Proposal for Qualifying Exam

David Weber

June 6, 2017

Exam Committee

Committee Chairperson:
Prof. James Bremer

Committee Members:
Prof. Naoki Saito
Prof. Thomas Strohmer
Prof. Yong Jae Lee
Prof. James Sharpnack

Exam Logistics

Date: Friday June 16th, 2017
Time: 11:00 AM
Location: MSB 2240

Proposed Research Talk

Title. Convolutional Neural Networks, the Scattering Transform, and Extensions

Abstract. Convolutional Neural Networks (CNNs) are a type of deep neural network which have performed well at image and audio classification. One approach to understanding the success of CNNs is Mallat’s scattering transform, which formalizes the observation that the filters learned by a CNN have wavelet-like structure. The resulting transform generates a representation that is approximately translation invariant and slowly varying to a wide class of deformations of the input patterns. These features are then used by a linear classifier to perform classification.

In this talk, we will explore the application of the scattering transform and a new variant to understanding the classification of objects using sonar. As we have an explicit model with a fast simulator for this problem, it provides a good test problem for understanding the properties of the scattering transform in the object domain. We obtain over 95% classification accuracy on binary discriminating shape and speed in the case of synthetic data generated by the fast simulator, and a 95% detection rate for unexploded ordinance, with a false positive of 25% in the case of real data.

We construct a new variant of the scattering transform, the shearlet scattering transform (shattering transform), using the shearlet frame instead of Morlet wavelets that are often used in the standard scattering transform. This 2D transform has the advantage of sparsifying the signals containing curvilinear singularities rather than point singularities.

Introduction

Deep neural networks, and convolutional neural networks in particular have proven quite effective at discerning hierarchical patterns in large datasets [LeCun et al., 2015]. Some examples include image classification [Krizhevsky et al., 2012], face recognition [Sun et al., 2013], speech recognition [Dahl et al., 2013], and the game of Go [Silver et al., 2016], among many others. Clearly, something is going very right in the design of CNNs. However, the principles that account for this success are still somewhat elusive. Briefly, a CNN is a deep neural network where each layer is composed of a number of convolutions. Beyond this broad shared
framework, there are many variations, such as the shape of the convolution, the introduction of max or min pooling, the depth, the kinds of nonlinearities, etc. Each of these features can be varied between layers in a dizzying array of parameters and hyper-parameters.

The idea of a CNN is not new; its predecessor, the Neocognitron, was invented in 1980 by K Fukushima [Fukushima, 1980]. Its structure was inspired by the structure of the visual system, in particular V1, where the layers of simple and complex cells form columns of locally connected structure, which are responsive to increasingly invariant representations [Hubel and Wiesel, 1968]. The first part of the visual cortex, preceding V1, has receptive fields that consist of regions of activation surrounded by inactivation. The simple cells, consist of edge detectors, where each cell has a particular location and orientation [Hubel et al., 1995]. The complex cells have less locational sensitivity, and a sensitivity to motion.

The ties between wavelets and the visual system have been developed for decades. In the early 80’s, several authors demonstrated that the simple cells are well modeled by Gabor filters (e.g. [Daugman, 1980] [Marcelja, 1980]). This has motivated the use of images generated using Gabor filters to determine properties of human visual processing; for example Field used Gabor filters to demonstrate that human vision relies on continuity to detect features [Field et al., 1993]. In the 90’s, this Gabor-like structure was experimentally demonstrated to develop sparsity [Field, 1987] [Olshausen and Field, 1996].

Harmonic analysis has come and gone in the neural network literature, much like the field as a whole; in the 90’s, when networks of only one hidden layer were the norm, there was some success in developing wavelet neural networks, where either the data was first processed with wavelets, or the dilation and translation parameters for the wavelets were trained simultaneously with the neuron weights [Szu et al., 1992] [Zhang, 1997].

