Mathematical Methods for Machine Learning and Signal Processing SS 2019

Lecture 21: Scattering Transform

V. Morgenshtern

Agenda:

1. Signal representations for pattern recognition
2. Invariance and stability
3. Why not Fourier modulus?
4. The scattering representation
5. The scattering representation for images
6. Energy propagation
7. Inverting the scattering transform
8. Each layer is a frame expansion
9. Energy conservation
10. Local translation invariance
11. Lipschitz continuity to deformations

This lecture is an overview of the scattering transform. For details please see [1, 2] and [3, 4] for a more complete treatment.

1 Signal representations for pattern recognition

A fundamental topic in image and audio processing is to find appropriate metrics to compare images and sounds. One way to construct a metric is to use the Euclidean distance on a (nonlinear) signal representation Φ, applied to signals x, x':

$$d(x, x') = \|\Phi(x) - \Phi(x')\|.$$

This puts all the structure of the problem in the construction of a signal representation, whose role is to encode the relevant signal information and to capture the right notion of similarity for a given task.

Object and sounds are perceived and recognized in the human brain in a fraction of a second, under a variety of physical transformations, including translations, rotations, illumination changes or other deformations. The question is how to construct the signal representation Φ that would be invariant to all these transformations, and at the same time preserve enough data about the underlying signals such that important task, such as classification or image recognition can be solved.

Scattering operators build invariant, stable, and informative signal representations by cascading wavelet modulus decompositions followed by a lowpass averaging filter. Scattering representations have the structure of a convolutional network. Rather than being learnt, the scattering network is obtained from the invariance, stability, and informative requirements, which lead to wavelet filter banks and to point-wise non-linearities.

2 Invariance and stability

Invariance to group action: Assume $x \in \mathcal{L}^2(\mathbb{R}^d)$. Let G denote a group of transformations (translations, rotations, etc.) of $\mathbb{R}^d$ and $L_\phi x(u) = x(\phi(u))$ denote the action of an element $\phi \in G$. The invariance to the action of G is obtained by requiring

$$\Phi(L_\phi x) = \Phi(x) \text{ for all } \phi \in G \text{ and } x \in \mathcal{L}^2(\mathbb{R}^d).$$

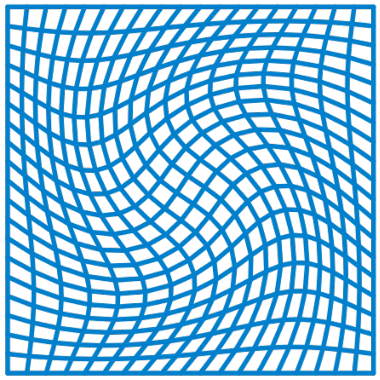
We will say that $\Phi(x)$ are the features of the signal x.

In this lecture, we will primarily be interested in building a representation that is invariant to translation. Let $x_c(u) = x(u - c)$ denote the translation of signal $x(\cdot)$ by $c \in \mathbb{R}^d$. Then, we want:

$$\Phi(x_c) = \Phi(x) \text{ for all } c \in \mathbb{R}^d \text{ and } x \in \mathcal{L}^2(\mathbb{R}^d).$$

The theory can be generalized to other groups as well. In practice, translation invariance means the features corresponding to each one of the following images are identical:

Stability to deformations: A diffeomorphism, denoted $\phi \in \mathrm{Diff}(\mathbb{R}^d)$, is a map $\phi : \mathbb{R}^d \to \mathbb{R}^d$ that is differentiable and has a differentiable inverse. For example, here is the image of a rectangular grid on a square under a diffeomorphism from the square onto itself:

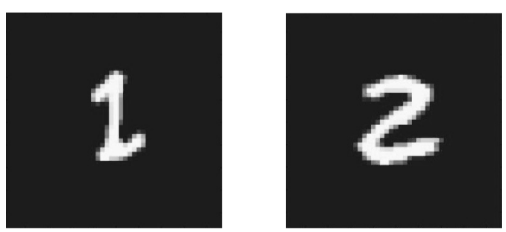


Stability to deformations is expressed as a Lipschitz continuity condition with respect to a metric $\|\phi\|$ on the space of diffeomorphisms measuring the amount of deformation:

$$\|\Phi(L_\phi x) - \Phi(x)\| \leq C\|x\|\|\phi\|, \ \phi \in \text{Diff}(\mathbb{R}^d) \text{ and } x \in \mathcal{L}^2(\mathbb{R}^d).$$

In practice, stability to deformations means that the feature corresponding to each of these images are close together:

Note that if the diffeomorphism is large, it may convert an image of digit 1 to the an image of digit 2:



This example shows that in general we want stability with respect to small diffeomorphisms only.

