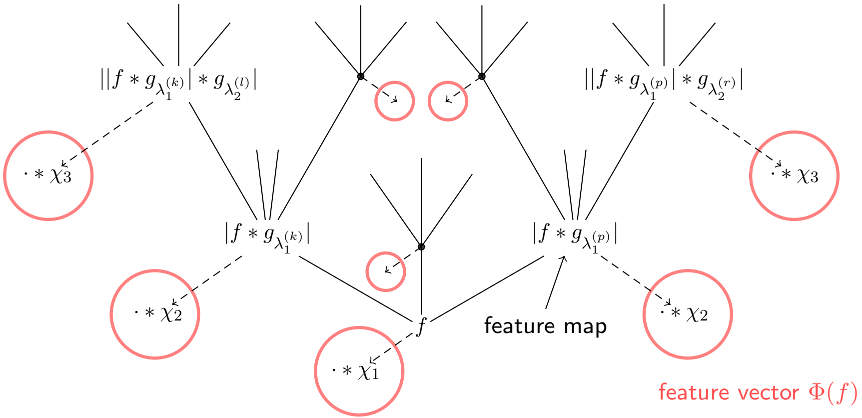
Scattering networks ([*Mallat, 2012*], [*Wiatowski and HB, 2015*])



Scattering networks ([*Mallat, 2012*], [*Wiatowski and HB, 2015*])



Scattering networks ([*Mallat, 2012*], [*Wiatowski and HB, 2015*])



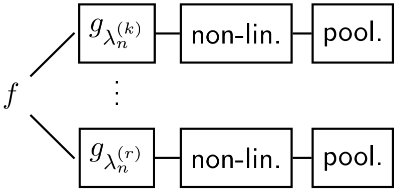
General scattering networks guarantee [*Wiatowski & HB, 2015*]

- (vertical) **translation invariance**
- **small deformation sensitivity**

essentially irrespective of filters, non-linearities, and poolings!

Building blocks

Basic operations in the n-th network layer

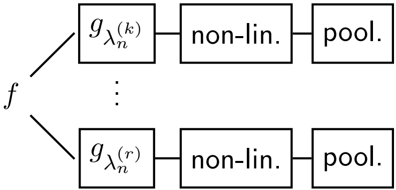


Filters: Semi-discrete frame $\Psi_n := \{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$

$$A_n \|f\|_2^2 \leq \|f * \chi_n\|_2^2 + \sum_{\lambda_n \in \Lambda_n} \|f * g_{\lambda_n}\|^2 \leq B_n \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$

Building blocks

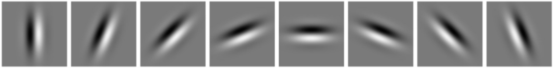
Basic operations in the n-th network layer



Filters: Semi-discrete frame $\Psi_n := \{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$

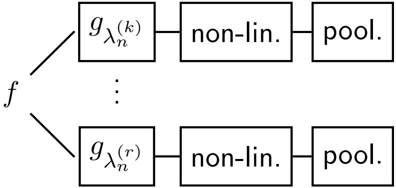
$$A_n \|f\|_2^2 \leq \|f * \chi_n\|_2^2 + \sum_{\lambda_n \in \Lambda_n} \|f * g_{\lambda_n}\|^2 \leq B_n \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$

e.g.: Structured filters



Building blocks

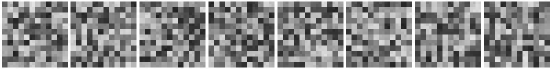
Basic operations in the n-th network layer



Filters: Semi-discrete frame $\Psi_n := \{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$

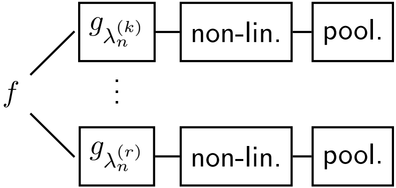
$$A_n \|f\|_2^2 \leq \|f * \chi_n\|_2^2 + \sum_{\lambda_n \in \Lambda_n} \|f * g_{\lambda_n}\|^2 \leq B_n \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$

e.g.: Unstructured filters



Building blocks

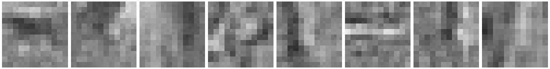
Basic operations in the n-th network layer



Filters: Semi-discrete frame $\Psi_n := \{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$

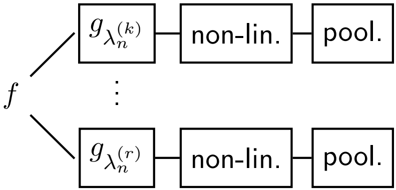
$$A_n \|f\|_2^2 \leq \|f * \chi_n\|_2^2 + \sum_{\lambda_n \in \Lambda_n} \|f * g_{\lambda_n}\|^2 \leq B_n \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$

e.g.: Learned filters



Building blocks

Basic operations in the n-th network layer

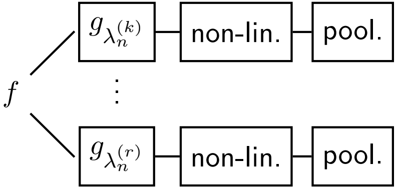


Non-linearities: Point-wise and Lipschitz-continuous

$$\|M_n(f) - M_n(h)\|_2 \leq L_n \|f - h\|_2, \quad \forall\, f, h \in L^2(\mathbb{R}^d)$$

Building blocks

Basic operations in the n-th network layer



Non-linearities: Point-wise and Lipschitz-continuous

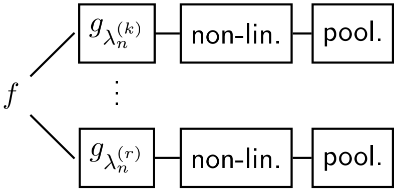
$$\|M_n(f) - M_n(h)\|_2 \leq L_n \|f - h\|_2, \quad \forall\, f, h \in L^2(\mathbb{R}^d)$$

$\Rightarrow$ Satisfied by virtually **all** non-linearities used
in the **deep learning literature**!

ReLU: $L_n = 1$; modulus: $L_n = 1$; logistic sigmoid: $L_n = \frac{1}{4}$; ...