In 2012, Stephane Mallat and Joan Bruna published both theoretical results [Mallat, 2012] and numerical implementations [Bruna and Mallat, 2011] tying together convolutional neural networks and wavelet theory. They demonstrated that scattering networks of wavelets and modulus nonlinearities, are translation invariant in the limit of infinite scale, and Lipschitz continuous under non-uniform translation, i.e. $T_\tau(f)(x) = f(x - \tau(x))$ for $\tau$ with bounded gradient. Numerically, they achieved state of the art on image and texture classification problems.

More recent work from Wiatowski and Bölcskei at ETH Zürich have generalized the Lipschitz continuity result from wavelet transforms to frames, and more importantly, established that increasing the depth of the network also leads to translation invariant features [Wiatowski and Bölcskei, 2015]. There have been a number of follow up papers, including a discrete version of Wiatowski’s result [Wiatowski et al., 2016], a related method on graphs [Chen et al., 2014], and an inverse problem used to solve the phase retrieval problem [Waldspurger et al., 2015]. There have been a number of papers using the scattering transform in such problems as fetal heart rate classification [Chudáček et al., 2014], age estimation from face images [Chang and Chen, 2015], and voice detection in the presence of transient noise [Dov and Cohen, 2014].

In this talk, I will apply both the scattering transform of Mallat et al. and a version based off of the theory of Wiatowski et al. to the problem of classification of underwater objects using sonar. This is motivated by previous work done by Professor Saito on the detection of unexploded ordinance using synthetic aperture sonar [Marchand et al., 2007]. Coming from this, we have a dataset of 14 different objects at various distances and rotations, partially submerged on the ocean floor, as shown in Fig. 1. Additionally, we can generate synthetic examples using a fast solver for the Helmholtz equation in two regions of differing speed, provided by Dr. Ian Sammis and Professor Bremer. We use this to examine more closely the dependence of both classification and the output of the scattering transform on both material properties and shape variations.

The approximate translation invariance and Lipschitz continuity under small diffeomorphisms of the scattering transform mean that for classes which are invariant under these transformations, members of the same class will be close together in the resulting space. So long as transforming from one class to another requires a $\tau$ with large derivative, then the classes will be well separated. Accordingly we use a linear classifier on the output of the scattering transform. Additionally, because the scattering transform concentrates energy

![Figure 1: Example targets on shore and in the target environment, from Kargl, 2015](image-url)
at lower scales and wavelets in general encourage sparsity for smooth signals with singularities, we use a sparse version of logistic regression as our linear classifier called LASSO [Hastie et al., 2015]. Specifically, we use a julia wrapper around the glmnet code of [Friedman et al., 2010]. As a baseline classification scheme to compare the classification performance, we use logistic regression on the absolute value of the Fourier transform of the signal.

Deep Convolutional Neural Networks

There are many ways to put together a CNN. For simplicity of presentation, we will consider data $U_0 \in \mathbb{R}^{D_0 \times 1}$ which is 1D, such as audio signals. Instead of a vector, $U_0$ could be a matrix (images), or a tensor (volumetric data), etc. with only minor changes. At layer $m$, there are $K_m$ filters, or receptive fields, $\{G_m^i\}_{i=1}^{K_m} \subset \mathbb{R}^{D_m \times K_{m-1}}$. A small number of adjacent rows of $G_m^i$ are non-zero, corresponding to small local support in the input. There is some Lipschitz continuous nonlinearity $M_m$ applied elementwise, such as a rectified linear unit (ReLU), hyperbolic tangent, or logistic sigmoid that is to the output of the convolution. Finally, there is some pooling function $R_m : \mathbb{R}^{D_m \times K_m} \to \mathbb{R}^{D_m \times K_m}$, which reduces $D_m \approx D_{m-1}/2$, often by returning only the maximum of $2^1$ elements, or their mean. This plays the role of either striding or pooling in the literature. Putting all of these together, we can recursively calculate the matrix $U_m$ at layer $m$ via

$$U_m = R_m \left( M_m \left( G_m^1 \ast_c U_{m-1} \right), \ldots, M_m \left( G_m^{K_m} \ast_c U_{m-1} \right) \right).$$

where by $\ast_c$ performs convolution occurs strictly along the columns, and then sums the result, i.e. if $g_1, \ldots, g_{K_{m-1}}$ are the columns of $G_m^i$ and $u_1, \ldots, u_{K_{m-1}}$, the columns of $U_{m-1}$, then

$$G_m^i \ast U_{m-1} := \sum_{i=1}^{K_{m-1}} g_i \ast u_i$$

The weights in the receptive fields $G_m^i$ are learned, but the support size is fixed. A single row $U_m$ corresponds to choosing a path consisting of a single filter from each layer.

