Sparse Approximation: from Image Restoration to High Dimensional Classification

Bin Dong

Beijing International Center for Mathematical Research
Beijing Institute of Big Data Research
Peking University

Workshop on Frame Theory and Sparse Representation for Complex Data
Institute of Mathematical Science
National University of Singapore
29 May - 2 June 2017
Outlines

I. Brief review of image restoration models

II. Wavelet frame transforms and differential operators under variational and PDE framework

III. Sparse approximation for high-dimensional data classification

IV. Conclusions and Future work
Image Restoration Model

- Image Restoration Problems

\[ f = Au + \eta \]

- Denoising, when \( A \) is identity operator
- Deblurring, when \( A \) is some blurring operator
- Inpainting, when \( A \) is some restriction operator
- CT/MR Imaging, when \( A \) is partial Radon/Fourier transform

- Challenges: large-scale & ill-posed
Image Restoration Models: A Quick Review

- Image restoration: \( f = Au + \eta \)
- Variational and Optimization Models
  \[
  \min_u \lambda R(u) + \| Au - f \|^2
  \]
  - Total variation (TV) and generalizations: \( R(u) = \| \nabla u \|_1 \) or \( \| Du \|_1 \)
  - Wavelet frame based: \( R(u) = \| Wu \|_1 \) or \( \| Wu \|_0 \)
  - Others: total generalized variation, low rank, NLM, BM3D, dictionary learning, etc.
- PDEs and Iterative Algorithms
  - Perona-Malik equation, shock-filtering (Rudin & Osher), etc
  \[
  u_t = \sum_{\ell=1}^{L} \frac{\partial \alpha_\ell}{\partial x} \Phi_\ell(Du, u) - A^*(Au - f), \quad \text{with } D = \left( \frac{\partial \beta_1}{\partial x_1}, \ldots, \frac{\partial \beta_L}{\partial x_L} \right)
  \]
  - Iterative shrinkage algorithm
  \[
  u^k = \widetilde{W}^\top S_{\alpha_{k-1}}(Wu^{k-1}) - A^\top (Au^{k-1} - f), \quad k = 1, 2, \ldots
  \]
- What do they have in common?

Bridging discrete and continuum

WAVELET FRAME TRANSFORMS
AND DIFFERENTIAL OPERATORS

MRA-Based Tight Wavelet Frames

- Refinable and wavelet functions
  \[ \phi = 2^d \sum a_0[k] \phi(2 \cdot -k) \quad \psi_\ell = 2^d \sum a_{\ell}[k] \phi(2 \cdot -k), \quad \ell = 1, 2, \ldots, q. \]

- Unitary extension principle (UEP)
  \[ \sum_{\ell=0}^{q} |\hat{a}_\ell(\xi)|^2 = 1 \quad \text{and} \quad \sum_{\ell=0}^{q} \hat{a}_\ell(\xi)\hat{a}_\ell(\xi + \nu) = 0, \]
  \[ \nu \in \{0, \pi\}^d \setminus \{0\} \text{ and } \xi \in [-\pi, \pi]^d \]

- Discrete 2D transformation: \( W u = \{W_{l,i} u : 0 \leq l \leq L - 1, 0 \leq i_1, i_2 \leq r \} \)
  \[ W_{l,i} u := a_{l,i}[-\cdot] \otimes u, \]

- Perfect reconstruction: \( W^T W = I \)

- Further reading: [Dong and Shen, MRA-Based Wavelet Frames and Applications, IAS Lecture Notes Series, 2011]
Connections: Motivation

- Difference operators in wavelet frame transform:

\[
\begin{align*}
    h_{0,1} &= \frac{1}{4} \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}, \quad h_{1,0} = \frac{1}{4} \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}, \quad h_{1,1} = \frac{1}{4} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \\
    \text{Transform} \quad W u &= \{ h_{0,1}[-\cdot] \ast u; h_{1,0}[-\cdot] \ast u; h_{1,1}[-\cdot] \ast u \} \\
    \text{Approximation} \\
    h_{0,1}[-\cdot] \ast u &\approx \frac{1}{2} \delta u_x, \quad h_{1,0}[-\cdot] \ast u \approx \frac{1}{2} \delta u_y, \quad h_{1,1}[-\cdot] \ast u \approx \frac{1}{4} \delta^2 u_{xy}
\end{align*}
\]

- Thus,

\[
\frac{2}{\delta} W u \approx \nabla u
\]

\[
|\nabla u| \approx \left( \frac{1}{4} \left[ (D_x^+ u_{i,j})^2 + (D_x^+ u_{i,j+1})^2 + (D_y^+ u_{i,j})^2 + (D_y^+ u_{i+1,j})^2 \right] \\
+ \left[ \frac{(D_x^+ u_{i,j} + D_y^- u_{i,j+1})^2}{4} + \frac{(D_x^+ u_{i,j} + D_y^+ u_{i+1,j})^2}{4} \right] \right)^{1/2}
\]