We will be interested in small diffeomorphisms that are not too far from the identity transformation. It is convenient to represent such diffeomorphism as $u \to u - \tau(u)$, where $\tau(u)$ is a vector displacement field satisfying $\|\nabla \tau\|_\infty < 1$. The diffeomorphism acts on $x \in \mathcal{L}^2(\mathbb{R}^d)$ by composition:

$$L[\tau]x(u) = x(u - \tau(u))$$

3 Why not Fourier modulus?

Translation invariant representation can be obtained from Fourier modulus operator. However, the resulting representations are not Lipschitz continuous to deformations, as we will see next.

Let $\hat{x}(\omega)$ be the Fourier transform of $x(u)$. Since $\hat{x_c}(\omega) = e^{-ic\dot{\omega}}\hat{x}(w)$, it results that $|\hat{x_c}| = |\hat{x}|$ does not depend upon c. We see that the feature representation $\Phi(x) = |\hat{x}(w)|$ is invariant to translations, as desired.

However, Fourier transform modulus is unstable to small deformations at high frequencies. This can be seen as follows. Let $x(u) = e^{i\epsilon u}\theta(u)$ be a modulated version of a lowpass window $\theta(u)$ with the central frequency at ϵ. Let $\tau(u) = su$ denote a linear displacement (difformation) field where $|s|$ is small. Then the deformed version of x is

$$x_\tau(u) = L_\tau x(u) = x((1+s)u)$$

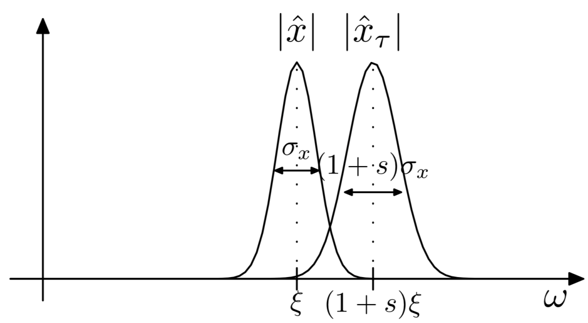
has the central frequency at $(1+s)\epsilon$. If $\sigma_\theta^2 = \int |\omega|^2 |\theta(\omega)|^2 \, d\omega$ measures the frequency spread of θ, then

$$\sigma_x^2 = \int |\omega - \epsilon|^2 |\hat{x}(\omega)|^2 \, d\omega = \sigma_\theta^2,$$

and

$$\begin{aligned} \sigma_{x_\tau}^2 &= (1+s)^{-d} \int (\omega - (1+s)\epsilon)^2 \left|\hat{x}((1+s)^{-1}\omega)\right|^2 d\omega \\ &= \int |(1+s)(\omega - \epsilon)|^2 |\hat{x}(\omega)|^2 \, d\omega = (1+s)^2 \sigma_x^2. \end{aligned}$$

It follows that if the distance between the central frequency of x and x_τ, $s\epsilon$, is large compared to their frequency spreads, $(2+s)\sigma_s$, then the frequency supports of x and x_τ are nearly disjoint:

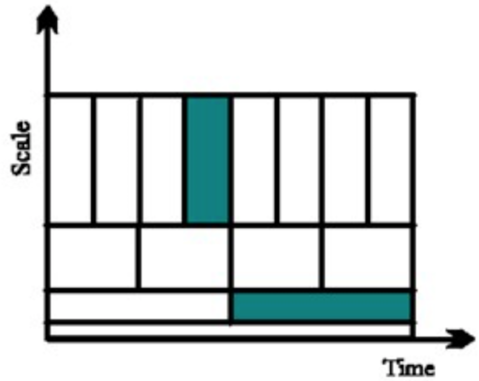


and hence

$$\| |\hat{x_\tau}| - |\hat{x}| \| \sim \|x\|,$$

which shows that $\Phi(x) = |\hat{x}|$ is not Lipschitz continuous to deformations since ϵ can be arbitrarily large.