Say we have $M$ layers, with the last layer being fully connected instead of having the above structure, and we have $N$ examples $\{U_0^j\}_{j=1}^N \subset \mathbb{R}^{D_0 \times 1}$ with labels $\{y_j\}_{j=1}^N \subset \mathbb{R}^{D_M \times 1}$, where $D_M$ is the number of classes. A commonly used objective is [Krizhevsky et al., 2012]

$$\min_{G_m^i} \sum_{j=1}^K \|y_j - U_M\|_2^2 + \gamma \sum_{n=1}^N \sum_{i=1}^{D_m} \|G_m^n\|_2^2$$

The first term corresponds to minimizing the error, while the second is a constraint on the total weight size; [Krizhevsky et al., 2012] found the inclusion of the weight constraint essential to training.

Previous Results

A generalized scattering network (hereafter referred to simply as a scattering network), as defined in [Wiatowski and Bölcskei, 2015], has an architecture that is a continuous analog of a CNN. For a diagram, see Fig. 4. First, at layer $m$, we start with a family of generators $\{g_{\lambda_m}\}_{\lambda_m \in \Lambda_m} \subset \mathcal{L}^1(\mathbb{R}^d) \cap \mathcal{L}^2(\mathbb{R}^d)$ for a translation invariant frame $\Psi_m = \{\psi_{b,\lambda_m}\}_{b \in \mathbb{Z}^d, \lambda_m \in \Lambda_m} = \{T_t g_{\lambda_m}\}_{b \in \mathbb{Z}^d, \lambda_m \in \Lambda_m}$, with frame bounds $a_m$ and $b_m$, that is

$$a_m \|f\|_2^2 \leq \sum_{\lambda_m \in \Lambda_m} \|T_t g_{\lambda_m}, f\|_2^2 \leq b_m \|f\|_2^2$$

for all $f \in \mathcal{L}^2(\mathbb{R}^d)$. Here, $T_t[f](x) = f(x-t)$ is the translation operator, $I[f](x) = f(-x)$ is the involution operator, and $\Lambda_m$ is some countable discrete index set, such as $\mathbb{Z}$. This index set needs to tile the frequency plane in some way, for example by indexing scales and rotations. The frames $g_{\lambda_m}$ correspond to the receptive

1 when we increase the input dimension, this will go up to $2^d$ for input dimension $d$
2 A more thorough exposition can be found at arxiv:1512.06293 [cs.IT].
Given the conditions above, for Theorem 2.

Theorem 1. For frequency shift weak admissibility condition and then they show that proof, see [Wiatowski and Bölcskei, 2015] Appendix F, G and H; first they show that \( \Phi_f \) is non-expansive, and defining the generator corresponding to index \( \lambda = (j, h) \) is \( g_x = a^{\|\cdot\|}(a^j h^{-1} x) \).

To get from layer \( m \) to layer \( m+1 \), we define the function \( u_m : \Lambda_m \to \mathbb{R}^d \) by
\[
  u_m(\lambda_m)(f(z)) = m_m[f \ast g_{\lambda_m}](r_m z)
\]
Using this to define the output at a layer coming along a path of indices \( q \in \Lambda_m := \Lambda_1 \times \ldots \times \Lambda_m \). Then the output at layer \( m \) is given by the path operator \( u \), defined recursively using \( u_i \) by
\[
  u(q)(f) = u_m(\lambda_m)u_{m-1}(\lambda_{m-1}) \cdots u_1(\lambda_1)f
\]
Choosing \( d = 1 \) corresponds to audio as in the CNN section. Unlike a CNN, a scattering network has infinite depth, and thus each layer has output. The output is created by choosing some \( \lambda_m \in \Lambda_m \), and defining the averaging function \( \chi_{n-1} : \Lambda_m \to \mathbb{R}^d \), which is removed from the set of generating functions. The resulting set of features at the \( n \)th layer is \( \Phi_m[f] := \left\{ u[q]f \ast \chi_{n-1} \right\}_{q \in \Lambda_n} \) and for the entire network \( \Phi[f] := \left\{ \Phi_m[f] \right\}_{m=0}^{\infty} \).

