Sparse Modeling of Data and its Relation to Deep Learning

Michael Elad

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

October 31st – November 1st, 2019
This Lecture Presents ...

A Theoretical Explanation of Deep-Learning (DL) Architectures based on Sparse Data Modeling

Context:

- Theoretical explanation for DL has become the holy-grail of data-sciences – this event is all about this
- There is a growing volume of such contributions
- Our work presents another chapter in this “growing book” of knowledge
- The overall dream: A coherent and complete theory for deep-learning
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

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

Ali Rahimi: NIPS 2017 Test-of-Time Award

“Machine learning has become alchemy”

Yan LeCun

Machine learning has become alchemy

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
A Theory for DL?

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.

Stefano Soatto’s team (UCLA): Analyzed the Stochastic Gradient Descent (SGD) algorithm, connecting it to the Information Bottleneck objective.

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

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

Stephane Mallat (ENS) & Joan Bruna (NYU): Proposed the scattering transform and emphasized the treatment of invariances in the input data.
Where Are We in this Map?

What About You?

- Eran Malach (SGD, generalization, deep generative model)
- Haim Sompolinsky (data manifold, geometry)
- Sanjeev Arora (Loss func. connectivity, optimization & generalization)
- Tomaso Poggio (approximation, optimization, generalization)
- Jeffery Pennington (rotation-based NN, batch-normalization)
- Surya Ganguli (point networks, dynamics of learning)
- Naftali Tishbi (information bottleneck)
- Yasaman Bahri (training & generalization)

Our work?

We start by **modeling the data** and show how it reflects on the choice of the **architectures** and on their expected performance.
Interesting Observation

- Languages used: Signal Processing, Control Theory, Information Theory, Harmonic Analysis, Sparse Representation, Quantum Physics, PDE, Machine learning, Theoretical CS, Neuroscience, ...

  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 theoretical view:

  Yaniv Romano  Vardan Papyan  Jeremias Sulam  Aviad Aberdam

... and our theory is the best 😊
Another underlying idea that accompanies us

Generative modeling of data sources enables
- A systematic algorithm development, &
- A theoretical analysis of their performance
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?

Using models

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

- Good models should be simple while matching the signals

  - Simplicity
  - Reliability

- Models are almost always imperfect
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

Segmentation
Classification

Recognition
Clustering
Identification

Sensor-Fusion
Summarizing
Synthesis

Anomaly detection

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
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 $\alpha$ 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 *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

\[
\binom{2000}{15} \approx 2.4e + 37 \text{ possibilities}
\]

- If each of these takes 1 nano-sec to test, will take \(~7.5e20\) 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

\begin{align*}
\min_{\alpha} & \|\alpha\|_0 \\
\text{s.t.} & \quad \|D\alpha - y\|_2 \leq \varepsilon
\end{align*}

- 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

Relaxation methods

- Basis-Pursuit

Greedy methods

- Thresholding
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

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

Thresholding

Multiply $y$ by $D^T$ and apply shrinkage:

$$\hat{\alpha} = P_\beta\{D^Ty\}$$
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**
  
  \[ \begin{pmatrix} D^T \end{pmatrix} \begin{pmatrix} D \\ \end{pmatrix} = \begin{pmatrix} D^T D \end{pmatrix} \]
  
  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, ...)
**Basis-Pursuit Success**

**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? …
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).
Convolutional Sparse Coding (CSC)

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

- \( m \) 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
- An image with \( N \) pixels
- The \( i \)-th filter of small size \( n \)

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:

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

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

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

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

\[
\mathbf{D}_L
\]

\[
\mathbf{D} =
\]

\[
\mathbf{D}
\]
Why CSC?

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

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

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

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

\[ \mathbf{R}_i \mathbf{X} \mathbf{Y}_i = (2(2^n - 1) \eta \mu) \]

\[ \mathbf{V}_{i+1} \]
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 \( \|\Gamma\|_0 < \frac{1}{4} (1 + \frac{1}{\mu}) \)

- Assuming that \( m = 2 \) and \( n = 64 \) we have that [Welch, ’74]
  \[ \mu \geq 0.063 \]

- Success of pursuits is guaranteed as long as
  \[ \|\Gamma\|_0 < \frac{1}{4} (1 + \frac{1}{\mu}) \leq 0.2 \]

- 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 *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 \( \lambda = 4 \| E \|_{2,\infty}^p \), if

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

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

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


From CSC to Multi-Layered CSC

Convolutional sparsity (CSC) assumes an inherent structure is present in natural signals.

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

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

- The atoms of $D_1 D_2$ are combinations of atoms from $D_1$ - these are now molecules.
- Thus, this model offers different levels of abstraction in describing $X$.