In contrast, remember that the wavelet transform coefficients can be understood in the frequency domain as averaging operations with the windows that look as follows in the time-frequency plane:



Therefore, at high frequencies, the wavelet transform provides averaging with wide windows:

$$\int_{\mathbb{R}^d} (x \star \psi_\lambda)(u) du = \int_{\mathbb{R}^d} x(t)\psi_\lambda(u-t)dt,$$

which in the Fourier domain is equal to:

$$\hat{x}(\omega)\hat{\psi}_\lambda(\omega).$$

Scattering transform is a multi-layer network that is based on convolutions with wavelets. Because wavelets become wider in the frequency domain at high frequencies, the representation is stable to deformations.

Besides deformation instabilities, the Fourier modulus looses too much information. For example, a Dirac delta $\delta(u)$ and a linear chirp e^{iu^2} are two signals having Fourier transforms whose moduli are equal and constant.

4 The scattering representation

Scattering representations construct invariant, stable, and informative signal representations by cascading wavelet modulus decompositions followed by a lowpass filter. A wavelet decomposition operator at scale J is defined as

$$W_\lambda x(u) = x \star \psi_\lambda(u) = \int x(v)\psi_\lambda(u-v)dv.$$

where and $\psi_\lambda(u) = 2^{-dj}\psi(2^{-j}r^{-1}u)$ and $\lambda = 2^j r$, with $j < J$ and $r \in G$ belongs to a finite rotation group G of $\mathbb{R}^d$. We say, $\lambda \in \Lambda_J$ where $\Lambda_J = \{2^j r : r \in G/\{\pm 1\}, |\lambda| \leq 2^J\}$. Each rotated and dilated wavelet thus extracts the energy of x located at a given scale and orientation given by λ. As a side remark, note that the larger j is, the wider $\psi_\lambda(u)$ is.

The difficulty is that the wavelet coefficients are not translation invariant. We could obtain try to build a translation invariant representation by averaging the coefficients over $\mathbb{R}^d$. However such averaging does not produce any information because wavelets have zero-mean:

$$\int_{\mathbb{R}^d} (x \star \psi_\lambda)(u) du = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} x(t)\psi_\lambda(u-t)dtdu = \int_{\mathbb{R}^d} x(t) \underbrace{\left(\int_{\mathbb{R}^d} \psi_\lambda(u-t)du\right)}_{0} dt = 0.$$

A translation invariant measure can be extracted out of each wavelet sub-band λ by introducing a non-linearity which restores a non-zero, informative average value. This is for instance achieved by computing the complex modulus and averaging the result

$$\int |x \star \psi_\lambda| \, du.$$

The information lost by this averaging is recovered by a new wavelet decomposition

$$\{|x \star \psi_\lambda| \star \psi'_\lambda\}_{\lambda' \in \Lambda_J}$$

of $|x \star \psi_\lambda|$, which produces new invariants by iterating the same procedure.

Let $U[\lambda]x = |x \star \psi_\lambda|$ denote the wavelet modulus operator corresponding to the subband λ. Any sequence $p = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ defines a path, i.e., the ordered product of non-linear and non-commuting operators

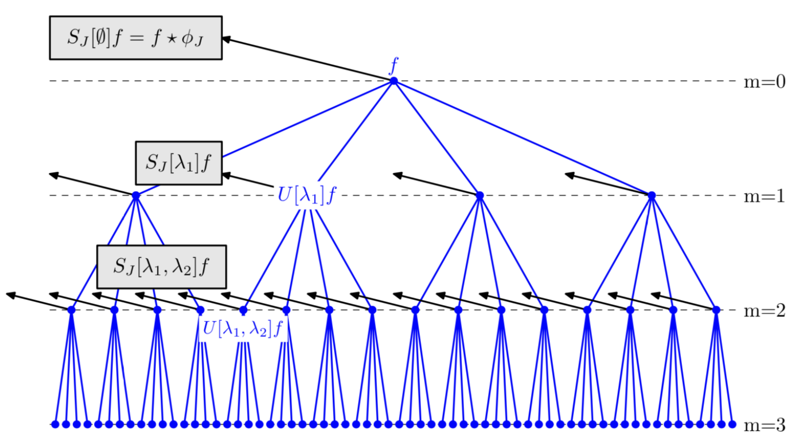
$$U[p]x = U[\lambda_m] \ldots U[\lambda_2] U[\lambda_1] x = |||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \ldots \star \psi_{\lambda_m}|$$

with $U[0]x = x$.

