Sparse Modeling in Image Processing and Deep Learning

Michael Elad

Computer Science Department
The Technion - Israel Institute of Technology
Haifa 32000, Israel

New Deep Learning Techniques
February 5-9, 2018

The research leading to these results has been received funding from the European union's Seventh Framework Program (FP/2007-2013) ERC grant Agreement ERC-SPARSE- 320649
Another underlying idea that will accompany us

Generative modeling of data sources enables
○ A systematic algorithm development, &
○ A theoretical analysis of their performance
Multi-Layered Convolutional Sparse Modeling
Our Data is Structured

- We are surrounded by various diverse sources of massive information
- Each of these sources have an internal structure, which can be exploited
- This structure, when identified, is the engine behind our ability to process this data

Michael Elad
The Computer-Science Department
The Technion
Models

- A model: a **mathematical** description of the underlying signal of interest, describing our **beliefs** regarding its **structure**

- The following is a partial list of commonly used models for images:
  - Principal-Component-Analysis
  - Gaussian-Mixture
  - Markov Random Field
  - Laplacian Smoothness
  - DCT concentration
  - Wavelet Sparsity
  - Piece-Wise-Smoothness
  - C2-smoothness
  - Besov-Spaces
  - Total-Variation
  - Beltrami-Flow

- Good models should be simple while matching the signals.

- Models are almost always imperfect.
Data Models and Their Use

- Almost any task in data processing requires a model – true for denoising, deblurring, super-resolution, inpainting, compression, anomaly-detection, sampling, recognition, separation, and more

- Sparse and Redundant Representations offer a new and highly effective model – we call it **Sparseland**

- We shall describe this and descendant versions of it that lead all the way to ... **deep-learning**
Multi-Layered Convolutional Sparse Modeling
A New Emerging Model

Sparseland

Signal Processing
- Wavelet Theory
- Multi-Scale Analysis
- Signal Transforms

Machine Learning

Mathematics
- Approximation Theory
- Linear Algebra
- Optimization Theory

Semi-Supervised Learning
Compression
Interpolation
Inference (solving inverse problems)
Source-Separation
Prediction
Denoising
Anomaly detection
Sensor-Fusion
Summarizing
Synthesis
Recognition
Clustering
Identification

Michael Elad
The Computer-Science Department
The Technion
The *Sparseland* Model

- Task: model image patches of size 8×8 pixels
- We assume that a **dictionary** of such image patches is given, containing 256 **atom** images
- The *Sparseland* model assumption: every image patch can be described as a linear combination of **few** atoms
The *Sparseland* Model

Properties of this model:

**Sparsity and Redundancy**

- We start with a 8-by-8 pixels patch and represent it using 256 numbers
  - This is a redundant representation
- However, out of those 256 elements in the representation, only 3 are non-zeros
  - This is a sparse representation
- Bottom line in this case: 64 numbers representing the patch are replaced by 6 (3 for the indices of the non-zeros, and 3 for their entries)
Chemistry of Data

We could refer to the *Sparseland* model as the *chemistry* of information:

- Our dictionary stands for the Periodic Table containing all the elements.
- Our model follows a similar rationale: Every molecule is built of few elements.
Sparseland: A Formal Description

- Every column in $\mathbf{D}$ (dictionary) is a prototype signal (atom).
- The vector $\alpha$ is generated with few non-zeros at arbitrary locations and values.
- This is a generative model that describes how (we believe) signals are created.

A Dictionary $\mathbf{D}$

A sparse vector $\alpha$

$\mathbf{X} = \mathbf{D} \alpha$
Difficulties with Sparseland

- Problem 1: Given a signal, how can we find its atom decomposition?

- A simple example:
  - There are 2000 atoms in the dictionary
  - The signal is known to be built of 15 atoms

\[
\binom{2000}{15} \approx 2.4 \times 10^{37} \text{ possibilities}
\]

- If each of these takes 1 nano-sec to test, will take \(~7.5 \times 10^{20}\) years to finish !!!!!!!