The chief difference between a scattering transform and a CNN is the summation that occurs across filters in a CNN as in Eq. (1). In a scattering transform, each filter is only applied to a specific path of filters in the layer below, generating a tree-like structure; it is possible to arrive at this in a CNN by simply fixing all but one column to 0.

There are two additional conditions that restrict the various operators in a given layer simultaneously. The first is the weak admissibility condition, which requires that the upper frame bound \( b_m \) be sufficiently small compared to the sub-sampling factor and Lipschitz constants:
\[
  \max \left\{ b_m, \frac{b_m \gamma_m^2}{r_m^d} \right\} \leq 1 \tag{3}
\]
this can be achieved by scaling \( g_m \), and suggests a possible role of the weight constraint in Eq. (2).

The second is that the nonlinearities must commute with the translation operator, so \( m_m T_\tau[f] = T_\tau m_m[f] \). Most nonlinearities used for CNN’s are pointwise i.e., \( m_m[f](x) = \rho_m(f(x)) \), so they certainly commute with \( T_\tau \). Given these constraints, we can now state the results of [Wiatowski and Bölcskei, 2015] precisely; the first is that the resulting features \( \Phi_m \) deform stably with respect to small frequency and space deformations:

**Theorem 1.** For frequency shift \( \omega \in C(\mathbb{R}^d, \mathbb{R}) \) and space shift \( \tau \in C^1(\mathbb{R}^d, \mathbb{R}) \), define the operator \( F_{\tau,\omega}[f](x) := e^{i\tau \cdot \omega}(x)f(x - \tau(x)) \). If \( \|D\tau\|_{\infty} \leq \frac{1}{2b} \), then there exists a \( C > 0 \) independent of the choice of \( \Phi \) s.t. for all \( f \in \mathcal{L}^2(\mathbb{R}^d) \),
\[
  \|\Phi[F_{\tau,\omega}f] - \Phi[f]\|_2 := \left| \sum_{n=1}^{\infty} \sum_{q \in \Lambda_n} \| \chi_n \ast u[q](F_{\tau,\omega}f) - \chi_n \ast u[q]f \|_2 \right| \leq C(\|\tau\|_{\infty} + \|\omega\|_{\infty})\|f\|_2 \tag{4}
\]
where \( \mathcal{L}^2(\mathbb{R}^d) \) is the set of \( \mathcal{L}^2(\mathbb{R}^d) \) functions whose Fourier transforms are band limited to \([-a, a] \).

Their next result, translation invariance that increases with depth, is distinct from the one shown by Mallat where translation invariance increases with resolution:

**Theorem 2.** Given the conditions above, for \( f \in \mathcal{L}^2(\mathbb{R}^d) \) the \( n \)th layer’s features satisfy
\[
  \Phi_m[T_{\tau}f] = T_{\tau,\cdot m} \left[ \Phi_m[f] \right] \tag{5}
\]
If there is also a global bound $K$ on the decay of the Fourier transforms of the output features $\chi_m$: 

$$|\hat{\chi}_m(\omega)| \leq K$$

then we have the stronger result 

$$\sum_{k=1}^{n} \|\Phi_m[T_t f] - \Phi_m[f]\|_2 \leq \frac{2\pi |t| K}{r_1 \cdots r_m} \|f\|_2.$$ 