Many applications in image and audio recognition require locally translation invariant representations, which also keep spatial or temporal information intact beyond a certain scale 2^J. A windowed scattering transform computes a locally translation invariant representation by computing a lowpass average at scale 2^J with a lowpass filter $\phi_{2^J}(u) = 2^{-2J}\phi(2^{-J}u)$. For each path $p = (\lambda_1, \ldots, \lambda_m)$ with $\lambda_i \in \Lambda_J$ we defined the windowed scattering transform as

$$S_J[p]x(u) = U[p]x \star \phi_{2^J}(u) = \int U[p]x(v)\phi_{2^J}(u - v)dv.$$

A Scattering transform has the structure of a convolutional network, but its filters are given by wavelets instead of being learned:



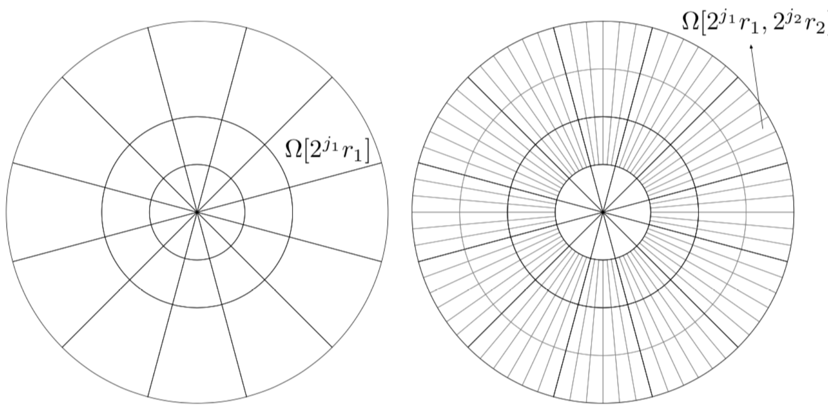
5 The scattering representation for images

To make the presentation more concrete, let's consider the special case of scattering transform for images. We start with a Morlet wavelet

$$\psi(u) = \alpha(e^{iu.\epsilon} - \beta)e^{-|u|^2/(2\sigma^2)}$$

where $\beta \ll 1$ is adjusted so that $\int \psi(u)du = 0$. Here is the real part of ψ, the imaginary part of ψ, and the Fourier modulus of $\left|\hat{\psi}(\omega)\right|$ for a $\epsilon = (0, 3\pi/4)$, which encodes the angle:

For a fixed position u, windowed scattering coefficients $S_J[p]x(u)$ of order $m = 1, 2$ are displayed as piecewise constant images over a disk representing the Fourier support of the image x. This frequency disk is partitioned into sectors $\{\Omega[p]\}$ indexed by the path p. The image value is $S_J[p]x(u)$ on the frequency sectors $\Omega[p]$ shown in the image below:



For $m = 1$, a scattering coefficient $S_J[\lambda_1]x(u)$ depends upon local Fourier transform energy of x over the support of $\hat{\psi}_{\lambda_1}$. Its values is displayed over a sector $\Omega[\lambda_1]$ which approximates the frequency support of $\hat{\psi}_{\lambda_1}$. For $\lambda_1 = 2^{-j_1}r_1$, there are K rotated sectors located in an annulus of scale 2^{-j_1}, corresponding to each $r_1 \in G$, as shown in the image on the left. The area are proportional to $\|\psi_{\lambda_1}\|^2 \sim K^{-1}2^{-j_1}$.