- More rigorously [Choi, Dong and Zhang, preprint, 2017]

**Proposition 2.2**: Let a tensor product framelet function \( \psi_\alpha \in L_2(\mathbb{R}^2) \) have vanishing moments of order \( \alpha \) with \( |\alpha| \leq s \), and let \( \text{supp} (\psi_\alpha) = [a_1, a_2] \times [b_1, b_2] \). For \( n \in \mathbb{N} \) and \( k \in \mathbb{Z}^2 \) with \( \text{supp}(\psi_{\alpha,n-1,k}) \subseteq \Omega \), we have

\[
\langle u, \psi_{\alpha,n-1,k} \rangle = (-1)^{|\alpha|2} |\alpha|^{1-n} \langle \partial^\alpha u, \varphi_{\alpha,n-1,k} \rangle
\]

for every \( u \in W^s_1(\Omega) \).

\[
\int_{\mathbb{R}^2} \varphi_\alpha dx \neq 0, \quad \text{supp}(\varphi_\alpha) = \text{supp}(\psi_\alpha)
\]
Connections: Analysis Based Model and Variational Model

- [Cai, Dong, Osher and Shen, JAMS, 2012]:

\[ \lambda \| W u \|_1 + \frac{1}{2} \| Au - f \|_2^2 \rightarrow \lambda \| D(u) \|_1 + \frac{1}{2} \| Au - f \|_2^2 \]

For any differential operator when proper parameter is chosen.

**Theorem.** Let the objective functionals of the analysis based model and the variational model be \( E_n(u) \) and \( E(u) \) respectively, then:

1. \( E_n(u) \rightarrow E(u) \) for each \( u \in W^s_1(\Omega) \);
2. \( E_n(u_n) \rightarrow E(u) \) for every sequence \( u_n \rightarrow u \). Consequently, \( E_n \)
   \( \Gamma \)-converges to \( E \);
3. If \( u_n^* \) is an \( \epsilon \)-optimal solution to \( E_n \), i.e. \( E_n(u_n^*) \leq \inf_u E_n(u) + \epsilon \),
   then
   \[ \limsup_n E_n(u_n^*) \leq \inf_u E(u) + \epsilon. \]

- Image segmentation: [Dong, Chien and Shen, 2010]
- Surface reconstruction from point clouds: [Dong and Shen, 2011]
- Standard Dilation: [Dong and Shen, 2011]
- Piecewise Linear WFT
Relations: Wavelet Shrinkage and Nonlinear PDEs

- [Dong, Jiang and Shen, MMS, 2017]

\[ u^k = \tilde{W}^\top S_{\alpha^{k-1}}(Wu^{k-1}), \quad k = 1, 2, \ldots \]

\[ u_t = \sum_{\ell=1}^{L} \frac{\partial^{\alpha_{\ell}}}{\partial x^{\alpha_{\ell}}} \Phi_{\ell}(Du, u), \quad \text{with } Du = \left( \frac{\partial^{\beta_1}}{\partial x^{\beta_1}}, \ldots, \frac{\partial^{\beta_L}}{\partial x^{\beta_L}} \right) \]

- Theoretical justification available for quasilinear parabolic equations.
- Lead to new PDE models such as:

\[ u_{tt} + Cu_t = \sum_{\ell=1}^{L} (-1)^{1+|\beta_{\ell}|} \frac{\partial^{\beta_{\ell}}}{\partial x^{\beta_{\ell}}} \left[ g_{\ell}(u, \frac{\partial^{\beta_1} u}{\partial x^{\beta_1}}, \ldots, \frac{\partial^{\beta_L} u}{\partial x^{\beta_L}}) \frac{\partial^{\beta_{\ell}}}{\partial x^{\beta_{\ell}}} u \right] - \kappa A^\top (Au - f) \]

\[ u^k = (I - \mu A^\top A)W^\top S_{\alpha^{k-1}}(Wu^{k-1}) + \mu A^\top f \]

where

\[ S_{\alpha^{k-1}}(Wu^{k-1}) = \{ S_{\alpha_{\ell}, n}(W_l u^{k-1}) : 0 \leq l \leq \text{Lev} - 1, 1 \leq \ell \leq L \} \]

\[ S_{\alpha_{\ell}, n}(d_{1,n}, d_{2,n}) = d_{\ell,n} \left( 1 - \frac{4\tau}{h^2} g \left( \frac{4(d_{1,n})^2 + 4(d_{2,n})^2}{h^2} \right) \right) \]
Summary

SPARSE APPROXIMATION IN HIGH-DIMENSIONAL DATA CLASSIFICATION

Introduction

What is data science? Extracting knowledge from data to make intelligent observations and decisions.