The proof for Eq. (5) follows almost immediately from the condition that $m_m$ and $T_t$ commute: first note that 

$$u_m[\lambda_m] T_t f(x) = m_m[T_t f * g_{\lambda_m}](r_m x) = m_m[T_t (f * g_{\lambda_m})](r_m x)$$

$$= m_m((f * g_{\lambda_m})(r_m x - t) = m_m((f * g_{\lambda_m}) (r_m (x - \frac{t}{r_m})))$$

$$= T_t / r_m u_m[\lambda_m]$$

Then Eq. (5) follows by repeated application along the path $q$. For the second result see Appendix I and E of [Wiatowski and Bölcskei, 2015]. This is more technical, and relies on Parseval’s formula, the previous result, the operator norm of $u_m[\lambda_m]$, and the weak admissibility condition.

To compare with the result from [Mallat, 2012], that says that for an admissible mother wavelet $\psi$, the scattering transform achieves perfect translation invariance as the lower bound on the scale goes to infinity:

$$\lim_{J \to \infty} \|\Phi_J[T_t f] - \Phi_J[f]\| = 0$$

### Sonar Detection

The problem that we will be investigating with the scattering transform techniques is the classification of objects partially buried on the sea floor using sonar data. This is motivated by using an unmanned underwater vehicle (UUV) equipped with sonar to detect unexploded ordinance. For this problem, we have both real and synthetic examples. The real examples consist of 14 objects buried at various distances and rotations in a well groomed sandy shallow oceanbed. The path of a UUV was simulated using a rail, and at short intervals along it the field of objects was pinged. So for each rotation of the object relative to the rail, there is a 2D wavefield, where each signal corresponds to a location on the rail and the observation time.

The synthetic examples come from considering the 2D Helmholtz equation regions of differing speeds:

$$\Delta u + k_i^2 u = 0 \quad \text{in } \Omega$$

$$\Delta v + k_i^2 v = 0 \quad \text{in } \Omega^c$$

$$u - v = g \quad \text{on } \partial \Omega$$

$$\partial_n u - \partial_n v = \partial_n g \quad \text{on } \partial \Omega$$

$$\sqrt{|x|} \left( \partial_{|x|} - i k_i \right) v(x) \to 0 \text{ as } |x| \to \infty,$$

which gives the response to a sinusoidal signal with frequency $\omega$ on an object with $k_i = \omega / c_i$, where $c_i$ is the speed of sound in the material, ranging from $343 m/s$ in air, to $1503 m/s$ in water, to $5100 m/s$ in aluminum. It is worth noting that this is idealized in several ways: the model we use is 2D, rather than 3D, the material is modelled as a fluid with only one layer, instead of a solid with multiple different components, and there is no representation of the ocean floor itself.

The signals sent out by the UUVs are not pure sinusoids. One can approximate the response to multifrequency signals (e.g. Gabor functions or chirps) by integrating across frequencies. We use a fast solver created by Ian Sammis and James Bremer to synthesize a set of examples, where we can more explicitly test the dependence of the scattering transform on the material properties (represented by the speed) and geometry. The current repository we use was created by Vincent Bodin, a former summer intern supervised by Professor Saito.
There are two ways we have approached classifying these sonar waveforms. Since at each rail location a 1D waveform is recorded, one can either treat each point in space as a different 1D signal, or gather all points on the rail in a single 2D wavefield. In the 1D case, there are far more examples (on the order of 1600 per waveform for the real case, or 480 for the synthetic), but less invariant structure. Specifically, translation in the object domain doesn’t correspond to a translation in the signal domain for the 1D case. The only reason we can expect any success from the scattering transform is the first theorem above of Lipschitz change under diffeomorphism. In the 2D case, translation parallel to the observation rail in the object domain corresponds to translation in the signal domain. However, translation not parallel to the rail modulates both amplitude and location in the signal domain. So the 2D case still relies on the signals not having too large a diffeomorphism to change between examples in the same class.