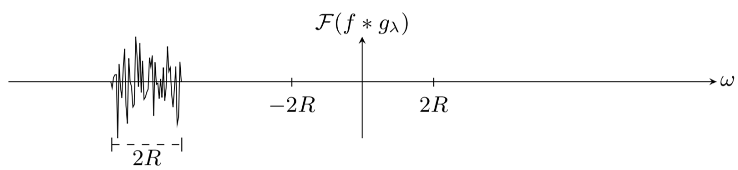
Second order scattering coefficients $S_J[\lambda_1, \lambda_2]x(u)$ are computed with a second wavelet transform which performs a second frequency subdivision. These coefficients are displayed over frequency sectors $\Omega[\lambda_1, \lambda_2]$ which subdivide the sectors $\Omega[\lambda_1]$ of the first wavelets $\hat{\psi}_{\lambda_1}$, as illustrated in the image on the right. For $\lambda_2 = 2^{-j_2}r_2$, the scale 2^{j_2} divides the radial axis and the resulting sectors are subdivided into K angular sectors corresponding to the different r_2. The scale and the angular subdivisions are adjusted so that the area of each $\Omega[\lambda_1, \lambda_2]$ is proportional to $\||\psi_{\lambda_1}| \star \psi_{\lambda_2}\|^2$.

6 Energy propagation

The table below shows the percentage of energy $\sum p$ of length $m\|S_J[p]x\|^2/\|x\|^2$ of scattering coefficients on frequency-decreasing paths of length m, depending upon J. The average values are computed on the Caltech-101 database, with zero mean and unit variance images:

J	$m=0$	$m=1$	$m=2$	$m=3$	$m=4$	$m \leq 3$
1	95.1	4.86	-	-	-	99.96
2	87.56	11.97	0.35	-	-	99.89
3	76.29	21.92	1.54	0.02	-	99.78
4	61.52	33.87	4.05	0.16	0	99.61
5	44.6	45.26	8.9	0.61	0.01	99.37
6	26.15	57.02	14.4	1.54	0.07	99.1
7	0	73.37	21.98	3.56	0.25	98.91

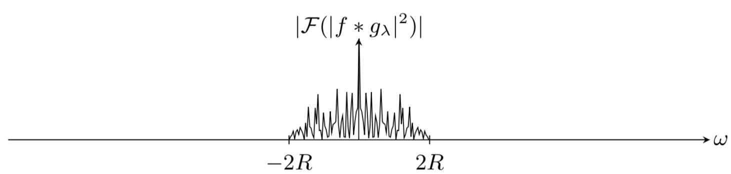
The propagated energy $\|U[p]x\|^2$ decays because $U[p]x$ is a progressively lower frequency signal as the path length increases. Indeed, each modulus computes a regular envelop of oscillating wavelet coefficients. The modulus can thus be interpreted as s non-linear "demodulator" which pushes the wavelet coefficient energy towards lower frequencies. To understand consider a hi-pass signal $f \star \psi_\lambda$:



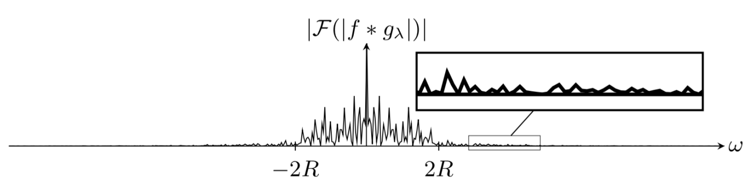
Then the Fourier transform of its modulus squared

$$\mathcal{F}(|f \star \psi_\lambda|^2) = \mathcal{F}(f \star \psi_\lambda) \star \overline{\mathcal{F}(f \star \psi_\lambda)} = R_{\hat{f}\hat{\psi}_\lambda}(\omega)$$

which is supported on the interval $[-2R, 2R]$:



Fourier transform of the modulus (without the square), is concentrated on $[-2R, 2R]$, but it's support is wider:

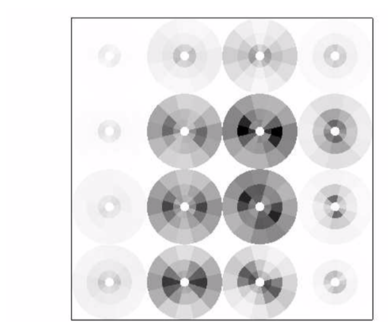


As a result, an important portion of the energy of $U[p]x$ is then captured by the low pass filter ϕ_{2^J} which outputs $S_J[p]x = U[p]x \star \phi_{2^J}$. Hence fewer energy is propagated to the next layer.

