Spread and Sparse: Learning Interpretable Transforms for Bandlimited Signals on Digraphs

Rasoul Shafipour

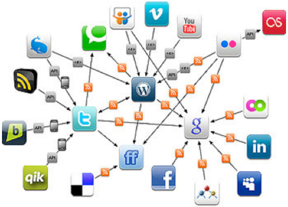
Dept. of Electrical and Computer Engineering
University of Rochester
rshafipo@ece.rochester.edu
http://www.ece.rochester.edu/~rshafipo/

Co-author: Gonzalo Mateos
Acknowledgment: NSF Awards CCF-1750428 and ECCS-1809356

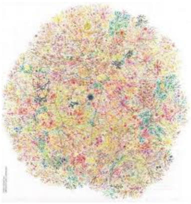
Pacific Grove, CA, October 30, 2018

Network Science analytics

Online social media



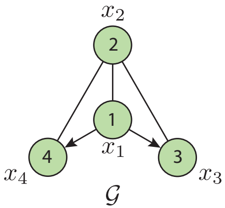
Internet



Clean energy and grid analytics

- Network as graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$: encode pairwise relationships
- Desiderata: Process, analyze and learn from network data [Kolaczyk'09]
- Interest here not in $\mathcal{G}$ itself, but in data associated with nodes in $\mathcal{V}$
  - $\Rightarrow$ The object of study is a graph signal
  - $\Rightarrow$ Ex: Opinion profile, buffer levels, neural activity, epidemic

Graph signal processing and Fourier transform



- Directed graph (digraph) $\mathcal{G}$ with adjacency matrix $\mathbf{A}$
  - $\Rightarrow A_{ij} =$ Edge weight from node i to node j
- Define a signal $\mathbf{x} \in \mathbb{R}^N$ on top of the graph
  - $\Rightarrow x_i =$ Signal value at node i

- Associated with $\mathcal{G}$ is the underlying undirected $\mathcal{G}^u$
  - $\Rightarrow$ Laplacian marix $\mathbf{L} = \mathbf{D} - \mathbf{A}^u$, eigenvectors $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_N]$
- Graph Signal Processing (GSP): exploit structure in $\mathbf{A}$ or $\mathbf{L}$ to process $\mathbf{x}$
- Graph Fourier Transform (GFT): $\tilde{\mathbf{x}} = \mathbf{V}^T\mathbf{x}$ for undirected graphs
  - $\Rightarrow$ Decompose $\mathbf{x}$ into different modes of variation
  - $\Rightarrow$ Inverse (i)GFT $\mathbf{x} = \mathbf{V}\tilde{\mathbf{x}}$, eigenvectors as frequency atoms

Our work in context

- Spectral analysis and filter design [Tremblay et al'17], [Isufi et al'16]

  $\Rightarrow$ GFT as a promising tool in neuroscience [Huang et al'16]

- Noteworthy GFT approaches
  - Jordan decomposition of $\mathbf{A}$ [Sandryhaila-Moura'14], [Deri-Moura'17]
  - Lovaśz extension of the graph cut size [Sardellitti et al'17]
  - Basis selection for spread modes [Shafipour et al'18]
  - Generalized variation operators and inner products [Girault et al'18]
- Dictionary learning (DL) for GSP
  - Parametric dictionaries for graph signals [Thanou et al'14]
  - Dual graph-regularized DL [Yankelevsky-Elad'17]
  - Joint topology- and data-driven prediction [Forero et al'14]
- **Our contribution:** digraph (D)GFT (dictionary) design
  - Orthonormal basis signals (atoms) offer notions of frequency
  - Frequencies are distributed as even as possible in $[0, f_{\max}]$
  - Sparsely represents bandlimited graph signals

Signal variation on digraphs

- Total variation of signal $\mathbf{x}$ with respect to $\mathbf{L}$

$$\text{TV}(\mathbf{x}) = \mathbf{x}^T \mathbf{L} \mathbf{x} = \sum_{i,j=1, j>i}^{N} A_{ij}^{u} (x_i - x_j)^2$$

$\Rightarrow$ Smoothness measure on the graph $\mathcal{G}^u$

- For Laplacian eigenvectors $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_N]$ $\Rightarrow$ $\text{TV}(\mathbf{v}_k) = \lambda_k$

$\Rightarrow$ $0 = \lambda_1 < \cdots \leq \lambda_N$ can be viewed as frequencies

- Directed variation for signals over digraphs ($[x]_+ = \max(0, x)$)

$$\text{DV}(\mathbf{x}) := \sum_{i,j=1}^{N} A_{ij} [x_i - x_j]_+^2$$

$\Rightarrow$ Captures signal variation (flow) along directed edges

$\Rightarrow$ Consistent, since $\text{DV}(\mathbf{x}) \equiv \text{TV}(\mathbf{x})$ for undirected graphs