Building blocks

Basic operations in the n-th network layer



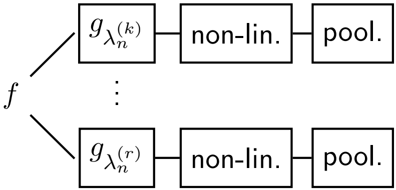
Pooling: In continuous-time according to

$$f \mapsto S_n^{d/2} P_n(f)(S_n \cdot),$$

where $S_n \geq 1$ is the **pooling factor** and $P_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ is R_n-Lipschitz-continuous

Building blocks

Basic operations in the n-th network layer



Pooling: In continuous-time according to

$$f \mapsto S_n^{d/2} P_n(f)(S_n \cdot),$$

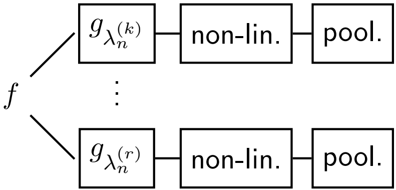
where $S_n \geq 1$ is the **pooling factor** and $P_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ is R_n-Lipschitz-continuous

$\Rightarrow$ **Emulates** most **poolings** used in the **deep learning literature**!

e.g.: Pooling by **sub-sampling** $P_n(f) = f$ with $R_n = 1$

Building blocks

Basic operations in the n-th network layer



Pooling: In continuous-time according to

$$f \mapsto S_n^{d/2} P_n(f)(S_n \cdot),$$

where $S_n \geq 1$ is the **pooling factor** and $P_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ is R_n-Lipschitz-continuous

$\Rightarrow$ **Emulates** most **poolings** used in the **deep learning literature**!

e.g.: Pooling by **averaging** $P_n(f) = f * \phi_n$ with $R_n = \|\phi_n\|_1$

Vertical translation invariance

Theorem (Wiatowski and HB, 2015)

Assume that the filters, non-linearities, and poolings satisfy

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N}.$$

Let the pooling factors be $S_n \geq 1$, $n \in \mathbb{N}$. *Then,*

$$|||\Phi^n(T_t f) - \Phi^n(f)||| = \mathcal{O}\left(\frac{\|t\|}{S_1 \dots S_n}\right),$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, $n \in \mathbb{N}$.

Vertical translation invariance

Theorem (Wiatowski and HB, 2015)

Assume that the filters, non-linearities, and poolings satisfy

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N}.$$

Let the pooling factors be $S_n \geq 1$, $n \in \mathbb{N}$. *Then,*

$$|||\Phi^n(T_t f) - \Phi^n(f)||| = \mathcal{O}\left(\frac{\|t\|}{S_1 \dots S_n}\right),$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, $n \in \mathbb{N}$.

$\Rightarrow$ Features become **more invariant** with **increasing** network **depth**!

Vertical translation invariance

Theorem (Wiatowski and HB, 2015)

Assume that the filters, non-linearities, and poolings satisfy

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N}.$$

Let the pooling factors be $S_n \geq 1$, $n \in \mathbb{N}$. Then,

$$|||\Phi^n(T_t f) - \Phi^n(f)||| = \mathcal{O}\left(\frac{\|t\|}{S_1 \dots S_n}\right),$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, $n \in \mathbb{N}$.

Full translation invariance: If $\lim_{n\to\infty} S_1 \cdot S_2 \cdot \ldots \cdot S_n = \infty$, then

$$\lim_{n\to\infty} |||\Phi^n(T_t f) - \Phi^n(f)||| = 0$$

Vertical translation invariance

Theorem (Wiatowski and HB, 2015)

Assume that the filters, non-linearities, and poolings satisfy

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N}.$$

Let the pooling factors be $S_n \geq 1$, $n \in \mathbb{N}$. Then,

$$|||\Phi^n(T_t f) - \Phi^n(f)||| = \mathcal{O}\left(\frac{\|t\|}{S_1 \ldots S_n}\right),$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, $n \in \mathbb{N}$.

The condition

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N},$$

is **easily satisfied** by **normalizing** the filters $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$.

Vertical translation invariance

Theorem (Wiatowski and HB, 2015)

Assume that the filters, non-linearities, and poolings satisfy

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N}.$$

Let the pooling factors be $S_n \geq 1$, $n \in \mathbb{N}$. *Then,*

$$|||\Phi^n(T_t f) - \Phi^n(f)||| = \mathcal{O}\left(\frac{\|t\|}{S_1 \ldots S_n}\right),$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, $n \in \mathbb{N}$.

$\Rightarrow$ applies to **general** filters, non-linearities, and poolings

Philosophy behind invariance results

Mallat's "horizontal" translation invariance [*Mallat, 2012*]:

$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d$$

"Vertical" translation invariance:

$$\lim_{n\to\infty} |||\Phi^n(T_t f) - \Phi^n(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d$$

Philosophy behind invariance results

Mallat's "horizontal" translation invariance [*Mallat, 2012*]:

$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d$$

- features become invariant in every network layer, but needs $J \to \infty$

"Vertical" translation invariance:

$$\lim_{n\to\infty} |||\Phi^n(T_t f) - \Phi^n(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d$$

- features become more invariant with increasing network depth

Philosophy behind invariance results

Mallat's "horizontal" translation invariance [*Mallat, 2012*]:

$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d$$

- features become invariant in every network layer, but needs $J \to \infty$
- applies to wavelet transform and modulus non-linearity without pooling

"Vertical" translation invariance:

$$\lim_{n\to\infty} |||\Phi^n(T_t f) - \Phi^n(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d$$

- features become more invariant with increasing network depth
- applies to general filters, general non-linearities, and general poolings

Non-linear deformations

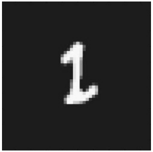
Non-linear deformation $(F_\tau f)(x) = f(x - \tau(x))$, where $\tau : \mathbb{R}^d \to \mathbb{R}^d$

For "small" τ:

Non-linear deformations

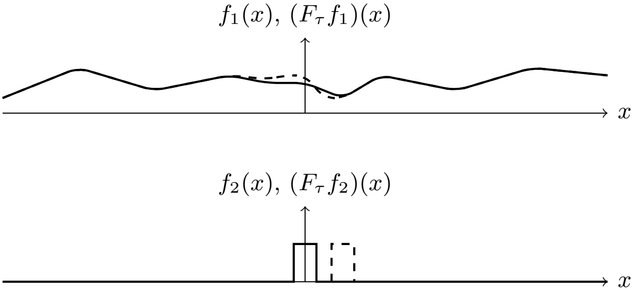
Non-linear deformation $(F_\tau f)(x) = f(x - \tau(x))$, where $\tau : \mathbb{R}^d \to \mathbb{R}^d$

For "large" τ:

Deformation sensitivity for signal classes

Consider $(F_\tau f)(x) = f(x - \tau(x)) = f(x - e^{-x^2})$



For given τ the amount of deformation induced
can depend drastically on $f \in L^2(\mathbb{R}^d)$

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound [*Mallat, 2012*]:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \le C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- The signal class H_W and the corresponding norm $\|\cdot\|_W$ depend on the mother wavelet (and hence the network)

Our deformation sensitivity bound:

$$|||\Phi(F_\tau f) - \Phi(f)||| \le C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- The signal class $\mathcal{C}$ (band-limited functions, cartoon functions, or Lipschitz functions) is independent of the network

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound [*Mallat, 2012*]:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- Signal class description complexity implicit via norm $\|\cdot\|_W$

Our deformation sensitivity bound:

$$|||\Phi(F_\tau f) - \Phi(f)||| \leq C_\mathcal{C}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- Signal class description complexity explicit via $C_\mathcal{C}$
  - L-band-limited functions: $C_\mathcal{C} = \mathcal{O}(L)$
  - cartoon functions of size K: $C_\mathcal{C} = \mathcal{O}(K^{3/2})$
  - M-Lipschitz functions $C_\mathcal{C} = \mathcal{O}(M)$