1. Broad applications
   - Environmental Data
   - Astronomical data
   - Sales data
   - Game data
   - Webpage data

2. Variety
   - Network data
   - Webpage data
   - Text
   - Image
   - Video
   - Audio
Different area has different focus. Some has tight link with another.
- Broader links?
Introduction

Importance of the merge:
- Combining merits
- New insights on classical problems
Introduction

- Typical big data set: with $n \times p$ huge
  \[ X \in K^{n \times p}, \quad K = \mathbb{Z}, \mathbb{R}, \mathbb{C}, \text{ etc.} \]
- Classical v.s. modern

Classical: $n < p$

Modern: $n > p$
Introduction

- Typical big data set: with $n \times p$ huge
  $$X \in K^{n \times p}, \quad K = \mathbb{Z}, \mathbb{R}, \mathbb{C}, \text{ etc.}$$
- Classical v.s. modern

Classical: $n < p$

Modern: $n > p$
Introduction

- **Sparsity is key**
  - What is sparsity for general data sets?
    - Essential information is of much lower dimension than the dimension of the data itself.
  - How do we harvest sparsity?
    - Sparse under certain (nonlinear) transformation.
  - Examples:
    - PCA and its siblings
    - Low rank approximation
    - Wavelet frame transform
    - Dictionary learning
    - Isomaps, LLE, diffusion maps
    - Autoencoder
    - … etc.
Nonlinear Classification

MODERN SCENARIO

Nonlinear Classifier

- When we have enough observations, nonlinear classifier leads to more accurate classification.
PDE Method

- Ginzburg–Landau (GL) functional [Andrea and Flenner, 2011]

\[ E(u) = \frac{\epsilon}{2} \langle u, L_s u \rangle + \frac{1}{4\epsilon} \| u^2 - 1 \|^2_{2,G} + \frac{\mu}{2} \| u|_{\Gamma} - f \|^2_{2,G} \]

where

\[ L_s = I - D^{-1/2}AD^{-1/2} \]

\[ \| f \|_{p,G} := \left( \sum_{k=1}^{K} |f[k]|^p d[k] \right)^{1/p} \]
**PDE Method**

- Splitting $E$ to a sum of convex and concave parts

\[ E(u) = E_1(u) - E_2(u) \]

\[ E_1(u) = \frac{\epsilon}{2} \int |\nabla u(x)|^2 \, dx + \frac{c}{2} \int |u(x)|^2 \, dx, \]

\[ E_2(u) = -\frac{1}{4\epsilon} \int (u(x)^2 - 1)^2 \, dx + \frac{c}{2} \int |u(x)|^2 \, dx - \int \frac{\lambda(x)}{2} (u(x) - u_0(x))^2 \, dx. \]

- Convex splitting scheme

\[ \frac{u^{n+1} - u^n}{dt} = -\frac{\partial E_1}{\partial u}(u^{n+1}) + \frac{\partial E_2}{\partial u}(u^n) \]

- At each iteration, we need to solve a Laplace equation on graph.
- Fast graph Laplacian solver is needed, such as Nystrom’s method.
Wavelet Frame Method

- **Key idea:** Eigenfunctions of Laplace-Beltrami operator (graph Laplacian in discrete setting) are understood as **Fourier basis** on manifolds (graphs in discrete setting) and the associated eigenvalues as **frequency components**.

- Spectrum of Laplace-Beltrami operator on \( \{ \mathcal{M}, g \} \)
  \[
  \Delta u + \lambda u = 0, \quad u|_{S} = 0.
  \]
  Eigenvalues and eigenfunctions: \( 0 < \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \)
  \[
  \langle u_p, u_{p'} \rangle_{L_2(\mathcal{M})} = \int_{\mathcal{M}} u_p(x) u_{p'}(x) dx = \delta_{p, p'}
  \]

- **Fourier transform** \( \hat{f}[p] = \langle f, u_p \rangle_{L_2(\mathcal{M})} \)

- **Plancherel and Parseval's identities**
  \[
  \langle f, g \rangle_{L_2(\mathcal{M})} = \langle \hat{f}, \hat{g} \rangle_{\ell_2(\mathbb{Z}^+)} \text{ for } f, g \in L_2(\mathcal{M})
  \]
  \[
  \|f\|_{L_2(\mathcal{M})}^2 = \|\hat{f}\|_{\ell_2(\mathbb{Z}^+)}^2.
  \]
Wavelet Frame Method

- Asymptotic properties of eigenfunctions and eigenvalues:
  - Weyl’s asymptotic formula (1912): \( \lambda_p \approx p^{\frac{2}{m}} \)
  - Uniform bound (Grieser, 2002): \( \|u_p\|_{L_\infty(M)} \leq C\lambda_p^{\frac{m-1}{4}} \)