$X = \sum$ atoms, molecules, cells, tissue, body-parts ...
Intuition: From Atoms to Molecules

\[
\mathbf{D}_{\text{eff}} = \mathbf{D}_1 \mathbf{D}_2 \mathbf{D}_3 \cdots \mathbf{D}_K \quad \rightarrow \quad \mathbf{x} = \mathbf{D}_{\text{eff}} \mathbf{\Gamma}_K
\]

- This is a special *Sparseland* (indeed, a CSC) model
- However: A key property in our model: the intermediate representations are required to be sparse as well
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 99.09% sparse
• \( D_3 \): 1024 filters of size \( 7 \times 7 \times 128 \), with 99.89% sparse

\( D_1 D_2 \) (15×15)

\( D_1 D_2 D_3 \) (28×28)
ML-CSC: Pursuit

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

\[
\begin{align*}
\mathbf{X} &= \mathbf{D}_1 \Gamma_1 & \|\Gamma_1\|_{0,\infty}^s &\leq \lambda_1 \\
\Gamma_1 &= \mathbf{D}_2 \Gamma_2 & \|\Gamma_2\|_{0,\infty}^s &\leq \lambda_2 \\
&\vdots & \vdots \\
\Gamma_{K-1} &= \mathbf{D}_K \Gamma_K & \|\Gamma_K\|_{0,\infty}^s &\leq \lambda_K
\end{align*}
\]

- Or, more realistically for noisy signals,

\[
\begin{align*}
\text{Find} & \quad \{\Gamma_j\}_{j=1}^K \quad s.t. \quad \begin{cases}
\|\mathbf{Y} - \mathbf{D}_1 \Gamma_1\|_2 &\leq \varepsilon \\
\Gamma_1 &= \mathbf{D}_2 \Gamma_2 & \|\Gamma_2\|_{0,\infty}^s &\leq \lambda_2 \\
&\vdots & \vdots \\
\Gamma_{K-1} &= \mathbf{D}_K \Gamma_K & \|\Gamma_K\|_{0,\infty}^s &\leq \lambda_K
\end{cases}
\end{align*}
\]
A Small Taste: Pursuit

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

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

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

The simplest pursuit algorithm (single-layer case) is the THR algorithm, which operates on a given input signal $Y$ 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 !!!
Armed with this view of a generative source model, we may ask new and daring theoretical questions.
Success of the Layered-THR

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

then the **Layered Hard THR** (with the proper thresholds) **finds the correct supports** and $\|\Gamma_i^{LT} - \Gamma_i\|_{2,\infty}^p \leq \varepsilon_i^L$, where we have defined $\varepsilon_0^L = \|E\|_{2,\infty}^p$ and

$$\varepsilon_i^L = \sqrt{\|\Gamma_i\|_{0,\infty}^p} \cdot \left( \varepsilon_{i-1}^L + \mu(D_i)(\|\Gamma_i\|_{0,\infty}^s - 1)|\Gamma_i^{\max}| \right)$$

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?

○ Lets use the Basis Pursuit instead ...

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

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

\[
\vdots
\]

\[
\Gamma_{K-1}^{\text{LBP}} = \min_{\Gamma_{K-1}} \frac{1}{2} \|\Gamma_{K-2}^{\text{LBP}} - D_{K-1} \Gamma_{K-1}\|_2^2 + \lambda_{K-1} \|\Gamma_{K-1}\|_1
\]

\[
\Gamma_K^{\text{LBP}} = \min_{\Gamma_K} \frac{1}{2} \|\Gamma_{K-1}^{\text{LBP}} - D_K \Gamma_K\|_2^2 + \lambda_K \|\Gamma_K\|_1
\]

\[
\text{(DCP}_\lambda^\varepsilon\text{): Find } \{\Gamma_j\}_{j=1}^K \text{ s.t.}
\]

\[
\begin{align*}
\|Y - D_1 \Gamma_1\|_2 &\leq \varepsilon & \|\Gamma_1\|_{0,\infty} &\leq \lambda_1 \\
\Gamma_1 &= D_2 \Gamma_2 & \|\Gamma_2\|_{0,\infty} &\leq \lambda_2 \\
\vdots & & \vdots & \\
\Gamma_{K-1} &= D_K \Gamma_K & \|\Gamma_K\|_{0,\infty} &\leq \lambda_K
\end{align*}
\]

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

---

**Problems:**

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

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

Layered Iterative Soft-Thresholding Algorithm (ISTA):

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

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

Can be seen as a very deep residual neural network

[He, Zhang, Ren, & Sun ‘15]
Where are the Labels?

Answer 1:

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

We presented the ML-CSC as a machine that produces signals $X$.

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

$\Gamma_i$ is $L_{0,\infty}$ sparse.
Where are the Labels?

Answer 2:

- In fact, this model could be augmented by a synthesis of the corresponding label by:

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

- This assumes that knowing the representations suffices for classification \(\rightarrow\) supervised mode

- Thus, a successful pursuit algorithm can lead to an accurate recognition if the network is augmented by a FC classification layer

- In fact, we can analyze theoretically the classification accuracy and the sensitivity to adversarial noise – see later

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 – unsupervised and 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 new approach for understanding CNN and getting global optimality.
Fresh from the Oven

My team’s work proceeds along the above-described line of thinking:


On a Personal Note ...

Disclaimer: I am biased, so take my words with a grain of salt ...

Conjecture: Sparse modeling of data is at the heart of Deep-Learning architectures, and as such it is one of the main avenues for developing theoretical foundations for this field.

Elad (‘19)

My research activity (past, present & future) is dedicated to establishing this connection and addressing various aspects of it (applicative & theoretical)
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.

Start the Professional Certificate Program

Courses in the Professional Certificate Program

Courses:

- Sparse Representations in Signal and Image Processing: Fundamentals
- Sparse Representations in Image Processing: From Theory to Practice

Instructors:

- Yaniv Romano
- Michael Elad
More on these (including these slides and the relevant papers) can be found in http://www.cs.technion.ac.il/~elad