- So, are we stuck?
Atom Decomposition Made Formal

\[ \min_\alpha \|\alpha\|_0 \quad \text{s.t.} \quad x = D\alpha \]

\[ \min_\alpha \|\alpha\|_0 \quad \text{s.t.} \quad \|D\alpha - y\|_2 \leq \varepsilon \]

Approximation Algorithms

- Relaxation methods
  - Basis-Pursuit
- Greedy methods
  - Thresholding/OMP

- \( L_0 \) – counting number of non-zeros in the vector
- This is a projection onto the Sparseland model
- These problems are known to be NP-Hard problem
Pursuit Algorithms

\[
\min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad \|D\alpha - y\|_2 \leq \varepsilon
\]

Approximation Algorithms

Basis Pursuit

Change the \( L_0 \) into \( L_1 \) and then the problem becomes convex and manageable

\[
\min_{\alpha} \|\alpha\|_1 \quad \text{s.t.} \quad \|D\alpha - y\|_2 \leq \varepsilon
\]

Matching Pursuit

Find the support greedily, one element at a time

Thresholding

Multiply \( y \) by \( D^T \) and apply shrinkage:

\[
\hat{\alpha} = \mathcal{P}_\beta\{D^Ty\}
\]
Difficulties with *Sparseland*

- There are various pursuit algorithms
- Here is an example using the Basis Pursuit (L₁):

Surprising fact: Many of these algorithms are often accompanied by **theoretical guarantees** for their success, if the unknown is sparse enough.
The Mutual Coherence

- Compute

  \[ D^T D \]

  Assume normalized columns

- The **Mutual Coherence** \( \mu(D) \) is the largest off-diagonal entry in absolute value

- We will pose all the theoretical results in this talk using this property, due to its simplicity

- You may have heard of other ways to characterize the dictionary (Restricted Isometry Property - RIP, Exact Recovery Condition - ERC, Babel function, Spark, …)
Theorem: Given a noisy signal $y = D\alpha + v$ where $\|v\|_2 \leq \varepsilon$ and $\alpha$ is sufficiently sparse,

$$\|\alpha\|_0 < \frac{1}{4} \left( 1 + \frac{1}{\mu} \right)$$

then Basis-Pursuit:

$$\min_{\alpha} \|\alpha\|_1 \quad \text{s.t.} \quad \|D\alpha - y\|_2 \leq \varepsilon$$

leads to a stable result:

$$\|\hat{\alpha} - \alpha\|_2^2 \leq \frac{4\varepsilon^2}{1 - \mu(4\|\alpha\|_0 - 1)}$$

Comments:
- If $\varepsilon = 0 \rightarrow \hat{\alpha} = \alpha$
- This is a worst-case analysis – better bounds exist
- Similar theorems exist for many other pursuit algorithms

Donoho, Elad & Temlyakov (’06)
Difficulties with *Sparseland*

- Problem 2: Given a family of signals, how do we find the dictionary to represent it well?
  - Solution: Learn! Gather a large set of signals (many thousands), and find the dictionary that sparsifies them.
  - Such algorithms were developed in the past 10 years (e.g., K-SVD), and their performance is surprisingly good.
  - We will not discuss this matter further in this talk due to lack of time.

Michael Elad
The Computer-Science Department
The Technion
Difficulties with *Sparseland*

- Problem 3: Why is this model suitable to describe various sources? e.g., Is it good for images? Audio? Stocks? ...

- General answer: Yes, this model is extremely effective in representing various sources
  - **Theoretical answer:** Clear connection to other models
  - **Empirical answer:** In a large variety of signal and image processing (and later machine learning), this model has been shown to lead to state-of-the-art results

Michael Elad
The Computer-Science Department
The Technion
Difficulties with *Sparseland*?

- Problem 1: Given an image patch, how can we find its atom decomposition?
- Problem 2: Given a family of signals, how do we find the dictionary to represent it well?
- Problem 3: Is this model flexible enough to describe various sources? E.g., Is it good for images? audio? …
A New Massive Open Online Course

Sparse Representations in Signal and Image Processing

Learn the theory, tools and algorithms of sparse representations and their impact on signal and image processing.

Courses in the Professional Certificate Program

- Sparse Representations in Signal and Image Processing: Fundamentals
  Learn about the field of sparse representations by understanding its fundamental theoretical and algorithmic foundations.
  Learn more

- Sparse Representations in Image Processing: From Theory to Practice
  Learn about the deployment of the sparse representation model to signal and image processing.
  Learn more

Instructors

Yaniv Romano
Michael Elad
When handling images, *Sparseland* is typically deployed on small overlapping patches due to the desire to train the model to fit the data better.