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound [*Mallat, 2012*]:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

Our deformation sensitivity bound:

$$|||\Phi(F_\tau f) - \Phi(f)||| \leq C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- Decay rate $\alpha > 0$ of the deformation error is signal-class-specific (band-limited functions: $\alpha = 1$, cartoon functions: $\alpha = \frac{1}{2}$, Lipschitz functions: $\alpha = 1$)

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound [*Mallat, 2012*]:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- The bound depends explicitly on higher order derivatives of τ

Our deformation sensitivity bound:

$$|||\Phi(F_\tau f) - \Phi(f)||| \leq C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- The bound implicitly depends on derivative of τ via the condition $\|D\tau\|_\infty \leq \frac{1}{2d}$

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound [*Mallat, 2012*]:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- The bound is *coupled* to horizontal translation invariance

$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d$$

Our deformation sensitivity bound:

$$|||\Phi(F_\tau f) - \Phi(f)||| \leq C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- The bound is *decoupled* from vertical translation invariance

$$\lim_{n\to\infty} |||\Phi^n(T_t f) - \Phi^n(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d$$

CNNs in a nutshell

CNNs used in practice employ potentially hundreds of layers and $10{,}000$s of nodes!

CNNs in a nutshell

CNNs used in practice employ potentially
hundreds of layers and $10{,}000$s of nodes!

e.g.: Winner of the ImageNet 2015 challenge [*He et al., 2015*]

- Network **depth**: 152 layers
- average # of **nodes** per layer: 472
- # of **FLOPS** for a single forward pass: 11.3 billion

CNNs in a nutshell

CNNs used in practice employ potentially
hundreds of layers and $10{,}000$s of nodes!

e.g.: Winner of the ImageNet 2015 challenge [*He et al., 2015*]

- Network **depth**: 152 layers
- average # of **nodes** per layer: 472
- # of **FLOPS** for a single forward pass: 11.3 billion

Such depths (and breadths) pose formidable computational
challenges in **training** and **operating** the network!

Topology reduction

Determine **how fast** the energy contained in the propagated signals (a.k.a. feature maps) decays across layers

Topology reduction

Determine **how fast** the energy contained in the propagated signals (a.k.a. feature maps) decays across layers

Guarantee **trivial null-space** for feature extractor Φ

Topology reduction

Determine **how fast** the energy contained in the propagated signals (a.k.a. feature maps) decays across layers

Guarantee **trivial null-space** for feature extractor Φ

Specify the **number of layers** needed to have "most" of the input signal energy be contained in the feature vector

Topology reduction

Determine **how fast** the energy contained in the propagated signals (a.k.a. feature maps) decays across layers

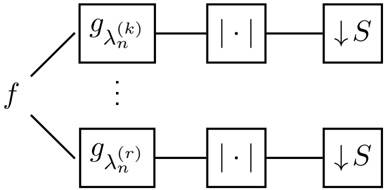
Guarantee **trivial null-space** for feature extractor Φ

Specify the **number of layers** needed to have “most” of the input signal energy be contained in the feature vector

For a fixed (possibly small) depth, **design CNNs** that capture “most” of the input signal energy

Building blocks

Basic operations in the n-th network layer



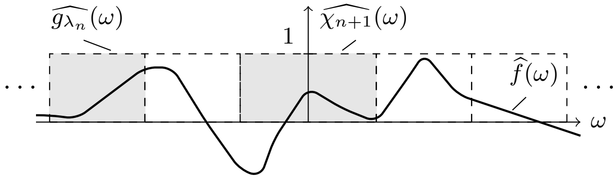
Filters: Semi-discrete frame $\Psi_n := \{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$

Non-linearity: Modulus $|\cdot|$

Pooling: Sub-sampling with pooling factor $S \geq 1$

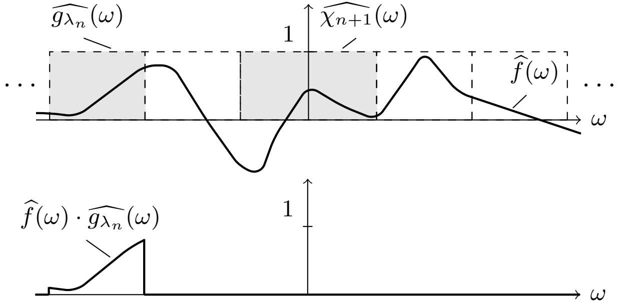
Demodulation effect of modulus non-linearity

Components of feature vector given by $|f * g_{\lambda_n}| * \chi_{n+1}$



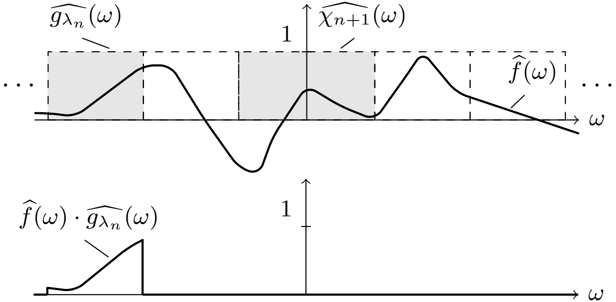
Demodulation effect of modulus non-linearity

Components of feature vector given by $|f * g_{\lambda_n}| * \chi_{n+1}$



Demodulation effect of modulus non-linearity

Components of feature vector given by $|f * g_{\lambda_n}| * \chi_{n+1}$

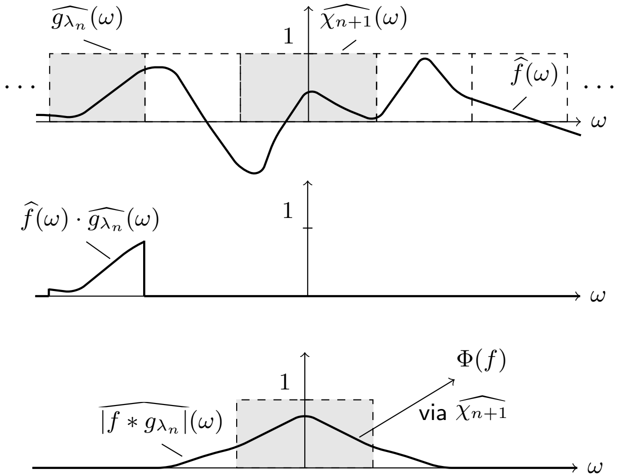


Modulus **squared**:

$$|f * g_{\lambda_n}(x)|^2 \quad \circ\!\!-\!\!\bullet \quad R_{\widehat{f} \cdot \widehat{g_{\lambda_n}}}(\omega)$$

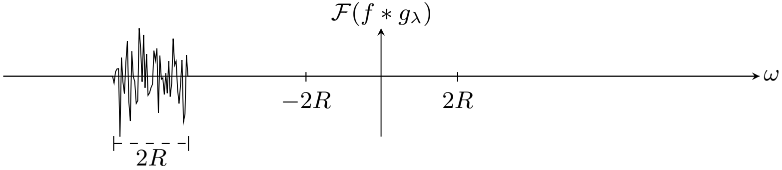
Demodulation effect of modulus non-linearity

Components of feature vector given by $|f * g_{\lambda_n}| * \chi_{n+1}$



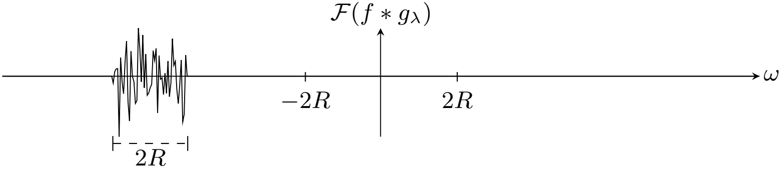
Do all non-linearities demodulate?