For the synthetic data, there are two primary problems of interest. The first is determining the importance of varying shape on the scattering transform. An example of the 1D scattering transform for a triangle is in Fig. 2. Since the energy at each layer decays exponentially, layers 1-3 have been scaled to match the intensity of the first layer. Note that only the zeroth layer has negative values; this is because the nonlinearity used by the scattering transform is an absolute value. In the figure, one can clearly see a time concentrated portion of the signal. At deeper layers, the energy is concentrated at the coarsest scale. The scattering transform was able to successfully discriminate between these similar shapes at over a 80% rate, while the AVFT achieved a 70%. While tractable, this problem is more difficult than the other problem, determining the importance of material. Even the AVFT was successful on this problem, achieving a 96% classification accuracy. The most important coordinate for the AVFT was the average. Initially, the scattering transform had the same results as the AVFT, but increasing the quality factor and the depth of the scattering transform improved its accuracy to 99%.

For the real dataset, the problem of interest is somewhat more ambiguous. In addition to a set of UXO’s and a set of arbitrary objects, there are some UXO replicas, not all of which are made of the same material. As we saw in the synthetic case, the difference in speed has a much clearer effect on classification accuracy than shape, so it is somewhat ambiguous how to treat these. However, even including these replicas in the UXO category gave us a false negative rate of 5% and a false positive rate of 24% (or 94% correctly identified UXO’s and 75% correctly identified others); in comparison, the AVFT had a false negative rate of 9% and a false positive rate of 50% (so no better than random guessing). It is not terribly surprising that the non-UXO’s are harder to classify as a group using a linear classifier, since they don’t share much in common—a SCUBA tank bears more resemblance to a UXO than it does to a rock.

**Shearlets**

Previous versions of the scattering transform have relied primarily on Morlet wavelets. A problem with this approach is that tensor products of wavelets are best suited to detect point singularities. However, in the sonar problem, and most images, the most relevant singularities are two dimensional curves, meaning that...
representations using tensor product wavelets will be much denser than is conceivably necessary. A natural solution to this problem are curvelets, which use a rotation operator and a parabolic scaling matrix:

\[ A_j = \begin{pmatrix} 2^j & 0 \\ 0 & 2^j \alpha/2 \end{pmatrix} \]

where \( \alpha \in (0, 2) \) controls the degree of anisotropy. If \( R_\theta \) is a rotation by \( \theta \), then curvelets are defined by \( \psi_{j, \theta}(x) = 2^{j/2} \psi(\mathcal{S}_k A_j x) \) [Candès, 2003]. Curvlets have been shown to be almost optimal in terms of sparsity for cartoon-like images (those which are smooth, except along a simple, differentiable curve) [Candès and Donoho, 2004]. However these suffer from numerical instability, since discritization doesn’t commute with the rotation operator for non-trivial rotations. Thus the shearlet was created, which uses shearing matrices, with \( k \in \mathbb{Z} \) and spacing factor \( c \):

\[ S_k = \begin{pmatrix} 1 & c^k \\ 0 & 1 \end{pmatrix} \]

instead of rotation matrices. Then if one chooses \( \psi \) and \( \phi \) appropriately, the resulting shearlet system is

\[ \psi_{j,k}(x) = 2^{j/4} \psi(\mathcal{S}_k A_j x) \]

which is convolved with the input to generate the transform. As \( k \to \infty \), \( \psi_{j,k} \) tiles the frequency plane. However, it does so with a directional bias, so when we truncate at finite \( k \), more of the first frequency axis is covered than the second. To counter this, two cones of shearlets are used, with the roles of \( x_1 \) and \( x_2 \) reversed; for example, in Fig. 3, \( \psi_{1,1,1} \) and \( \psi_{2,1,1} \) have the same scale and shearing, but different cones. The combination of \( S_k \) and \( A_j \) leave the discritization lattice intact if chosen well. The resulting transform is the compactly supported universal shearlet transform. The shearlets used in this paper are defined so that they are approximately classical shearlets:

\[ \hat{\psi}(\xi_1, \xi_2) \approx \psi_1(\xi_1) \psi_2(\xi_2/\xi_1) \]