The model assumption is: each patch in the image is believed to have a sparse representation w.r.t. a common local dictionary.

What is the corresponding global model? This brings us to ... the Convolutional Sparse Coding (CSC).
Multi-Layered Convolutional Sparse Modeling

Joint work with
Yaniv Romano
Vardan Panyan
Jeremias Sulam
Convolutional Sparse Coding (CSC)

\[ [X] = \sum_{i=1}^{m} d_i * [\Gamma_i] \]

- An image with \( N \) pixels
- The \( i \)-th filter of small size \( n \)
- \( m \) filters convolved with their sparse representations
- \( i \)-th feature-map: An image of the same size as \( X \) holding the sparse representation related to the \( i \)-filter
CSC in Matrix Form

- Here is an alternative global sparsity-based model formulation

\[ x = \sum_{i=1}^{m} c^i \Gamma^i = \begin{bmatrix} c^1 & \cdots & c^m \end{bmatrix} \begin{bmatrix} \Gamma^1 \\ \vdots \\ \Gamma^m \end{bmatrix} = D \Gamma \]

- \( c^i \in \mathbb{R}^{N \times N} \) is a banded and Circulant matrix containing a single atom with all of its shifts

\[ c^i = \begin{bmatrix} \vdots & \ddots & \vdots \\ \end{bmatrix} \]

- \( \Gamma^i \in \mathbb{R}^N \) are the corresponding coefficients ordered as column vectors
The CSC Dictionary

\[
\begin{bmatrix}
C^1 & C^2 & C^3
\end{bmatrix} = D
\]

\[D_L\]

\[D = \begin{bmatrix}
m \\
n
\end{bmatrix}\]
Why CSC?

\[ X = \mathbf{D} \Gamma \]

\[ R_i X = \Omega \gamma_i \]

\[ R_{i+1} X = \Omega \gamma_{i+1} \]

Every patch has a sparse representation w.r.t. to the same local dictionary (\( \Omega \)) just as assumed for images.
Classical Sparse Theory for CSC?

\[ \min_{\Gamma} \| \Gamma \|_0 \quad \text{s.t.} \quad \| Y - D \Gamma \|_2 \leq \varepsilon \]

Theorem: BP is guaranteed to “succeed” .... if \( \| \Gamma \|_0 < \frac{1}{4} \left( 1 + \frac{1}{\mu} \right) \)

- Assuming that \( m = 2 \) and \( n = 64 \) we have that [Welch, ’74]

\[
\mu \geq 0.063
\]

- Success of pursuits is

- Only few (4) non-zeros GLOBALLY are allowed!!! This is a very pessimistic result!

The classic Sparseland Theory does not provide good explanations for the CSC model.
Moving to Local Sparsity: *Stripes*

The main question we aim to address is this:

Can we generalize the vast theory of *Sparseland* to this new notion of local sparsity? For example, could we provide guarantees for success for pursuit algorithms?

\[
\ell_{0,\infty} \text{ Norm: } \| \Gamma \|_{0,\infty}^S = \max_i \| \chi_i \|_0
\]

\[
\min_{\Gamma} \| \Gamma \|_{0,\infty}^S \quad \text{s.t.} \quad \| Y - D \Gamma \|_2 \leq \varepsilon
\]

\[
\| \Gamma \|_{0,\infty}^S \text{ is low } \rightarrow \text{ all } \chi_i \text{ are sparse } \rightarrow \text{ every patch has a sparse representation over } \Omega
\]
Success of the Basis Pursuit

\[ \Gamma_{BP} = \min_{\Gamma} \frac{1}{2} \|Y - D\Gamma\|_2^2 + \lambda \|\Gamma\|_1 \]

**Theorem:** For \( Y = D\Gamma + E \), if \( \lambda = 4 \|E\|_{2,\infty}^p \), if

\[ \|\Gamma\|_{0,\infty}^s < \frac{1}{3} \left( 1 + \frac{1}{\mu(D)} \right) \]

then Basis Pursuit performs very-well:

1. The support of \( \Gamma_{BP} \) is contained in that of \( \Gamma \)
2. \( \|\Gamma_{BP} - \Gamma\|_{\infty} \leq 7.5 \|E\|_{2,\infty}^p \)
3. Every entry greater than \( 7.5 \|E\|_{2,\infty}^p \) is found
4. \( \Gamma_{BP} \) is unique