High-pass filtered signal:

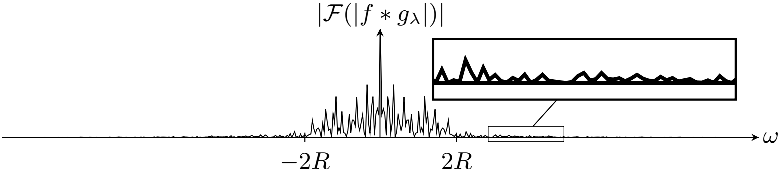


Do all non-linearities demodulate?

High-pass filtered signal:



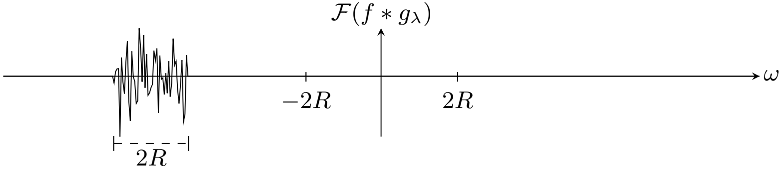
Modulus: **Yes**!



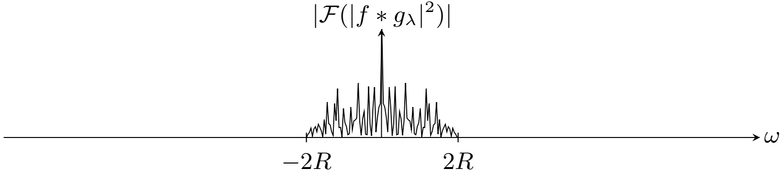
... but (small) tails!

Do all non-linearities demodulate?

High-pass filtered signal:



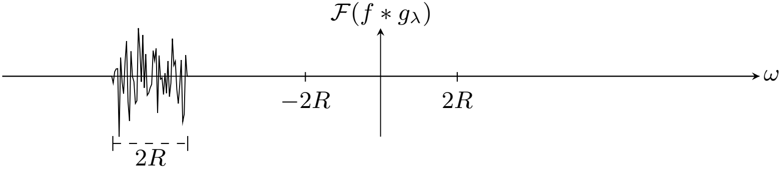
Modulus squared: **Yes, and sharply so**!



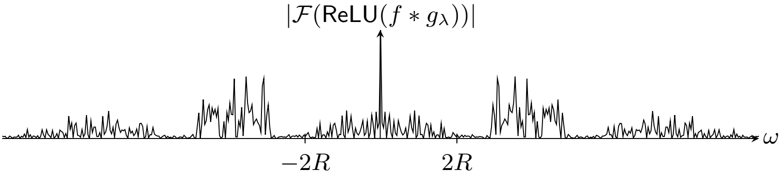
... but not Lipschitz-continuous!

Do all non-linearities demodulate?

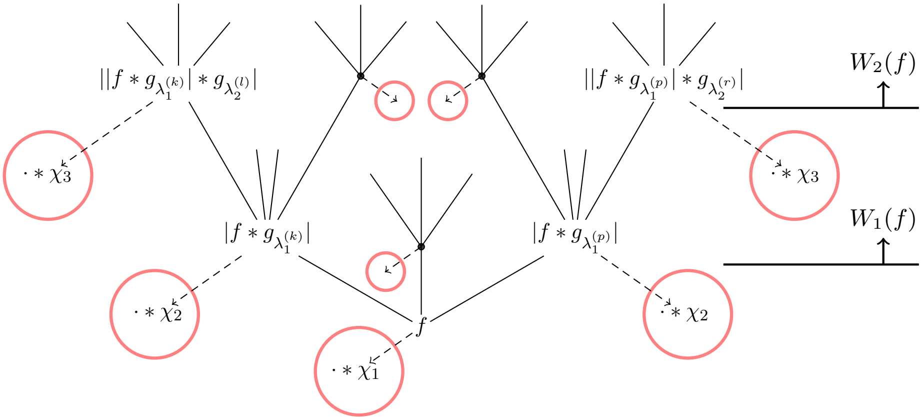
High-pass filtered signal:



Rectified linear unit: **No**!



First goal: Quantify feature map energy decay



Assumptions (on the filters)

i) **Analyticity**: For every filter g_{λ_n} there exists a (not necessarily canonical) orthant $H_{\lambda_n} \subseteq \mathbb{R}^d$ such that

$$\mathrm{supp}(\widehat{g_{\lambda_n}}) \subseteq H_{\lambda_n}.$$

ii) **High-pass**: There exists $\delta > 0$ such that

$$\sum_{\lambda_n \in \Lambda_n} |\widehat{g_{\lambda_n}}(\omega)|^2 = 0, \quad \text{a.e. } \omega \in B_\delta(0).$$

Assumptions (on the filters)

i) **Analyticity**: For every filter g_{λ_n} there exists a (not necessarily canonical) orthant $H_{\lambda_n} \subseteq \mathbb{R}^d$ such that

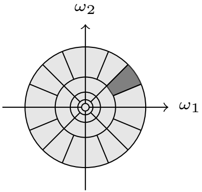
$$\mathrm{supp}(\widehat{g_{\lambda_n}}) \subseteq H_{\lambda_n}.$$

ii) **High-pass**: There exists $\delta > 0$ such that

$$\sum_{\lambda_n \in \Lambda_n} |\widehat{g_{\lambda_n}}(\omega)|^2 = 0, \quad \text{a.e. } \omega \in B_\delta(0).$$

$\Rightarrow$ Comprises various contructions of WH filters, wavelets, ridgelets, (α)-curvelets, shearlets

e.g.: analytic band-limited curvelets:



Input signal classes

Sobolev functions of order $s \geq 0$:

$$H^s(\mathbb{R}^d) = \left\{ f \in L^2(\mathbb{R}^d) \;\middle|\; \int_{\mathbb{R}^d} (1 + |\omega|^2)^s |\widehat{f}(\omega)|^2 \mathrm{d}\omega < \infty \right\}$$