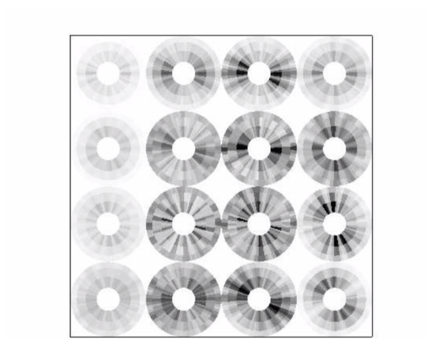
Another consequence is that the scattering energy propagates only along a subset of frequency decreasing paths. Since the envelope $|x \star \phi_\lambda|$ is more regular than $x \star \phi_\lambda$ it results that $|x \star \phi_\lambda| \star \phi'_\lambda$ is non-negligible only if ϕ'_λ is located at lower frequencies that ϕ_λ, i.e. when $|\lambda'| < |\lambda|$. Iterating on wavelet modulus operators thus propagates the scattering energy along frequency-decreasing paths $p = (\lambda_1, \ldots, \lambda_m)$ where $|\lambda_k| < |\lambda_{k-1}|$ for $1 \leq k < m$. Scattering coefficients along other path have a negligible energy. Numerically, it is therefore sufficient to compute the scattering transform along frequency-decreasing path only.

We see from the table above that most of the energy of scattering coefficients is concentrated in the first two layers. Here are the first two layers of scattering coefficients computed locally for an image from MNIST database:

(a)



(b)



7 Inverting the scattering transform

Preserving energy does not imply that the signal information is preserved. Since a scattering transform is calculated by iteratively applying $\mathfrak{U}$, inverting S_J requires to invert $\mathfrak{U}$. The wavelet transform $\mathcal{W}$ is a linear invertible operator, so inverting $\mathfrak{U}z = \{z \star \phi_{2^J}, |z \star \psi_\lambda|\}$ amounts to recovering the complex phases of wavelet coefficients removed by the modulus. The phase of Fourier coefficients cannot be recovered from their modulus in general, but wavelet coefficients are redundant, as opposed to Fourier coefficients. For particular wavelets, it has been proven that the phase of wavelet coefficients can be recovered from their modulus, and that $\mathfrak{U}$ has a continuous inverse.

Still, one can not exactly invert S_J because we discard information when computing the scattering coefficients $S_J[p]x = U[p] \star \phi_{2^J}$ of the last layer. Indeed, the propagated coefficients $|U[p]x \star \psi_\lambda|$ of the next layer are eliminated, because they are not invariant and have a negligible total energy. Initializing the inversion by assuming that the small coefficients in the last layer are zero produces an error. Numerical experiments on audio signals indicate that the scattering transform can be inverted successfully with good accuracy from $m = 2$ layers. Importantly, when the invariant scale 2^J becomes too large, the number of second order coefficients becomes too small, and accurate reconstruction becomes impossible. This is expected, because when 2^J is large, scattering transform destroys the information about the global position of objects.

8 Each layer is a frame expansion

Each layer in the scattering network thus consists of a filter bank:

$$\mathcal{W}_J x = \{x \star \phi_J, (W_\lambda x)_{\lambda \in \Lambda_J}\}.$$

Its norm is defines as

$$\|\mathcal{W}_J x\|^2 = \|x \star \phi_J\|^2 + \sum_{\lambda \in \Lambda_J} \|W_\lambda x\|^2.$$

The filter bank defines a frame of $L^2(\mathbb{R}^d)$ whose bounds are characterized by the following Littlewood-Paley condition:

Theorem 1. If there exists $\epsilon > 0$ such that for almost all $\omega \in \mathbb{R}^d$ and all $J \in \mathbb{Z}$

$$1 - \epsilon \leq \left|\hat{\phi}(2^J w)\right|^2 + \frac{1}{2} \sum_{j \leq J} \sum_{r \in G} \left|\hat{\psi}(2^j r w)\right|^2 \leq 1,$$

then $\mathcal{W}_J$ is a frame with bounds given by:

$$(1 - \epsilon)\|x\|^2 \leq \|\mathcal{W}_J x\|^2 \leq \|x\|^2, \; x \in L^2(\mathbb{R}^d). \tag{1}$$

9 Energy conservation

The windowed scattering representation is obtained by cascading a basic propagator operator,

$$\mathfrak{U}_J = \{x \star \phi_{2^J}, |U[\lambda]x|_{\lambda \in \Lambda_J}\}$$

