Sparse Modeling
in
Image Processing and Deep Learning

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

Monday July 9th 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
This Lecture is About ...

A Proposed Theory for Deep-Learning (DL)

Explanation:

- DL has been extremely successful in solving a variety of learning problems
- DL is an empirical field, with numerous tricks and know-how, but almost no theoretical foundations
- A theory for DL has become the holy-grail of current research in Machine-Learning and related fields
Who Needs Theory?

We All Do !!

... because ... A theory

- ... could bring the next rounds of ideas to this field, breaking existing barriers and opening new opportunities
- ... could map clearly the limitations of existing DL solutions, and point to key features that control their performance
- ... could remove the feeling with many of us that DL is a “dark magic”, turning it into a solid scientific discipline

Ali Rahimi: NIPS 2017 Test-of-Time Award

“Machine learning has become alchemy”

Yan LeCun

Understanding is a good thing... but another goal is inventing methods. In the history of science and technology, engineering preceded theoretical understanding:

- Lens & telescope → Optics
- Steam engine → Thermodynamics
- Airplane → Aerodynamics
- Radio & Comm. → Info. Theory
- Computer → Computer Science

Michael Elad
The Computer-Science Department
The Technion
A Theory for DL?

Stephane Mallat (ENS) & Joan Bruna (NYU): Proposed the scattering transform (wavelet-based) and emphasized the treatment of invariances in the input data.

Richard Baraniuk & Ankit Patel (RICE): Offered a generative probabilistic model for the data, showing how classic architectures and learning algorithms relate to it.

Raja Giryes (TAU): Studied the architecture of DNN in the context of their ability to give distance-preserving embedding of signals.

Gitta Kutyniok (TU) & Helmut Bolcskei (ETH): Studied the ability of DNN architectures to approximate families of functions.

Rene Vidal (JHU): Explained the ability to optimize the typical non-convex objective and yet get to a global minima.

Naftali Tishby (HUJI): Introduced the Information Bottleneck (IB) concept and demonstrated its relevance to deep learning.

Stefano Soatto’s team (UCLA): Analyzed the Stochastic Gradient Descent (SGD) algorithm, connecting it to the IB objective.
So, is there a Theory for DL?

The answer is tricky:

There are already various such attempts, and some of them are truly impressive

... but ...

none of them is complete
Interesting Observations

○ Theory origins: Signal Proc., Control Theory, Info. Theory, Harmonic Analysis, Sparse Represent., Quantum Physics, PDE, Machine learning ...

Ron Kimmel: “DL is a dark monster covered with mirrors. Everyone sees his reflection in it ...”

David Donoho: “... these mirrors are taken from Cinderella's story, telling each that he is the most beautiful”

○ Today’s talk is on our proposed theory:

... and yes, our theory is the best 😊
This Lecture: More Specifically

Sparseland
Sparse Representation Theory

CSC
Convolutional Sparse Coding

ML-CSC
Multi-Layered Convolutional Sparse Coding

Sparsity-Inspired Models → Deep-Learning

Another underlying idea that accompanies us

Generative modeling of data sources enables

- A systematic algorithm development, &
- A **theoretical analysis** of their performance

**Disclaimer:** Being a lecture on the theory of DL, this lecture is ... theoretical ... and mathematically oriented

Michael Elad
The Computer-Science Department
The Technion
Our eventual goal in today’s talk is to present the ...

**Multi-Layered Convolutional Sparse Modeling**

So, let's use this as our running title, parse it into words, and explain each of them.
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 the ability to process data
- How to identify structure?

Michael Elad
The Computer-Science Department
The Technion
Effective removal of noise (and many other tasks) relies on a proper modeling of the signal.

**Fact 1:** This signal contains AWGN $\mathbb{N}(0,1)$

**Fact 2:** The clean signal is believed to be PWC.
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.
- Simplicity ↔ Reliability
- Models are almost always imperfect.
What this Talk is all About?