Input signal classes

Sobolev functions of order $s \geq 0$:

$$H^s(\mathbb{R}^d) = \left\{ f \in L^2(\mathbb{R}^d) \;\middle|\; \int_{\mathbb{R}^d} (1 + |\omega|^2)^s |\widehat{f}(\omega)|^2 \mathrm{d}\omega < \infty \right\}$$

$H^s(\mathbb{R}^d)$ contains a wide range of **practically relevant** signal classes

Input signal classes

Sobolev functions of order $s \geq 0$:

$$H^s(\mathbb{R}^d) = \left\{ f \in L^2(\mathbb{R}^d) \;\middle|\; \int_{\mathbb{R}^d} (1 + |\omega|^2)^s |\widehat{f}(\omega)|^2 \mathrm{d}\omega < \infty \right\}$$

$H^s(\mathbb{R}^d)$ contains a wide range of **practically relevant** signal classes

- square-integrable functions $L^2(\mathbb{R}^d) = H^0(\mathbb{R}^d)$

Input signal classes

Sobolev functions of order $s \geq 0$:

$$H^s(\mathbb{R}^d) = \left\{ f \in L^2(\mathbb{R}^d) \;\middle|\; \int_{\mathbb{R}^d} (1 + |\omega|^2)^s |\widehat{f}(\omega)|^2 \mathrm{d}\omega < \infty \right\}$$

$H^s(\mathbb{R}^d)$ contains a wide range of **practically relevant** signal classes

- square-integrable functions $L^2(\mathbb{R}^d) = H^0(\mathbb{R}^d)$
- L-band-limited functions $L^2_L(\mathbb{R}^d) \subseteq H^s(\mathbb{R}^d)$, $\forall L > 0$, $\forall s \geq 0$

Input signal classes

Sobolev functions of order $s \geq 0$:

$$H^s(\mathbb{R}^d) = \left\{ f \in L^2(\mathbb{R}^d) \;\middle|\; \int_{\mathbb{R}^d} (1 + |\omega|^2)^s |\widehat{f}(\omega)|^2 \mathrm{d}\omega < \infty \right\}$$

$H^s(\mathbb{R}^d)$ contains a wide range of **practically relevant** signal classes

- square-integrable functions $L^2(\mathbb{R}^d) = H^0(\mathbb{R}^d)$
- L-band-limited functions $L^2_L(\mathbb{R}^d) \subseteq H^s(\mathbb{R}^d)$, $\forall L > 0$, $\forall s \geq 0$
- cartoon functions [*Donoho, 2001*] $\mathcal{C}_{\mathrm{CART}} \subseteq H^s(\mathbb{R}^d)$, $\forall s \in [0, \frac{1}{2})$

Handwritten digits from MNIST database [*LeCun & Cortes, 1998*]

Exponential energy decay

Theorem

Let the filters be **wavelets** *with mother wavelet*

$$\mathit{supp}(\widehat{\psi}\,) \subseteq [r^{-1}, r], \qquad r > 1,$$

or **Weyl-Heisenberg (WH) filters** *with prototype function*

$$\mathit{supp}(\widehat{g}) \subseteq [-R, R], \qquad R > 0.$$

Then, for every $f \in H^s(\mathbb{R}^d)$, there exists $\beta > 0$ such that

$$W_n(f) = \mathcal{O}\Big(a^{-\frac{n(2s+\beta)}{2s+\beta+1}}\Big),$$

where $a = \frac{r^2+1}{r^2-1}$ in the wavelet case, and $a = \frac{1}{2} + \frac{1}{R}$ in the WH case.

Exponential energy decay

Theorem

Let the filters be **wavelets** *with mother wavelet*

$$\mathit{supp}(\widehat{\psi}\,) \subseteq [r^{-1}, r], \qquad r > 1,$$

or **Weyl-Heisenberg (WH) filters** *with prototype function*

$$\mathit{supp}(\widehat{g}) \subseteq [-R, R], \qquad R > 0.$$

Then, for every $f \in H^s(\mathbb{R}^d)$, there exists $\beta > 0$ such that

$$W_n(f) = \mathcal{O}\Big(a^{-\frac{n(2s+\beta)}{2s+\beta+1}}\Big),$$

where $a = \frac{r^2+1}{r^2-1}$ in the wavelet case, and $a = \frac{1}{2} + \frac{1}{R}$ in the WH case.

$\Rightarrow$ decay factor a is **explicit** and can be **tuned** via r, R

Exponential energy decay

Exponential energy decay:

$$W_n(f) = \mathcal{O}\left(a^{-\frac{n(2s+\beta)}{2s+\beta+1}}\right)$$

Exponential energy decay

Exponential energy decay:

$$W_n(f) = \mathcal{O}\left(a^{-\frac{n(2s+\beta)}{2s+\beta+1}}\right)$$

- $\beta > 0$ determines the **decay** of $\widehat{f}(\omega)$ (as $|\omega| \to \infty$) according to

$$|\widehat{f}(\omega)| \leq \mu(1+|\omega|^2)^{-(\frac{s}{2}+\frac{1}{4}+\frac{\beta}{4})}, \quad \forall\, |\omega| \geq L,$$

for some $\mu > 0$, and L acts as an "effective bandwidth"

Exponential energy decay

Exponential energy decay:

$$W_n(f) = \mathcal{O}\Big(a^{-\frac{n(2s+\beta)}{2s+\beta+1}}\Big)$$

- $\beta > 0$ determines the **decay** of $\widehat{f}(\omega)$ (as $|\omega| \to \infty$) according to

$$|\widehat{f}(\omega)| \leq \mu(1+|\omega|^2)^{-(\frac{s}{2}+\frac{1}{4}+\frac{\beta}{4})}, \quad \forall\, |\omega| \geq L,$$

  for some $\mu > 0$, and L acts as an "effective bandwidth"
- **smoother** input signals (i.e., $s\uparrow$) lead to **faster** energy decay

Exponential energy decay

Exponential energy decay:

$$W_n(f) = \mathcal{O}\Big(a^{-\frac{n(2s+\beta)}{2s+\beta+1}}\Big)$$

- $\beta > 0$ determines the **decay** of $\widehat{f}(\omega)$ (as $|\omega| \to \infty$) according to

$$|\widehat{f}(\omega)| \leq \mu(1+|\omega|^2)^{-(\frac{s}{2}+\frac{1}{4}+\frac{\beta}{4})}, \quad \forall\, |\omega| \geq L,$$

  for some $\mu > 0$, and L acts as an "effective bandwidth"
- **smoother** input signals (i.e., $s\uparrow$) lead to **faster** energy decay
- **pooling** through sub-sampling $f \mapsto S^{1/2} f(S\cdot)$ leads to decay factor $\frac{a}{S}$