The propagator $\mathfrak{U}_J$ is non-expansive, since the wavelet transform decomposition $\mathcal{W}_j$ is nonexpansive from (1) and the modulus is also nonexpansiv ($\||a| - |b|\| \leq \|a - b\|$). As a result,

$$\|\mathfrak{U}_J x - \mathfrak{U}_J x'\|^2 = \|x \star \phi_j - x' \star \phi_j\|^2 + \sum_{\lambda \in \Lambda_J} \||W_\lambda x| - |W_\lambda x'|\|^2 \leq \|x - x'\|^2$$

For any path set Ω, the Euclidean norm defined by the scattering coefficients $S_J[p], p \in \Omega$ is

$$\|S_J[\Omega]x\|^2 = \sum_{p \in \Omega} \|S_J[p]x\|^2.$$

We denote $\mathcal{P}_J$ the set of paths of any order up to scale 2^J, $\mathcal{P}_J = \cup_m \Lambda_J^m$. Since $S_J[\mathcal{P}_J]$ is constructed by cascading the non-expansive propagator $\mathfrak{U}_J$, it results that $S_J[\mathcal{P}_J]$ is also non-expansive:

Theorem 2. The windowed scattering transform is non-expansive:

$$\|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\| \leq \|x - x'\| \text{ for all } x, x' \in \mathcal{L}^2(\mathbb{R}^d).$$

In fact more is true: the scattering transform also preserves the signal energy, thus showing that all high-frequency infromation is encoded in scattering coefficients:

Theorem 3. A scattering wavelet ψ is said to be admissible if there exists $\eta \in \mathbb{R}^d$ and $\rho \in \mathcal{L}^2(\mathbb{R}^d) \geq 0$, with $|\hat{\rho}(\omega)| \leq \left|\hat{\phi}(2\omega)\right|$, $\hat{\rho}(0) = 1$, such that the function

$$\hat{\Psi}(\omega) = |\hat{\rho}(\omega - \eta)|^2 - \sum_{k=1}^{\infty} k \left(1 - \left|\hat{\rho}(2^{-k}(\omega - \eta))\right|\right)$$

satisfies

$$\inf_{1 \leq |w| \leq 2} \sum_{j=-\infty}^{\infty} \sum_{r \in G} \hat{\Psi}(2^{-j} r^{-1} \omega) \left|\hat{\Psi}(2^{-j} r^{-1} \omega)\right| > 0. \tag{2}$$

If ψ satisfies the frame condition (1) with $\epsilon = 1$ and is admissible, then for all $x \in \mathcal{L}^2(\mathbb{R}^d)$:

$$\lim_{m \to \infty} \|U[\Lambda_J^m]\|^2 = \lim_{m \to \infty} \sum_{n \geq m} \|S_J[\Lambda_J^n]x\|^2 = 0,$$

and

$$\|S_J[\mathcal{P}_J]x\| = \|x\|.$$

The proof of the theorem relies on the effect explained above that the scattering energy propagates progressively towards the low frequencies, thanks to the demodulation effect of the complex modulus.

10 Translation invariance

Theorem 4. Let $x_c(u) = x(u - c)$. Then for admissible scattering wavelets satisfying (2) we have:

$$\lim_{J \to \infty} \|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x_c\| = 0$$

for all $x \in \mathcal{L}^2(\mathbb{R}^d)$ and all $c \in \mathbb{R}^d$.

It is important to pay attention to the following. We see that scattering transform is invariant to translations. Therefore, if x and x_c are translations of one another, then $S_J x = S_J x_c$ and therefore $\|S_J x - S_J x_c\| = 0$. How is this consistent with energy conservation $\|S_J(x - x_c)\| = \|x - x_c\| \neq 0$? If S_J was a linear operator, we would have $0 = \|S_J x - S_J x_c\| = \|S_J(x - x_c)\| \neq 0$. However, importantly S_J is nonlinear and can simultaneously be translation invariant and energy preserving.