Local noise (per patch)

Papyan, Sulam & Elad (‘17)
Multi-Layered Convolutional Sparse Modeling
Quick Recall: The Forward Pass

\[ f(Y) = \text{ReLU}(b_2 + W_2^T \text{ReLU}(b_1 + W_1^T Y)) \]

\[ Z_2 \in \mathbb{R}^{N \times m_2} \quad b_2 \in \mathbb{R}^{N \times m_2} \quad W_2^T \in \mathbb{R}^{N \times m_2 \times N \times m_1} \]

\[ b_1 \in \mathbb{R}^{N \times m_1} \quad W_1^T \in \mathbb{R}^{N \times m_1 \times N} \]

\[ Y \in \mathbb{R}^N \]
Convolutional sparsity (CSC) assumes an inherent structure is present in natural signals.

We propose to impose the same structure on the representations themselves.

\[
\begin{align*}
X & \in \mathbb{R}^N \\
D_1 & \in \mathbb{R}^{N \times N m_1} \\
\Gamma_1 & \in \mathbb{R}^{N m_1} \\
D_2 & \in \mathbb{R}^{N m_1 \times N m_2} \\
\Gamma_2 & \in \mathbb{R}^{N m_2}
\end{align*}
\]
Intuition: From Atoms to Molecules

- We can chain all the dictionaries into one effective dictionary:
  \[
  D_{\text{eff}} = D_1 D_2 D_3 \cdots D_K \rightarrow x = D_{\text{eff}} \Gamma_K
  \]

- This is a special \textit{Sparseland} (indeed, a CSC) model

- However:
  - A key property in this model: sparsity of the \textit{intermediate representations}
  - The effective atoms: \textit{atoms}
A Small Taste: Model Training (MNIST)

MNIST Dictionary:
- \(D_1\): 32 filters of size \(7 \times 7\), stride of 2 (dense)
- \(D_2\): 128 filters of size \(5 \times 5 \times 32\), stride of 1 - 99.09% sparse
- \(D_3\): 1024 filters of size \(7 \times 7 \times 128\), 99.89% sparse

\(D_1, D_2, D_1D_2, D_1D_2D_3\)
A Small Taste: Model Training (CiFAR)

\[ D_1 (5 \times 5 \times 3) \quad D_1 D_2 (13 \times 13) \quad D_1 D_2 D_3 (32 \times 32) \]
ML-CSC: Pursuit

- **Deep-Coding Problem (DCP)** (dictionaries are known):

  \[
  \begin{align*}
  \mathbf{X} &= \mathbf{D}_1 \mathbf{\Gamma}_1 & \|\mathbf{\Gamma}_1\|_{0,\infty}^s &\leq \lambda_1 \\
  \mathbf{\Gamma}_1 &= \mathbf{D}_2 \mathbf{\Gamma}_2 & \|\mathbf{\Gamma}_2\|_{0,\infty}^s &\leq \lambda_2 \\
  \vdots & & \vdots \\
  \mathbf{\Gamma}_{K-1} &= \mathbf{D}_K \mathbf{\Gamma}_K & \|\mathbf{\Gamma}_K\|_{0,\infty}^s &\leq \lambda_K
  \end{align*}
  \]

- Or, more realistically for noisy signals,

  \[
  \text{Find } \{\mathbf{\Gamma}_j\}_{j=1}^K \text{ s.t. } \begin{align*}
  \|\mathbf{Y} - \mathbf{D}_1 \mathbf{\Gamma}_1\|_2 &\leq \varepsilon & \|\mathbf{\Gamma}_1\|_{0,\infty}^s &\leq \lambda_1 \\
  \mathbf{\Gamma}_1 &= \mathbf{D}_2 \mathbf{\Gamma}_2 & \|\mathbf{\Gamma}_2\|_{0,\infty}^s &\leq \lambda_2 \\
  \vdots & & \vdots \\
  \mathbf{\Gamma}_{K-1} &= \mathbf{D}_K \mathbf{\Gamma}_K & \|\mathbf{\Gamma}_K\|_{0,\infty}^s &\leq \lambda_K
  \end{align*}
  \]
A Small Taste: Pursuit

\[ x = D_1 \Gamma_1 \]
\[ x = D_1 D_2 \Gamma_2 \]
\[ x = D_1 D_2 D_3 \Gamma_3 \]