- Wavelet system (semi-continuous) on manifold \( M \):
  \[
  X(\Psi) = \{ \psi_{j,n,y}^M \in L_2(M) : 1 \leq j \leq r, n \in \mathbb{Z}, y \in M \},
  \]
  where \( \psi_{j,n,y}^M \in L_2(M) \) is generated by \( \Psi = \{ \psi_j : 1 \leq j \leq r \} \subset L_2(\mathbb{R}) \) as

  \[
  \psi_{j,n,y}^M(x) = \sum_{p=0}^{\infty} \hat{\psi}_j(2^{-n} \lambda_p)u_p^*(y)u_p(x), \quad \text{with } n \in \mathbb{Z}, x \in M, y \in M,
  \]

  Dilation  Translation
  where \( \hat{\psi}_j \) denotes that Fourier transform of \( \psi_j \in L_2(\mathbb{R}) \).

- Question: how to construct \( \psi_j \) so that \( X(\Psi) \) is a tight frame on \( M \)?
Wavelet Frame Method

- Further restriction on $\psi_j$:

Given $\hat{\phi}(2\xi) = \hat{a}(\xi)\hat{\phi}(\xi)$ and $a_j \in \ell_0(\mathbb{Z})$
let $\hat{\psi}_j(2\xi) := \hat{a}_j(\xi)\hat{\phi}(\xi)$, $1 \leq j \leq r$.

- Question: how to construct $a_j$ so that $X(\Psi)$ is a tight frame on $\mathcal{M}$?

- Benefits of such restriction
  - Grants a natural transition from continuum to discrete setting
  - Makes construction of tight frames on manifolds/graphs painless
  - Grants fast decomposition and reconstruction algorithms (Chebyshev polynomial approximation)
Wavelet Frame Method

- Sparsity based semi-supervised learning models

Model L2:  \[
\min_{u \in [0,1]} \| \nu \cdot W u \|_{1,G} + \frac{1}{2} \| u|_G - f \|_{2,G}^2,
\]

Exact Model:  \[
\min_{u \in [0,1]} \| \nu \cdot W u \|_{1,G} \quad \text{s.t.} \quad u|_G = f
\]

Robust Model:  \[
\min_{u \in [0,1]} \| \nu \cdot W u \|_{1,G} + \| u|_G - f \|_{1,G}
\]

- Transform \( W \) is the fast tight wavelet frame transform on graphs [B. Dong, ACHA, 2015].
Wavelet Frame Method

- Classification – Real Datasets
  - MNIST data set (http://yann.lecun.com/exdb/mnist/)
  - Banknote authentication dataset (UCI machine learning repository)

- Results: Model L2 [Dong, ACHA, 2015]

<table>
<thead>
<tr>
<th>Errors (%)</th>
<th>Our Method</th>
<th>Max-Flow</th>
<th>PAL</th>
<th>Binary MBO</th>
<th>GL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>2.76 (8.5 sec.)</td>
<td>1.52</td>
<td>1.56</td>
<td>1.64</td>
<td>1.75</td>
</tr>
<tr>
<td>Banknote</td>
<td>1.64 (2.9 sec.)</td>
<td>1.17</td>
<td>1.71</td>
<td>6.52</td>
<td>3.90</td>
</tr>
</tbody>
</table>

- Max-Flow & PAL: [Merkurjev, Bae, Bertozzi, and Tai, preprint, 2014]
- Binary MBO: [Merkurjev, Kostic, and Bertozzi, 2013]
- GL: [Bertozzi and Flenner, 2012]
Further Studies of Wavelet Frame Transform on Graphs

- High dimensional classification [Dong and Hao, SPIE 2015]
  - V.S. LDA methods: Leukemia (n=72, p=7129) and Lung (n=181, p=12533)

<table>
<thead>
<tr>
<th>Error % (Std %)</th>
<th>NSC</th>
<th>IR</th>
<th>ROAD</th>
<th>RS-ROAD</th>
<th>Exact Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukemia</td>
<td>8.51 (3.0)</td>
<td>4.27 (8.4)</td>
<td>6.35 (6.0)</td>
<td>4.46 (3.1)</td>
<td>5.57 (4.2)</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>10.44 (1.4)</td>
<td>3.47 (7.3)</td>
<td>1.37 (1.1)</td>
<td>0.93 (0.9)</td>
<td>0.59 (0.6)</td>
</tr>
</tbody>
</table>

- Application in super-resolution diffusion MRI [Yap, Dong, Zhang, Lung]
CONCLUDING REMARKS
Conclusions and Future Work

- **Conclusions**
  - Bridging wavelet frame transforms and differential operators
  - New insights, models/algorithms and applications
  - Sparse approximation for general data analysis

- **Future work**
  - Idea of “end-to-end” in classical problems such as imaging
  - Learning PDEs from data
Thanks for Your Attention and Questions?

http://bicmr.pku.edu.cn/~dongbin