DGFT with spread frequeny components

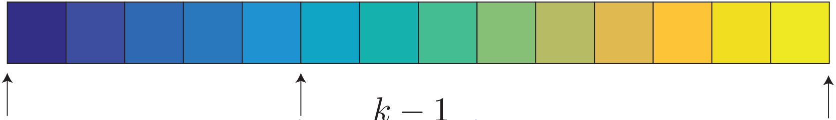
- Find N orthonormal bases capturing low, medium, and high *frequencies*
- Collect the desired bases in a matrix $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_N] \in \mathbb{R}^{N \times N}$

$$\text{DGFT:} \quad \tilde{\mathbf{x}} = \mathbf{U}^T \mathbf{x}$$

$\Rightarrow$ $\mathbf{u}_k$ represents the kth frequency mode with $f_k := \text{DV}(\mathbf{u}_k)$

- Similar to the DFT, seek N *evenly* distributed graph frequencies in $[0, f_{\max}]$

  $\Rightarrow$ $f_{\max}$ is the maximum DV of a unit-norm graph signal on $\mathcal{G}$



$f_{\min}$

$$f_k = \frac{k-1}{N-1} f_{\max}$$

$f_{\max}$

Spread frequencies in two steps

- **First:** Find $f_{\max}$ by solving

$$\mathbf{u}_{\max} = \underset{\|\mathbf{u}\|=1}{\operatorname{argmax}} \ \mathrm{DV}(\mathbf{u}) \quad \text{and} \quad f_{\max} := \mathrm{DV}(\mathbf{u}_{\max}).$$

- Let $\mathbf{v}_N$ be the dominant eigenvector of $\mathbf{L}$
  - $\Rightarrow$ Can 1/2-approximate $f_{\max}$ with $\tilde{\mathbf{u}}_{\max} = \underset{\mathbf{v}\in\{\mathbf{v}_N, -\mathbf{v}_N\}}{\operatorname{argmax}} \ \mathrm{DV}(\mathbf{v})$

- **Second:** Set $\mathbf{u}_1 = \mathbf{u}_{\min} := \frac{1}{\sqrt{N}}\mathbf{1}_N$ and $\mathbf{u}_N = \mathbf{u}_{\max}$ and minimize

$$\delta(\mathbf{U}) := \sum_{i=1}^{N-1} \left[\mathrm{DV}(\mathbf{u}_{i+1}) - \mathrm{DV}(\mathbf{u}_i)\right]^2$$

  - $\Rightarrow$ $\delta(\mathbf{U})$ is the *spectral dispersion function*
  - $\Rightarrow$ Minimized when *free* DV values form an arithmetic sequence

Spectral dispersion and sparsity minimization

- *Sparsify* a set of bandlimited signals $\mathbf{X} \in \mathbb{R}^{N \times P} \rightarrow$ Minimize $||\mathbf{U}^T\mathbf{X}||_1$
- **Problem:** given $\mathcal{G}$ and $\mathbf{X}$, find sparsifying DGFT with spread frequencies

$$\begin{aligned} \min_{\mathbf{U}} \quad & \Psi(\mathbf{U}) := \sum_{i=1}^{N-1} [\mathrm{DV}(\mathbf{u}_{i+1}) - \mathrm{DV}(\mathbf{u}_i)]^2 + \mu||\mathbf{U}^T\mathbf{X}||_1 \\ \text{subject to} \quad & \mathbf{U}^T\mathbf{U} = \mathbf{I} \\ & \mathbf{u}_1 = \mathbf{u}_{\min} \\ & \mathbf{u}_N = \mathbf{u}_{\max} \end{aligned}$$

  - Non-convex, orthogonality-constrained minimization
  - Non-differentiable $\Psi(\mathbf{U})$
  - Feasible since $\mathbf{u}_{\max} \perp \mathbf{u}_{\min}$
- Variable-splitting and a feasible method in the Stiefel manifold:
  (i) Obtain $f_{\max}$ (and $\mathbf{u}_{\max}$) by minimizing $-\mathrm{DV}(\mathbf{u})$ over $\{\mathbf{u} \mid \mathbf{u}^T\mathbf{u} = 1\}$
  (ii) Replace $\mathbf{U}^T\mathbf{X}$ with an auxiliary variable $\mathbf{Y} \in \mathbb{R}^{N \times P}$, enforce $\mathbf{Y} = \mathbf{U}^T\mathbf{X}$
  (iii) Adopt an alternating minimization scheme