Exponential energy decay

Exponential energy decay:

$$W_n(f) = \mathcal{O}\Big(a^{-\frac{n(2s+\beta)}{2s+\beta+1}}\Big)$$

- $\beta > 0$ determines the **decay** of $\widehat{f}(\omega)$ (as $|\omega| \to \infty$) according to

$$|\widehat{f}(\omega)| \leq \mu(1+|\omega|^2)^{-(\frac{s}{2}+\frac{1}{4}+\frac{\beta}{4})}, \quad \forall\, |\omega| \geq L,$$

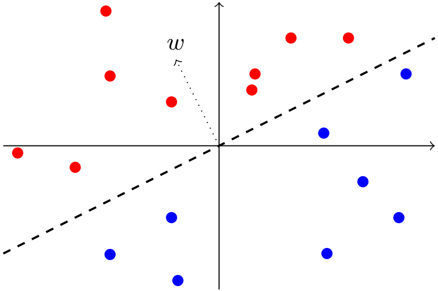
  for some $\mu > 0$, and L acts as an "effective bandwidth"
- **smoother** input signals (i.e., $s\uparrow$) lead to **faster** energy decay
- **pooling** through sub-sampling $f \mapsto S^{1/2} f(S\cdot)$ leads to decay factor $\frac{a}{S}$

What about **general** filters? $\Rightarrow$ **polynomial** energy decay!

... our second goal ... trivial null-space for Φ

Why trivial null-space?

Feature space



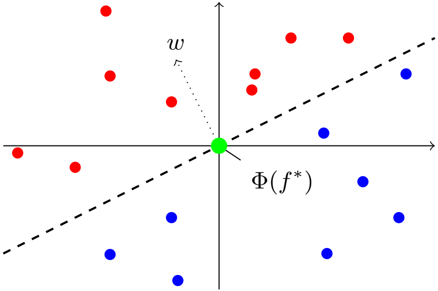
● : $\langle w, \Phi(f) \rangle > 0$

● : $\langle w, \Phi(f) \rangle < 0$

... our second goal ... trivial null-space for Φ

Why trivial null-space?

Feature space



$\bullet$: $\langle w, \Phi(f) \rangle > 0$

$\bullet$: $\langle w, \Phi(f) \rangle < 0$

Non-trivial null-space: $\exists\, f^* \neq 0$ such that $\Phi(f^*) = 0$

$\Rightarrow \langle w, \Phi(f^*) \rangle = 0$ **for all** w!

$\Rightarrow$ these f^* become **unclassifiable**!

... our second goal ...

Trivial null-space for feature extractor:

$$\{ f \in L^2(\mathbb{R}^d) \mid \Phi(f) = 0 \} = \{0\}$$

Feature extractor $\Phi(\cdot) = \bigcup_{n=0}^{\infty} \Phi^n(\cdot)$ shall satisfy

$$A\|f\|_2^2 \leq |||\Phi(f)|||^2 \leq B\|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d),$$

for some $A, B > 0$.

"Energy conservation"

Theorem

For the frame upper $\{B_n\}_{n\in\mathbb{N}}$ *and frame lower bounds* $\{A_n\}_{n\in\mathbb{N}}$*, define* $B := \prod_{n=1}^{\infty} \max\{1, B_n\}$ *and* $A := \prod_{n=1}^{\infty} \min\{1, A_n\}$*. If*

$$0 < A \leq B < \infty,$$

then

$$A\|f\|_2^2 \leq |||\Phi(f)|||^2 \leq B\|f\|_2^2, \qquad \forall\, f \in L^2(\mathbb{R}^d).$$

“Energy conservation”

Theorem

For the frame upper $\{B_n\}_{n\in\mathbb{N}}$ *and frame lower bounds* $\{A_n\}_{n\in\mathbb{N}}$*, define* $B := \prod_{n=1}^{\infty} \max\{1, B_n\}$ *and* $A := \prod_{n=1}^{\infty} \min\{1, A_n\}$*. If*

$$0 < A \leq B < \infty,$$

then

$$A\|f\|_2^2 \leq |||\Phi(f)|||^2 \leq B\|f\|_2^2, \qquad \forall\, f \in L^2(\mathbb{R}^d).$$

- For **Parseval** frames (i.e., $A_n = B_n = 1$, $n \in \mathbb{N}$), this yields

$$|||\Phi(f)|||^2 = \|f\|_2^2$$

“Energy conservation”

Theorem

For the frame upper $\{B_n\}_{n\in\mathbb{N}}$ *and frame lower bounds* $\{A_n\}_{n\in\mathbb{N}}$*, define* $B := \prod_{n=1}^{\infty} \max\{1, B_n\}$ *and* $A := \prod_{n=1}^{\infty} \min\{1, A_n\}$*. If*

$$0 < A \le B < \infty,$$

then

$$A\|f\|_2^2 \le |||\Phi(f)|||^2 \le B\|f\|_2^2, \qquad \forall\, f \in L^2(\mathbb{R}^d).$$

- For **Parseval** frames (i.e., $A_n = B_n = 1$, $n \in \mathbb{N}$), this yields

$$|||\Phi(f)|||^2 = \|f\|_2^2$$

- Connection to energy decay:

$$\|f\|_2^2 = \sum_{k=0}^{n-1} |||\Phi^k(f)|||^2 + \underbrace{W_n(f)}_{\to 0}$$

... and our third goal ...

For a given CNN, specify the **number of layers**
needed to capture "most" of the input signal energy

... and our third goal ...

For a given CNN, specify the **number of layers**
needed to capture "most" of the input signal energy

How many layers n are needed to have at least $((1-\varepsilon)\cdot 100)\%$ of the input signal energy be contained in the **feature vector**, i.e.,

$$(1-\varepsilon)\|f\|_2^2 \leq \sum_{k=0}^{n} |||\Phi^k(f)|||^2 \leq \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d).$$

Number of layers needed

Theorem

Let the frame bounds satisfy $A_n = B_n = 1$, $n \in \mathbb{N}$. Let the input signal f be L-band-limited, and let $\varepsilon \in (0,1)$. If

$$n \geq \left\lceil \log_a \left(\frac{L}{(1 - \sqrt{1-\varepsilon}\,)} \right) \right\rceil,$$

then

$$(1-\varepsilon)\|f\|_2^2 \leq \sum_{k=0}^{n} |||\Phi^k(f)|||^2 \leq \|f\|_2^2.$$

Number of layers needed

Theorem

Let the frame bounds satisfy $A_n = B_n = 1$, $n \in \mathbb{N}$. Let the input signal f be L-band-limited, and let $\varepsilon \in (0,1)$. If

$$n \geq \left\lceil \log_a\!\left(\frac{L}{(1-\sqrt{1-\varepsilon}\,)}\right)\right\rceil,$$

then

$$(1-\varepsilon)\|f\|_2^2 \leq \sum_{k=0}^{n} |||\Phi^k(f)|||^2 \leq \|f\|_2^2.$$

$\Rightarrow$ also guarantees **trivial null-space** for $\bigcup_{k=0}^{n} \Phi^k(f)$