\[ \sum_{j \in \mathbb{Z}} |\hat{\psi}_1(2^j \xi)|^2 = 1 \text{ for a.e. } \xi \in \mathbb{R} \]

\[ |\hat{\psi}_2(\xi - 1)|^2 + |\hat{\psi}_2(\xi)|^2 + |\hat{\psi}_2(\xi + 1)|^2 = 1 \text{ for a.e. } \xi \in [-1, 1] \]

i.e. \( \psi_1 \) satisfies a discrete Calderón condition and \( \{\hat{\psi}_2(\xi - \ell)\}_{\ell \in \mathbb{Z}} \) is a partition of unity. This results in wedge-like essential support in the Fourier domain. Further, the construction means that the shearlets generated by \( \psi \) are compactly supported, at the cost of only being approximately classical. More details on the construction can be found in [Kutyniok et al., 2016], and on shearlets in the text [Kutyniok and Labate, 2012].

**Shattering Transform**

With Shearlets introduced, we can now define the *shattering transform*, created by Professor Saito and myself. In each layer of the shattering transform, we use universal shearlet generated by shearlab [Kutyniok et al., 2016]. The nonlinearity we use in these examples is the ReLU, which has a Lipschitz constant of 1 when all variables are real, and we subsample by a factor of 2 every layer. Estimating frame bounds can be difficult; [Kittipoom et al., 2012] have provided some estimates for the class used by shearlab. For the purpose of this talk, we will assume that the frame bound \( b_m \leq 1 \), so because the Lipschitz constant of the ReLU is 1, any subsampling rate larger than 1 will satisfy Eq. (3). In this talk, I will present comparisons between using the shattering transform, a tensor product version of Morlet wavelet scattering transform, and the absolute value of the Fourier transform on the synthetic data set. Currently, there are implementation issues with the shattering transform on examples as large as the real example.
My Contributions

The first project that I undertook was using the scattering transform as implemented in MATLAB by the Mallat lab on the real dataset to classify whether an object is unexploded ordinance or not, using the 1D approach. The scattering transform performed significantly better than using the AVFT across 10-fold cross validation on the real dataset, as well as on shape and material discrimination in the synthetic case. I created a new way of visualizing the results for the 1D scattering transform, both for selected coordinates and the actual data, as in Fig. 2. In the synthetic case, I have created a Julia wrapper for the Helmholtz equation solver to facilitate the creation of new examples.

For the 2D case, I wrote the Shearlet scattering transform using Shearlab for the shearlet transform in Julia, and compared its success at classification with a Morlet scattering transform and a 2D AVFT. On shape discrimination in the synthetic dataset, the shattering transform improved classification from 91% to 95%, using $2/3$ as many coefficients.

Future Plans

In addition to continued exploration of the shattering transform, there are a number of related problems I hope to investigate:

One theoretical problem that arises naturally in the synthetic data is the distinction between invariance properties in the object domain instead of the signal domain. The results thus far discuss translations and diffeomorphisms in the signal, but the real invariances of class in the sonar problem arise from translation and rotation of the object. To do so may require examining a different diffeomorphism in place of $F_{\omega, \tau}(f)(x) = e^{i\omega(x)}f(x - \tau(x))$, since amplitude shifts can’t necessarily be written in this form. To see this, consider $f(x) = e^{x^2}$, and $g(x) = Af(x)$, with $A > 1$. Then if we attempt to represent $g(x) = F_{\omega, \tau}(f)(x)$, then $\omega(x) = 0$, and $\tau$ is one of $x \pm \sqrt{x^2 - \log A}$, which is imaginary for $x \in (-\log A)^{1/2}, (\log A)^{1/2})$. This simple example, combined with the usefulness of the scattering transform in the sonar problem suggests that the scattering transform is robust to a more general diffeomorphism which can represent the transformations in the object domain for the sonar scattering problem.