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 \textit{Sparseland}

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

**Sparseland**

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

**Mathematics**
- Approximation Theory
- Linear Algebra
- Optimization Theory

**Machine Learning**

**Applications**
- Denoising
- Interpolation
- Prediction
- Source-Separation
- Segmentation
- Classification
- Recognition
- Clustering
- Identification
- Sensor-Fusion
- Summarizing
- Anomaly Detection
- Synthesis

- Semi-Supervised Learning
- Compression

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

- Task: model image patches of size $8 \times 8$ 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

---

Michael Elad
The Computer-Science Department
The Technion
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)
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 \( \mathbf{\alpha} \) is generated with few non-zeros at arbitrary locations and values
- This is a generative model that describes how (we believe) signals are created
Difficulties with \textit{Sparseland}

- Problem 1: Given a signal, how can we find its \textit{atom decomposition}?
- A simple example:
  - There are 2000 atoms in the dictionary
  - The signal is known to be built of 15 atoms
  \[
  \binom{2000}{15} \approx 2.4e + 37 \text{ possibilities}
  \]
  - If each of these takes 1 nano-sec to test, will take \(\sim 7.5e20\) 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:

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

- There are various pursuit algorithms
- Here is an example using the Basis Pursuit (L$_1$):
  
  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, ...)

---

Michael Elad
The Computer-Science Department
The Technion
Basis-Pursuit Success

Theorem: Given a noisy signal \( y = D\alpha + v \) where \( \|v\|_2 \leq \varepsilon \) and \( \alpha \) is sufficiently sparse, 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.
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 \textit{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? ...

\begin{itemize}
  \item \textbf{ALL ANSWERED POSITIVELY AND CONSTRUCTIVELY}
\end{itemize}
This Field has been rapidly GROWING ... 

- **Sparseland** has a great success in signal & image processing and machine learning tasks
- In the past 8-9 years, many books were published on this and closely related fields
A New Massive Open Online Course

Sparse Representations in Signal and Image Processing

Michael Elad
The Computer-Science Department
The Technion
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 Paryan  Jeremias Sulam
Convolutional Sparse Coding (CSC)

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

- \( m \) filters convolved with their sparse representations
- \( d_i \) is the \( m \)-th feature-map
- \( [\Gamma_i] \) is the sparse representation related to the \( i \)-th filter

An image with \( N \) pixels

- The \( i \)-th filter of small size \( n \)

i-th feature-map: An image of the same size as \( \mathbf{X} \) holding the sparse representation related to the \( i \)-filter

This model emerged in 2005-2010, developed and advocated by Yan LeCun and others. It serves as the foundation of Convolutional Neural Networks.
CSC in Matrix Form

- Here is an alternative global sparsity-based model formulation

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

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

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

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

\[
D_L
\]

\[
D =
\]
Why CSC?

\[ \mathbf{R}_i \mathbf{R}_{i+1} \mathbf{X} \]

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

\[ \mathbf{R}_i \mathbf{X} = \mathbf{\Omega} \mathbf{\gamma}_i \]

\[ \mathbf{R}_{i+1} \mathbf{X} = \mathbf{\Omega} \mathbf{\gamma}_{i+1} \]

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

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

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 guaranteed as long as \( \mu \)

- 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.
The main question we aim to address is this:

Can we generalize the vast theory of \textit{Sparseland} to this new notion of local sparsity? For example, could we provide guarantees for success for pursuit algorithms?
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 \( \lambda = 4\|E\|^p_{2,\infty} \), 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. \( \|\Gamma_{BP} - \Gamma\|_\infty \leq 7.5\|E\|^p_{2,\infty} \)
3. Every entry greater than \( 7.5\|E\|^p_{2,\infty} \) is found
4. \( \Gamma_{BP} \) is unique

This is a much better result – it allows few non-zeros locally in each stripe, implying a permitted \( O(N) \) non-zeros globally

Papyan, Sulam & Elad (‘17)
Multi-Layered Convolutional Sparse Modeling
Convolutional sparsity (CSC) assumes an inherent structure is present in natural signals.