Number of layers needed

Theorem

Let the frame bounds satisfy $A_n = B_n = 1$, $n \in \mathbb{N}$. Let the input signal f be L-band-limited, and let $\varepsilon \in (0,1)$. If

$$n \geq \left\lceil \log_a\left(\frac{L}{(1-\sqrt{1-\varepsilon}\,)}\right)\right\rceil,$$

then

$$(1-\varepsilon)\|f\|_2^2 \leq \sum_{k=0}^{n} |||\Phi^k(f)|||^2 \leq \|f\|_2^2.$$

- lower bound depends on
  - **description complexity** of input signals (i.e., bandwidth L)
  - **decay factor** (wavelets $a = \frac{r^2+1}{r^2-1}$, WH filters $a = \frac{1}{2} + \frac{1}{R}$)

Number of layers needed

Theorem

Let the frame bounds satisfy $A_n = B_n = 1$, $n \in \mathbb{N}$. Let the input signal f be L-band-limited, and let $\varepsilon \in (0,1)$. If

$$n \geq \left\lceil \log_a\left(\frac{L}{(1-\sqrt{1-\varepsilon}\,)}\right)\right\rceil,$$

then

$$(1-\varepsilon)\|f\|_2^2 \leq \sum_{k=0}^{n} |||\Phi^k(f)|||^2 \leq \|f\|_2^2.$$

- lower bound depends on
  - **description complexity** of input signals (i.e., bandwidth L)
  - **decay factor** (wavelets $a = \frac{r^2+1}{r^2-1}$, WH filters $a = \frac{1}{2} + \frac{1}{R}$)
- similar estimates for **Sobolev** input signals and for **general** filters (polynomial decay!)

Number of layers needed

Numerical example for bandwidth $L = 1$:

	$(1-\varepsilon)$					
	0.25	0.5	0.75	0.9	0.95	0.99
wavelets $(r = 2)$	2	3	4	6	8	11
WH filters $(R = 1)$	2	4	5	8	10	14
general filters	2	3	7	19	39	199

Number of layers needed

Numerical example for bandwidth $L = 1$:

	$(1-\varepsilon)$					
	0.25	0.5	0.75	0.9	0.95	0.99
wavelets $(r = 2)$	2	3	4	6	8	11
WH filters $(R = 1)$	2	4	5	8	10	14
general filters	2	3	7	19	39	199

Number of layers needed

Numerical example for bandwidth $L = 1$:

	$(1-\varepsilon)$					
	0.25	0.5	0.75	0.9	0.95	0.99
wavelets $(r = 2)$	2	3	4	6	8	11
WH filters $(R = 1)$	2	4	5	8	10	14
general filters	2	3	7	19	39	199

Recall: Winner of the ImageNet 2015 challenge [*He et al., 2015*]

- Network **depth**: 152 layers
- average # of **nodes** per layer: 472
- # of **FLOPS** for a single forward pass: 11.3 billion

... our fourth and last goal ...

For a fixed (possibly small) depth N, **design scattering networks** that capture "most" of the input signal energy

... our fourth and last goal ...

For a fixed (possibly small) depth N, **design scattering networks** that capture "most" of the input signal energy

Recall: Let the filters be **wavelets** with mother wavelet

$$\mathrm{supp}(\widehat{\psi}\,) \subseteq [r^{-1}, r], \qquad r > 1,$$

or **Weyl-Heisenberg filters** with prototype function

$$\mathrm{supp}(\widehat{g}) \subseteq [-R, R], \qquad R > 0.$$

... our fourth and last goal ...

For a fixed (possibly small) depth N, **design scattering networks** that capture "most" of the input signal energy

For fixed depth N, want to choose r in the wavelet and R in the WH case so that

$$(1-\varepsilon)\|f\|_2^2 \leq \sum_{k=0}^{N} |||\Phi^k(f)|||^2 \leq \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d).$$

Depth-constrained networks

Theorem

Let the frame bounds satisfy $A_n = B_n = 1$, $n \in \mathbb{N}$. Let the input signal f be L-band-limited, and fix $\varepsilon \in (0,1)$ and $N \in \mathbb{N}$. If, in the wavelet case,

$$1 < r \leq \sqrt{\frac{\kappa + 1}{\kappa - 1}},$$

or, in the WH case,

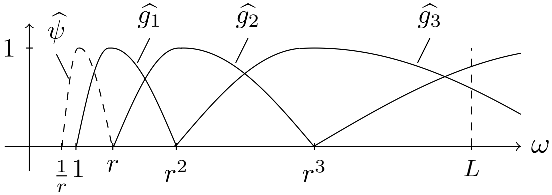
$$0 < R \leq \sqrt{\frac{1}{\kappa - \frac{1}{2}}},$$

where $\kappa := \left(\frac{L}{(1-\sqrt{1-\varepsilon}\,)}\right)^{\frac{1}{N}}$, then

$$(1-\varepsilon)\|f\|_2^2 \leq \sum_{k=0}^{N} |||\Phi^k(f)|||^2 \leq \|f\|_2^2.$$

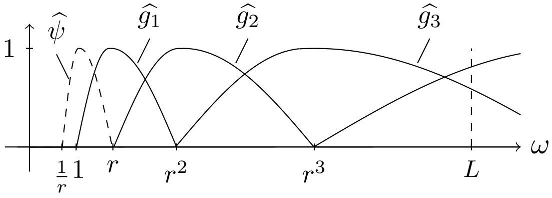
Depth-width tradeoff

Spectral supports of wavelet filters:



Depth-width tradeoff

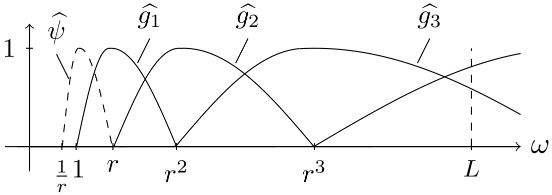
Spectral supports of wavelet filters:



Smaller depth $N \Rightarrow$ smaller "bandwidth" r of mother wavelet

Depth-width tradeoff

Spectral supports of wavelet filters:

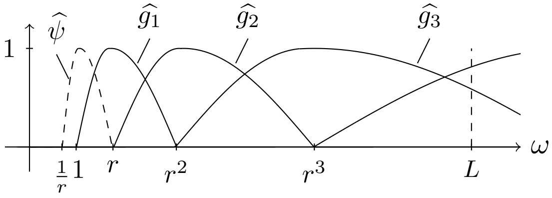


Smaller depth $N \Rightarrow$ smaller "bandwidth" r of mother wavelet

$\Rightarrow$ larger number of wavelets ($\mathcal{O}(\log_r(L))$) to cover the spectral support $[-L, L]$ of input signal

Depth-width tradeoff

Spectral supports of wavelet filters:



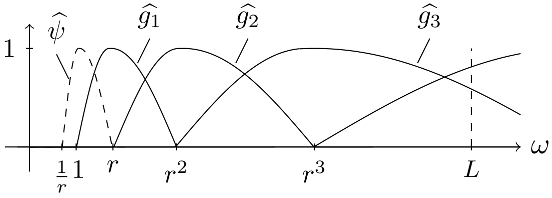
Smaller depth $N \Rightarrow$ smaller "bandwidth" r of mother wavelet