Update $\mathbf{U}$: Feasible method in the Stiefel manifold

- For fixed $\mathbf{Y} = \mathbf{Y}_k$, rewrite the problem of finding $\mathbf{U}_{k+1}$ as

$$\underset{\mathbf{U}}{\text{minimize}} \quad \phi(\mathbf{U}) := \delta(\mathbf{U}) + \frac{\lambda}{2}\left(\|\mathbf{u}_1 - \mathbf{u}_{\min}\|^2 + \|\mathbf{u}_N - \mathbf{u}_{\max}\|^2\right) + \frac{\gamma}{2}\|\mathbf{Y}_k - \mathbf{U}^T\mathbf{X}\|_F^2$$
$$\text{subject to} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}_N$$

  - Recall $\delta(\mathbf{U}) := \sum_{i=1}^{N-1}[\mathrm{DV}(\mathbf{u}_{i+1}) - \mathrm{DV}(\mathbf{u}_i)]^2$
  - Choose large enough $\lambda > 0$ to ensure $\mathbf{u}_1 = \mathbf{u}_{\min}$ and $\mathbf{u}_N = \mathbf{u}_{\max}$

- Let $\mathbf{U}_k$ be a feasible point at iteration k and the gradient $\mathbf{G}_k = \nabla\phi(\mathbf{U}_k)$

  $\Rightarrow$ Skew-symmetric matrix $\mathbf{B}_k := \mathbf{G}_k{\mathbf{U}_k}^T - \mathbf{U}_k{\mathbf{G}_k}^T$

- Update rule $\mathbf{U}_{k+1}(\tau) = \left(\mathbf{I} + \frac{\tau}{2}\mathbf{B}_k\right)^{-1}\left(\mathbf{I} - \frac{\tau}{2}\mathbf{B}_k\right)\mathbf{U}_k$

  $\Rightarrow$ Cayley transform preserves orthogonality (i.e., ${\mathbf{U}_{k+1}}^T\mathbf{U}_{k+1} = \mathbf{I}$)

Theorem (Wen-Yin'13) Iterates converge to a stationary point of smooth $\phi(\mathbf{U})$, while generating feasible points at every iteration

Update $\mathbf{Y}$: Soft thresholding

- For fixed $\mathbf{U} = \mathbf{U}_{k+1}$, rewrite the problem of finding $\mathbf{Y}_{k+1}$ as

$$\underset{\mathbf{Y}}{\text{minimize}} \quad \mu\|\mathbf{Y}\|_1 + \frac{\gamma}{2}\|\mathbf{Y} - {\mathbf{U}_{k+1}}^T\mathbf{X}\|_F$$

$\Rightarrow$ Proximal operator that is component-wise separable

- Update $\mathbf{Y}_{k+1}$ in closed form via soft-thresholding operations

$$\mathbf{Y}_{k+1} = \text{sign}({\mathbf{U}_{k+1}}^T\mathbf{X}) \circ \left[|{\mathbf{U}_{k+1}}^T\mathbf{X}| - \mu/\gamma\right]_+$$