While the Helmholtz equation is particular to sonar scattering, there are other problems which arise from observing the result of a PDE, such as photographs, seismic data, or audio. The simplification of the Helmholtz equation above serves as one example of a class of inverse problems, where we are trying to determine categorical information about an object by observing it indirectly. Providing bounds in terms of transformations in the object domain, rather than the signal domain, more naturally reflects the properties used for classification. Generalizing this in some way would speak to the effectiveness of CNNs in image recognition.

Another theoretical problem is the lack of lower bounds for the scattering transform. Neither Mallat and his group, nor Wiatowski & Bölcskei give lower bounds on $\|\Phi[F_{\omega, \tau}f] - \Phi[f]\|$ in terms of $\tau$ and $\omega$. If we can do so, this provides guarantees for when different classes will be well separated.

An essential step for the scattering transform is to show that it satisfies the frame bounds Eq. (3). As mentioned above, for the frames from shearlab, there are some loose frame bounds provided in [Kittipoom et al., 2012]. It should be a fairly straightforward problem to adapt these to prove Eq. (3). Additionally, for discrete frames, because the optimal frame bounds correspond to the eigenvalues of the frame operator, this enables us to explicitly test if a given frame satisfies Eq. (3), so long as the dimension is sufficiently small [Christensen, 2003].

CNNs and scattering transforms work well on Euclidian domains. However, there is much more data that is representable on graphs, such as epidemiological problems, trade networks, and fMRI, to name a few. There is an emerging literature of wavelets and harmonic analysis on graphs [Shuman et al., 2013], and a natural extension of the scattering transform is to apply it to graph domains using an existing graph wavelet transform, such as the generalized Haar-Walsh transform of [Irion and Saito, 2014].
Figure 4: Generalized scattering transform
Proposed Exam Syllabus

Harmonic Analysis (MAT 271):

Fourier Analysis [Hunter and Nachtergaele, 2001] [Mallat, 2009] [Stein and Shakarchi, 2003b]
Fourier transform, Fourier series, convergence, DFT, DCT, FFT, Shannon-Nyquest sampling theorem, Parseval/Plancherel theorems, Riemann-Lebesgue lemma, discrete & continuous uncertainty principles

Frames [Christensen, 2003]
Definition, frames of translates, dual frames, Gabor frames, short time Fourier transform,

Wavelets [Mallat, 2009]
Haar transform, Walsh transform, Multi-resolution analysis, Shannon-Littlewood-Paley wavelets, DWT, curvelets, shearlets, wavelet packets

Optimization & Machine learning (MAT 258A, MAT226A, MAT280):

Optimization [Boyd and Vandenberghe, 2004]
Lagrangian Duality, KKT conditions, linear programming, conjugate gradient, quasi-Newton, BFGS, augmented Lagrangian, Wolfe conditions.

Numerical Linear Algebra [Horn and Johnson, 2012]
QR factorization, LU decomposition, SVD, conditioning, algorithmic stability, Cholesky factorization, Krylov subspace, power method for eigenvalues

Machine learning [Bishop, 2006] [Hastie et al., 2015]
Support Vector Machines, Logistic Regression, Generative vs Discriminative models (linear discriminant analysis), neural networks, mutual information, relative entropy, Markov random field

Mathematical foundations of Big Data : (MAT280 Strohmer notes)
Principal component analysis, concentration of measure, clustering/community detection techniques, compressed sensing, lasso, basic spectral graph theory, diffusion maps

Analysis (MAT201AB, MAT205):

Real [Rudin, 1987][Hunter and Nachtergaele, 2001]
Banach spaces, Hilbert spaces, dominated convergence theorem, monotone convergence theorem, reproducing kernel Hilbert spaces.

Complex [Stein and Shakarchi, 2003a]:
Cauchy’s theorem, residue theorem, holomorphic, meromorphic, argument principle, Paley-Wiener theorem, Gamma function, conformal mapping

Probability & Statistics (MAT235AB) [Kallenberg, 1997]:
Basic inequalities (Markov’s, Chebyshev’s, etc), Borel-Cantelli lemma, Markov process, convergence in probability vs distribution vs $L^p$, strong law of large numbers, Portmanteau theorem

References


10