11 Lipschitz continuity to deformations

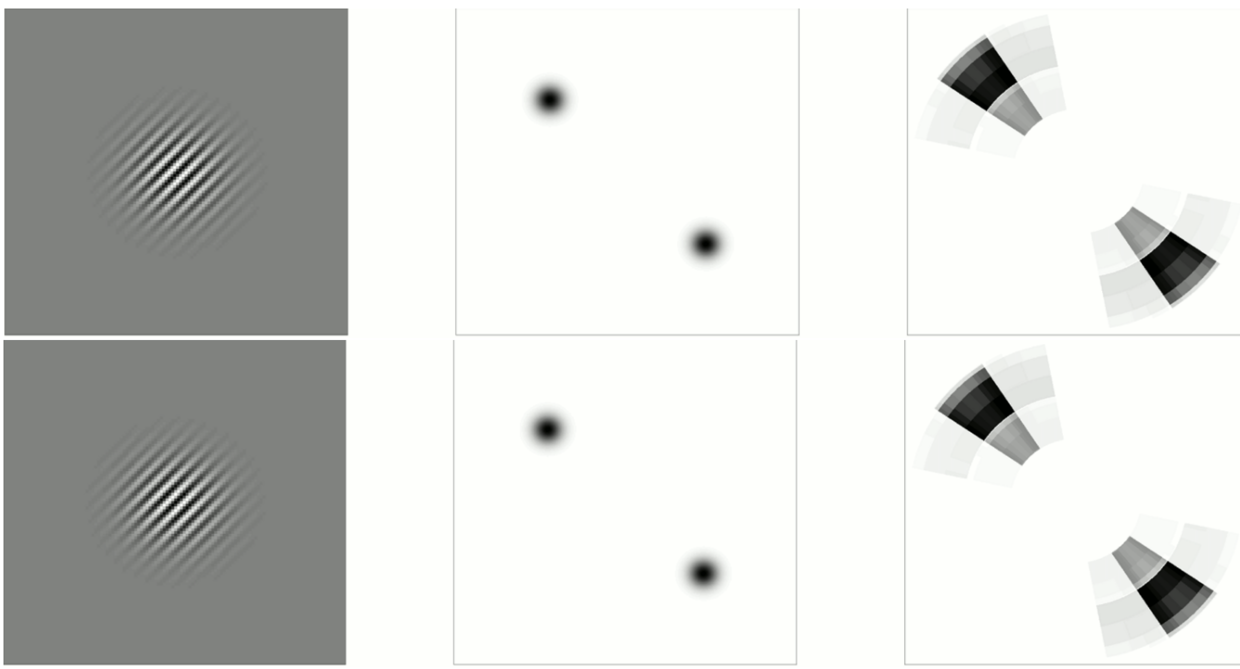
Theorem 5. For any compact set $\Omega \in \mathbb{R}^d$ there exists C such that for all $x \in \mathcal{L}^2(\mathbb{R}^d)$ supported in Ω with $\|U[\mathcal{P}_J]x\|_1 < \infty$ and for all $\tau \in C^2(\mathbb{R}^d)$ with $\|\nabla\tau\|_\infty \leq \frac{1}{2}$, then

$$\|S_J[\mathcal{P}_{J,m}]L[\tau]x - S_J[\mathcal{P}_{J,m}]x\| \leq C\|U[\mathcal{P}_J]x\|_1 \underbrace{(2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty + \|H\tau\|_\infty)}_{\text{norm of diffeomorphism}}$$

where $\mathcal{P}_{J,m} = \cup_{n<m}\Lambda_J^n$ and H is the Hessian.

This theorem shows that the size of the effect a diffeomorphism produces in the scattering domain is bounded by a term proportional to $2^{-J}\|\tau\|_\infty$, which corresponds to the local translation invariance, plus a deformation error proportional to $\|\nabla\tau\|_\infty$.

Stability to deformations is illustrated as follows:



In the two images the dots move by more than their size, but the scattering transform barely changes. A better illustraction can be found here: `https://www.di.ens.fr/data/scattering/image/`

References

[1] J. Bruna, *Scattering representations for recognition.* PhD thesis, Ecole Polytechnique, 2013.

[2] M. Andreux, T. Angles, G. Exarchakis, R. Leonarduzzi, G. Rochette, L. Thiry, J. Zarka, S. Mallat, J. Andén, E. Belilovsky, J. Bruna, V. Lostanlen, M. J. Hirn, E. Oyallon, S. Zhang, C. Cella, and M. Eickenberg, "Kymatio: Scattering transforms in python," *CoRR*, vol. abs/1812.11214, 2018.

[3] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[4] T. Wiatowski and H. Bölcskei, "A mathematical theory of deep convolutional neural networks for feature extraction," *IEEE Trans. Inf. Theory*, vol. 64, pp. 1845–1866, Mar. 2018.