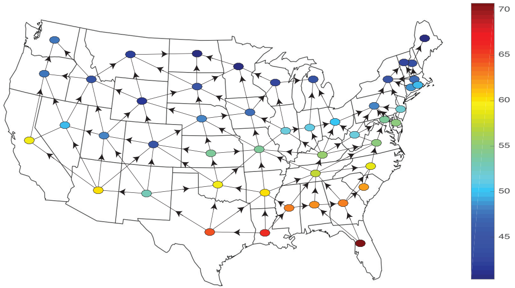
Algorithm

Input: Adjacency matrix $\mathbf{A}$, signals $\mathbf{X} \in \mathbb{R}^{N \times P}$, and λ, μ, γ, ϵ_1, $\epsilon_2 > 0$
Find $\mathbf{u}_{\max}$ by a similar feasible method and set $\mathbf{u}_{\min} = \frac{1}{\sqrt{N}} \mathbf{1}_N$
Initialize $k = 0$, $\mathbf{Y}_0 \in \mathbb{R}^{N \times P}$ at random
repeat
    $\mathbf{U}$**-update:** **Initialize** $t = 0$ and orthonormal $\hat{\mathbf{U}}_0 \in \mathbb{R}^{N \times N}$ at random
        repeat
            Compute gradient $\mathbf{G}_t := \nabla \phi(\hat{\mathbf{U}}_t) \in \mathbb{R}^{N \times N}$
            Form $\mathbf{B}_t = \mathbf{G}_t \hat{\mathbf{U}}_t{}^T - \hat{\mathbf{U}}_t \mathbf{G}_t{}^T$
            Select τ_t satisfying Armijo-Wolfe conditions
            Update $\hat{\mathbf{U}}_{t+1}(\tau_t) = (\mathbf{I}_N + \frac{\tau_t}{2} \mathbf{B}_t)^{-1} (\mathbf{I}_N - \frac{\tau_t}{2} \mathbf{B}_t) \hat{\mathbf{U}}_t$
        until $\|\hat{\mathbf{U}}_t - \hat{\mathbf{U}}_{t-1}\|_F / \|\hat{\mathbf{U}}_{t-1}\|_F \leq \epsilon_1$
        Return $\mathbf{U}_k = \hat{\mathbf{U}}_t$
    $\mathbf{Y}$**-update:** $\mathbf{Y}_{k+1} = \mathrm{sign}(\mathbf{U}_k{}^T \mathbf{X}) \circ (|\mathbf{U}_k{}^T \mathbf{X}| - \mu/\gamma)_+$.
    $k \leftarrow k + 1$.
until $\|\mathbf{U}_k{}^T \mathbf{X} - \mathbf{U}_{k-1}{}^T \mathbf{X}\|_1 / \|\mathbf{U}_{k-1}{}^T \mathbf{X}\|_1 \leq \epsilon_2$
Return $\hat{\mathbf{U}} = \mathbf{U}_k$.

- Overall run-time is $\mathcal{O}(N^3)$ per iteration

Numerical test: US average temperatures

- Graph of the $N = 48$ contiguous United States
  - $\Rightarrow$ Connect two states if they share a border
  - $\Rightarrow$ Set arc directions from lower to higher latitudes



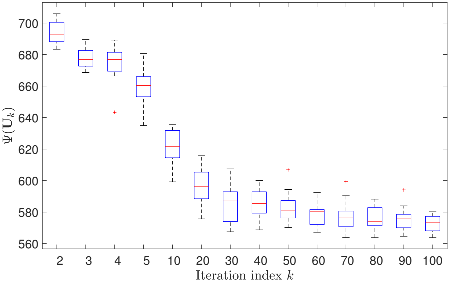
- Test graph signal $\mathbf{x} \rightarrow$ Average annual *temperature* of each state

Numerical test: Convergence behavior

- Average monthly temperature over ~ 60 years for each state

  $\Rightarrow$ Training signals $\mathbf{X} \in \mathbb{R}^{48 \times 12}$

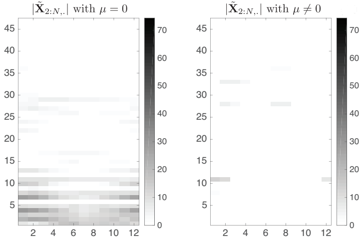
- First, use Monte-Carlo method to study the convergence properties
  - Plot $\boldsymbol{\Psi}(\mathbf{U}) = \delta(\mathbf{U}) + \mu ||\mathbf{U}^T \mathbf{X}||_1$ versus k for 10 different initializations



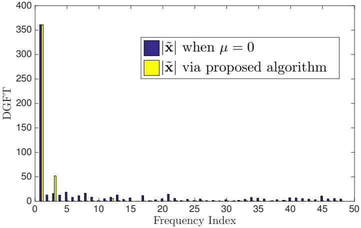
- Convergence is apparent, with limited variability on the solution

Numerical test: Spread and sparse

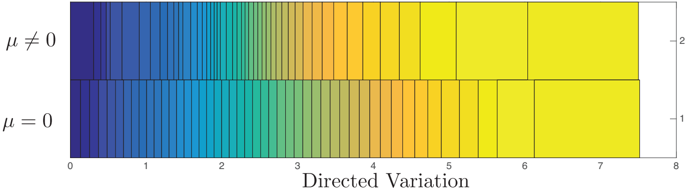
- Heat maps of the trained $\tilde{\mathbf{X}}$



- Spectral representation of test signal



- Distribution of all the frequencies



- **Tradeoff:** spectral dispersion for a sparser representation
  - Still attain well dispersed frequencies

Closing remarks

- Measure of directed variation to capture the notion of frequency on $\mathcal{G}$
- Find an orthonormal set of Fourier basis signals for digraphs
  - Span a maximal frequency range $[0, f_{\max}]$ as evenly as possible
  - Sparsify a training set of bandlimited graph signals
- Adopt alternating scheme via a feasible method and soft-thresholding

  i) Minimize smooth dispersion over the Stiefel manifold

  ii) Encourage sparsity of the representation via soft-thresholding
- Ongoing work and future directions
  - Provide convergence guarantees for the alternating scheme
  - Exploit knowledge on the signals being low, medium, or high-pass
  - Scalable and fast digraph Fourier transform?