\( \Gamma_1 \)
94.51% sparse (213 nnz)

\( \Gamma_2 \)
99.52% sparse (30 nnz)

\( \Gamma_3 \)
99.51% sparse (5 nnz)
The simplest pursuit algorithm (single-layer case) is the THR algorithm, which operates on a given input signal $Y$ by:

$$Y = D \Gamma + E$$

and $\Gamma$ is sparse

$$\hat{\Gamma} = P_\beta(D^T Y)$$
Consider this for Solving the DCP

- Layered thresholding (LT):
  - Estimate $\Gamma_1$ via the THR algorithm
    \[ \hat{\Gamma}_2 = \mathcal{P}_{\beta_2} \left( D_2^T \mathcal{P}_{\beta_1} (D_1^T Y) \right) \]
  - Estimate $\Gamma_2$ via the THR algorithm

- Now let's take a look at how Conv. Neural Network operates:
  \[ f(Y) = \text{ReLU}(b_2 + W_2^T \text{ReLU}(b_1 + W_1^T Y)) \]

The layered (soft nonnegative) thresholding and the CNN forward pass algorithm are the very same thing !!!
Theoretical Path

\[
\mathbf{M} = \mathbf{D}_1 \Gamma_1 \\
\Gamma_1 = \mathbf{D}_2 \Gamma_2 \\
\vdots \\
\Gamma_{K-1} = \mathbf{D}_K \Gamma_K \\
\Gamma_i \text{ is } L_{0,\infty} \text{ sparse}
\]

Armed with this view of a generative source model, we may ask new and daring theoretical questions

\[\mathcal{A} (\text{DCP}_\lambda^\varepsilon)\]

Layered THR (Forward Pass)

\[\left\{ \hat{\Gamma}_i \right\}_{i=1}^K\]

Maybe other?
Success of the Layered-THR

Theorem: If $\|\Gamma_i\|_{0,\infty}^S < \frac{1}{2} \left( 1 + \frac{1}{\mu(D_i)} \cdot \frac{|\Gamma_i^{\min}|}{|\Gamma_i^{\max}|} \right) - \frac{1}{\mu(D_i)} \cdot \frac{\varepsilon^i_{L}^{-1}}{|\Gamma_i^{\max}|}$

then the Layered Hard THR (with the proper thresholds) finds the correct supports and $\|\Gamma_i^{ LT} - \Gamma_i\|^p_{2,\infty} \leq \varepsilon^i_L$, where we have defined $\varepsilon^0_L = \|E\|^p_{2,\infty}$ and

$$\varepsilon^i_L = \sqrt{\|\Gamma_i\|^p_{0,\infty} \cdot (\varepsilon^i_{L}^{-1} + \mu(D_i)(\|\Gamma_i\|^S_{0,\infty} - 1)|\Gamma_i^{\max}|)}$$

The stability of the forward pass is guaranteed if the underlying representations are locally sparse and the noise is locally bounded.

Problems:
1. Contrast
2. Error growth
3. Error even if no noise

Papyan, Romano & Elad (’17)
Layered Basis Pursuit (BP)

- We chose the Thresholding algorithm due to its simplicity, but we do know that there are better pursuit methods – how about using them?
- Let's use the Basis Pursuit instead ...

\[
\Gamma_1^{LBP} = \min_{\Gamma_1} \frac{1}{2} \|Y - D_1 \Gamma_1\|_2^2 + \lambda_1 \|\Gamma_1\|_1
\]

\[
\Gamma_2^{LBP} = \min_{\Gamma_2} \frac{1}{2} \|\Gamma_1^{LBP} - D_2 \Gamma_2\|_2^2 + \lambda_2 \|\Gamma_2\|_1
\]

\[
\begin{align*}
\text{(DCP}_\lambda^\mathcal{E}) & : \text{Find } \{\Gamma_j\}_{j=1}^K \text{ s.t.} \\
\|Y - D_1 \Gamma_1\|_2 & \leq \mathcal{E} \quad \|\Gamma_1\|_{0,\infty}^s \leq \lambda_1 \\
\Gamma_1 & = D_2 \Gamma_2 \quad \|\Gamma_2\|_{0,\infty}^s \leq \lambda_2 \\
\vdots & \quad \vdots \\
\Gamma_{K-1} & = D_K \Gamma_K \quad \|\Gamma_K\|_{0,\infty}^s \leq \lambda_K
\end{align*}
\]