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

\[ X \in \mathbb{R}^N \quad D_1 \in \mathbb{R}^{N \times Nm_1} \quad \Gamma_1 \in \mathbb{R}^{Nm_1} \]

\[ D_2 \in \mathbb{R}^{Nm_1 \times Nm_2} \quad \Gamma_2 \in \mathbb{R}^{Nm_2} \]

Multi-Layer CSC (ML-CSC)
Intuition: From Atoms to Molecules

- We can chain the all the dictionaries into one effective dictionary
  \[ \mathbf{D}_{\text{eff}} = \mathbf{D}_1 \mathbf{D}_2 \mathbf{D}_3 \cdots \mathbf{D}_K \rightarrow \mathbf{x} = \mathbf{D}_{\text{eff}} \Gamma_K \]

- This is a special *Sparseland* (indeed, a CSC) model

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

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

- **Deep-Coding Problem (\(\text{DCP}_\lambda\))** (dictionaries are known):

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

- Or, more realistically for noisy signals,

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

\[ x = D_1 \Gamma_1 \]
\[ x = D_1 D_2 \Gamma_2 \]
\[ x = D_1 D_2 D_3 \Gamma_3 \]
ML-CSC: The Simplest Pursuit

The simplest pursuit algorithm (single-layer case) is the THR algorithm, which operates on a given input signal $\mathbf{Y}$ by:

$$\mathbf{Y} = \mathbf{D} \mathbf{\Gamma} + \mathbf{E}$$

and $\mathbf{\Gamma}$ is sparse

$$\hat{\mathbf{\Gamma}} = \mathcal{P}_\beta (\mathbf{D}^T \mathbf{Y})$$
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!!!
Armed with this view of a generative source model, we may ask new and daring theoretical questions
Success of the Layered-THR

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

then the **Layered Hard THR** (with the proper thresholds) **finds the correct supports** and \( \| \Gamma_i^{LT} - \Gamma_i \|_{2,\infty}^p \leq \epsilon_i \), where we have defined \( \epsilon_0^i = \| E \|_{2,\infty}^p \) and

\[
\epsilon_i = \sqrt{\| \Gamma_i \|_{0,\infty}^p \cdot (\epsilon_{L,i}^{i-1} + \mu(D_i) (\| \Gamma_i \|_{0,\infty}^S - 1) |\Gamma_i^{\max}|)}
\]

Papyan, Romano & Elad (’17)

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
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 ...

\[
\begin{align*}
\Gamma_1^{\text{LBP}} &= \min_{\Gamma_1} \frac{1}{2} \|Y - D_1 \Gamma_1\|_2^2 + \lambda_1 \|\Gamma_1\|_1 \\
\Gamma_2^{\text{LBP}} &= \min_{\Gamma_2} \frac{1}{2} \|\Gamma_1^{\text{LBP}} - D_2 \Gamma_2\|_2^2 + \lambda_2 \|\Gamma_2\|_1 \\
&\vdots
\end{align*}
\]

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

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:
   \[
   \| \Gamma_i^{LBP} - \Gamma_i \|_2,\infty^p \leq \varepsilon_i^L,
   \]
   where
   \[
   \varepsilon_L = 7.5^i \| E \|_{2,\infty}^p \Pi_j=1 \sqrt{\| \Gamma_j \|_{0,\infty}^p}
   \]
3. Every entry in \( \Gamma_i \) greater than \( \varepsilon_L^i / \sqrt{\| \Gamma_i \|_{0,\infty}^p} \) will be found

---

**Problems:**

1. **Contrast**
2. **Error growth**
3. **Error even if no noise**

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

Layered BP:
\[
\Gamma_j^{LBP} = \min_{\Gamma_j} \frac{1}{2} \left\| \Gamma_{j-1}^{LBP} - D_j \Gamma_j \right\|_2^2 + \xi_j \left\| \Gamma_j \right\|_1
\]

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

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]
Where are the Labels?

Answer 2:

- We fact, this model labels can be augmented by a synthesis 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.

- See more on this in our recent submission to NIPS 2018 (Available on ArXiv)

We presented the ML-CSC as a machine that produces signals \( X \).
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**: Two algorithms were proposed – see ArxiV (unsupervised) and submission to NIPS 2018 (supervised).
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 a novel multi-layer extension, showing tight connections to CNNs.
More on these (including these slides and the relevant papers) can be found in http://www.cs.technion.ac.il/~elad