$\Rightarrow$ larger number of wavelets ($\mathcal{O}(\log_r(L))$) to cover the spectral support $[-L, L]$ of input signal

$\Rightarrow$ larger number of filters in the first layer

Depth-width tradeoff

Spectral supports of wavelet filters:



Smaller depth N $\Rightarrow$ smaller "bandwidth" r of mother wavelet

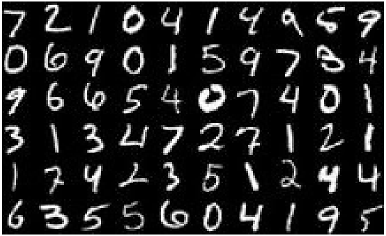
$\Rightarrow$ larger number of wavelets ($\mathcal{O}(\log_r(L))$) to cover the spectral support $[-L, L]$ of input signal

$\Rightarrow$ larger number of filters in the first layer

$\Rightarrow$ **depth-width tradeoff**

Yours truly

Experiment: Handwritten digit classification



- Dataset: MNIST database of handwritten digits [*LeCun & Cortes, 1998*]; 60,000 training and 10,000 test images
- Φ-network: $D = 3$ layers; same filters, non-linearities, and pooling operators in all layers
- Classifier: SVM with radial basis function kernel [*Vapnik, 1995*]
- Dimensionality reduction: Supervised orthogonal least squares scheme [*Chen et al., 1991*]

Experiment: Handwritten digit classification

Classification error in percent:

	Haar wavelet				Bi-orthogonal wavelet			
	abs	ReLU	tanh	LogSig	abs	ReLU	tanh	LogSig
n.p.	0.57	0.57	1.35	1.49	0.51	0.57	1.12	1.22
sub.	0.69	0.66	1.25	1.46	0.61	0.61	1.20	1.18
max.	0.58	0.65	0.75	0.74	0.52	0.64	0.78	0.73
avg.	0.55	0.60	1.27	1.35	0.58	0.59	1.07	1.26

Experiment: Handwritten digit classification

Classification error in percent:

	Haar wavelet				Bi-orthogonal wavelet			
	abs	ReLU	tanh	LogSig	abs	ReLU	tanh	LogSig
n.p.	0.57	0.57	1.35	1.49	0.51	0.57	1.12	1.22
sub.	0.69	0.66	1.25	1.46	0.61	0.61	1.20	1.18
max.	0.58	0.65	0.75	0.74	0.52	0.64	0.78	0.73
avg.	0.55	0.60	1.27	1.35	0.58	0.59	1.07	1.26

- modulus and ReLU perform better than tanh and LogSig

Experiment: Handwritten digit classification

Classification error in percent:

	Haar wavelet				Bi-orthogonal wavelet			
	abs	ReLU	tanh	LogSig	abs	ReLU	tanh	LogSig
n.p.	0.57	0.57	1.35	1.49	0.51	0.57	1.12	1.22
sub.	0.69	0.66	1.25	1.46	0.61	0.61	1.20	1.18
max.	0.58	0.65	0.75	0.74	0.52	0.64	0.78	0.73
avg.	0.55	0.60	1.27	1.35	0.58	0.59	1.07	1.26

- modulus and ReLU perform better than tanh and LogSig
- results with pooling ($S = 2$) are competitive with those without pooling, at significanly lower computational cost

Experiment: Handwritten digit classification

Classification error in percent:

	Haar wavelet				Bi-orthogonal wavelet			
	abs	ReLU	tanh	LogSig	abs	ReLU	tanh	LogSig
n.p.	0.57	0.57	1.35	1.49	0.51	0.57	1.12	1.22
sub.	0.69	0.66	1.25	1.46	0.61	0.61	1.20	1.18
max.	0.58	0.65	0.75	0.74	0.52	0.64	0.78	0.73
avg.	0.55	0.60	1.27	1.35	0.58	0.59	1.07	1.26

- modulus and ReLU perform better than tanh and LogSig
- results with pooling ($S = 2$) are competitive with those without pooling, at significanly lower computational cost
- state-of-the-art: 0.43 [*Bruna and Mallat, 2013*]
  - similar feature extraction network with directional, non-separable wavelets and no pooling
  - significantly higher computational complexity

Energy decay: Related work

[*Waldspurger, 2017*]: Exponential energy decay

$$W_n(f) = \mathcal{O}(a^{-n}),$$

for some **unspecified** $a > 1$.

- 1-D **wavelet** filters
- **every** network layer equipped with the **same** set of wavelets

Energy decay: Related work

[*Waldspurger, 2017*]: Exponential energy decay

$$W_n(f) = \mathcal{O}(a^{-n}),$$

for some **unspecified** $a > 1$.

- 1-D **wavelet** filters
- **every** network layer equipped with the **same** set of wavelets
- **vanishing moments** condition on the mother wavelet

Energy decay: Related work

[*Waldspurger, 2017*]: Exponential energy decay

$$W_n(f) = \mathcal{O}(a^{-n}),$$

for some **unspecified** $a > 1$.

- 1-D **wavelet** filters
- **every** network layer equipped with the **same** set of wavelets
- **vanishing moments** condition on the mother wavelet
- applies to 1-D **real-valued band-limited** input signals $f \in L^2(\mathbb{R})$

Energy decay: Related work

[*Czaja and Li, 2016*]: Exponential energy decay

$$W_n(f) = \mathcal{O}(a^{-n}),$$

for some **unspecified** $a > 1$.

- d-dimensional **uniform covering** filters (similar to Weyl-Heisenberg filters), but does not cover **multi-scale** filters (e.g. wavelets, ridgedelets, curvelets etc.)
- **every** network layer equipped with the **same** set of filters

Energy decay: Related work

[*Czaja and Li, 2016*]: Exponential energy decay

$$W_n(f) = \mathcal{O}(a^{-n}),$$

for some **unspecified** $a > 1$.

- d-dimensional **uniform covering** filters (similar to Weyl-Heisenberg filters), but does not cover **multi-scale** filters (e.g. wavelets, ridgedelets, curvelets etc.)
- **every** network layer equipped with the **same** set of filters
- **analyticity** and **vanishing moments** conditions on the filters

Energy decay: Related work

[*Czaja and Li, 2016*]: Exponential energy decay

$$W_n(f) = \mathcal{O}(a^{-n}),$$

for some **unspecified** $a > 1$.

- d-dimensional **uniform covering** filters (similar to Weyl-Heisenberg filters), but does not cover **multi-scale** filters (e.g. wavelets, ridgedelets, curvelets etc.)
- **every** network layer equipped with the **same** set of filters
- **analyticity** and **vanishing moments** conditions on the filters
- applies to d-dimensional **complex-valued** input signals $f \in L^2(\mathbb{R}^d)$