Deconvolutional networks

[Zeiler, Krishnan, Taylor & Fergus ‘10]
Success of the Layered BP

Theorem: Assuming that \( \|\Gamma_i\|_{0,\infty}^s < \frac{1}{3} \left( 1 + \frac{1}{\mu(D_i)} \right) \)
then the Layered Basis Pursuit performs very well:

1. The support of \( \Gamma_i^{LBP} \) is contained in that of \( \Gamma_i \)
2. The error is bounded: \( \|\Gamma_i^{LBP} - \Gamma_i\|_{2,\infty}^p \leq \epsilon_L^i \), where
   \[ \epsilon_L^i = 7.5^i \|E\|_{2,\infty}^p \prod_{j=1}^{i} \sqrt{\|\Gamma_j\|_{0,\infty}^p} \]
3. Every entry in \( \Gamma_i \) greater than \( \epsilon_L^i/\sqrt{\|\Gamma_i\|_{0,\infty}^p} \) will be found

Problems:
1. Contrast
2. Error growth
3. Error even if no noise

Papyan, Romano & Elad ('17)
Layered Iterative Thresholding

Layered BP:
\[ \Gamma_j^{LBP} = \min_{\Gamma_j} \frac{1}{2} \| \Gamma_{j-1}^{LBP} - D_j \Gamma_j \|^2 + \xi_j \| \Gamma_j \|_1 \]

Layered Iterative Soft-Thresholding:
\[ \Gamma_j^t = S_{\xi_j/c_j} \left( \Gamma_j^{t-1} + D_j^T (\hat{\Gamma}_{j-1} - D_j \Gamma_j^{t-1}) \right) \]

Note that our suggestion implies that groups of layers share the same dictionaries.

Can be seen as a very deep recurrent neural network

[Gregor & LeCun ‘10]
What About Learning?

- All these models rely on proper Dictionary Learning Algorithms to fulfil their mission:
  - Sparseland: We have unsupervised and supervised such algorithms, and a beginning of theory to explain how these work.
  - CSC: We have few and only unsupervised methods, and even these are not fully stable/clear.
  - ML-CSC: One algorithm has been proposed (unsupervised) – see ArxiV.
Where are the Labels?

\[ X = D_1 \Gamma_1 \]
\[ \Gamma_1 = D_2 \Gamma_2 \]
\[ \vdots \]
\[ \Gamma_{K-1} = D_K \Gamma_K \]

\( \Gamma_i \) is \( L_{0,\infty} \) sparse

**Answer 1:**
- We do not need labels because everything we show refers to the unsupervised case, in which we operate on signals, not necessarily in the context of recognition.

**Answer 2:**
- In fact, this model labels become everything we show, so to speak, of the corresponding label by:

\[ L(X) = \text{sign}(c + \sum_{j=1}^{K} w_j \Gamma_{jj}) \]

- This assumes that knowing the representations (or maybe their supports?) suffice for identifying the label.
- Thus, a successful pursuit algorithm can lead to an accurate recognition if the network is augmented by a FC classification layer.

We presented the ML-CSC as a machine that produces signals \( X \).
Time to Conclude
This Talk

**Take Home Message 1:**
Generative modeling of data sources enables algorithm development along with theoretically analyzing algorithms' performance.

A novel interpretation and theoretical understanding of CNN

**Take Home Message 2:**
The Multi-Layer Convolutional Sparse Coding model could be a new platform for understanding and developing deep-learning solutions.

We presented a theoretical study of the CSC model and showed that it aligns with theoretical CNN tasks while getting global optimality.

**Sparseland**

The desire to model data

**Novel View of Convolutional Sparse Coding**

**Multi-Layer Convolutional Sparse Coding**

We spoke about the importance of models in signal/image processing and described Sparseland in details.

We presented a theoretical study of the CSC model and how to operate locally while getting global optimality.

\[
X = D_1 \Gamma_1 \\
\Gamma_1 = D_2 \Gamma_2 \\
\vdots \\
\Gamma_{k-1} = D_k \Gamma_k \\
\Gamma_k \text{ is } \ell_{0,\infty} \text{ sparse}
\]
More on these (including these slides and the relevant papers) can be found in http://www.cs.technion.ac.il/~elad