From Shallow to Deep Sparsity with Convolutional Networks

Jeremias Sulam

CoSIP Intense Course on Deep Learning

Joint work with

Vardan Papyan  Yaniv Romano  Michael Elad

Supported by ERC Grant no. 320649
Sparse and Redundant Representations

Theory $\rightarrow$ Algorithms $\rightarrow$ Applications

Generative models to provide theoretically justified algorithms and performance

The end of this talk:

**Multi-Layer Convolutional Sparse Modeling**
1. **Modeling**
   Why do we need models?

2. **Sparse Modeling**
   What are the known guarantees, algorithms, applications?

3. **Convolutional Sparse Modeling**
   What happens to all the above if we now address the convolutional scenario?

4. **Multi-Layer Convolutional Sparse Modeling**
   Did someone say CNNs?
Why do we need Models?

“Nothing is more practical than a good theory” – Vladimir N. Vapnik
All data has inherent **structure** than can be exploited

This structure enables different **processing** tasks to be carried out
Modeling Example - JPEG

Discrete Cosine Trans.
Image Models

Fourier Smoothness

Energy

\[ L_p(x) = \lambda \| x \|_2^2 \]

Smoothness

\[ L_p(x) = \lambda \| Lx \|_2^2 \]

DCT Smoothness

\[ E = \|x\|_2 \]

Gaussian Mixture Models

Total Variation

\[ L_p(x) = \lambda \| \nabla x \|_1 \]

Sparse & Redundant Representations

\[ L_p(x) = \lambda \| y \|_0 \]

for \( D y = x \)

Wavelets

\[ L_p(x) = \lambda \| Wx \|_1 \]

PCA

Beltrami Flow

Deep CNNs

...?
Sparse Representations

“Numquam ponenda est pluralitas sine necessitate”

Occam’s razor

OCCAM'S RAZOR
A Parsimonious Shave Every Time!
How to find $\gamma_i$?

**Pursuit - Sparse Coding**

\[
(P_0) : \min_{\gamma} \|\gamma\|_0 \quad \text{s.t.} \quad x_i = D\gamma_i
\]

(BUT) Cannot be solved!
Sparse Representations

Characterization of the Dictionary

Mutual Coherence:

\[ \mu(D) = \max_{i \neq j} |d_i^T d_j| \]

[Donoho & Elad, 2003]

Uniqueness Guarantees

Given the system \( x = D\gamma \), if \( \|\gamma\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu(D)} \right) \), then \( \gamma \) is the sparsest solution.

[Donoho & Elad, 2003]
Assume now \( y = D\gamma + v \), with \( \|v\|_2 \leq \epsilon \)

\[
(P_0^\epsilon) : \min_{\gamma} \|\gamma\|_0 \quad \text{s.t.} \quad \|y - D\gamma\|_2^2 \leq \epsilon^2
\]

Restricted Isometry Property - RIP

\( D \) is said to satisfy \( k \)-RIP with constant \( \delta_k \) if

\[
(1 - \delta_k)\|\alpha\|_2^2 \leq \|D\alpha\|_2^2 \leq (1 + \delta_k)\|\alpha\|_2^2
\]

holds true for any \( \alpha \) with \( \|\alpha\|_0 = k \).

Have we lost hope in finding \( \gamma \)?

Stability

If the true representation \( \gamma \) satisfies \( \|\gamma\|_0 = k < \frac{1}{2} \left( 1 + \frac{1}{\mu(D)} \right) \), then

\[
\|\gamma - \hat{\gamma}\|_2^2 \leq \frac{4\epsilon^2}{1 - \delta_{2k}} \leq \frac{4\epsilon^2}{1 - (2k - 1)\mu(D)}
\]

since \( \delta_k \leq (k - 1)\mu(D) \).
Pursuit Algorithms

$$(P_0^c): \min_\gamma \| y - D\gamma \|^2_2 \quad \text{s.t.} \quad \|\gamma\|_0 \leq k$$

- **Greedy Algorithms**
  - (Orthogonal) Matching Pursuit
    
    Build support of $\gamma$ progressively, one iteration at a time
  - Hard Thresholding
  - Iterative Hard Thresholding

  $$\hat{\gamma}^{t+1} = H_k (\hat{\gamma}^t - \eta D^T (D\hat{\gamma}^t - y))$$

- **Relaxation Approaches**

  $$(P_1): \min_\gamma \| y - D\gamma \|^2_2 + \lambda \|\gamma\|_1 \quad \text{- Basis Pursuit (BP)}$$

  - Convex optimization tools
  - Soft Thresholding
  - Iterative Soft Thresholding

... and many other variations.
Pursuit Algorithms

These algorithms... do they work?

\[(P_0^\epsilon) : \min_{\gamma} \|\gamma\|_0 \quad \text{s.t.} \quad \|y - D\gamma\|_2^2 \leq \epsilon^2\]

**Theorem: Stability of OMP**

If \(y = D\gamma + v\), \(\|v\|_2 = \epsilon\), and \(\|\gamma\|_0 < \frac{1}{2} \left(1 + \frac{1}{\mu(D)}\right) - \frac{1}{\mu(D)} \frac{\epsilon}{|\Gamma_{\min}|}\), then OMP will

- Run for \(k\) iterations
- Find the correct support
- Stable solution

\[\|\hat{\gamma}_{\text{OMP}} - \gamma\|_2^2 \leq \frac{\epsilon^2}{1 - \mu(D)(\|\gamma\|_0 - 1)}\]

✓ Perfect reconstruction in the noiseless case (\(\epsilon = 0\))
Pursuit Algorithms

These algorithms... do they work?

\[(P_1^\epsilon) : \min_{\gamma} \|\gamma\|_1 \quad \text{s.t.} \quad \|y - D\gamma\|_2^2 \leq \epsilon^2\]

**Theorem: Stability of BPDN**

If \(y = D\gamma + v\), \(\|v\|_2 = \epsilon\), and \(\|\gamma\|_0 \leq \frac{1}{4} \left(1 + \frac{1}{\mu(D)}\right)\), then BPDN will

- **Stable solution**
  \[
  \|\hat{\gamma}_{BP} - \gamma\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(D)(4\|\gamma\|_0 - 1)}
  \]

✓ **Perfect reconstruction** in the noiseless case (\(\epsilon = 0\))

All these results... how pessimistic ("limiting") are they?

Average performance results are available too, showing much better bounds

[Donoho (‘04)] [Candes et.al. (04)] [Tanner et.al. (05)] [E. (06)] [Tropp et.al. (06)] ... [Candes et. al. (09)]
Pursuit Algorithms

What about the simplest pursuits?

Stability of Hard Thresholding

\[ \hat{\gamma} = \mathcal{H}_\lambda (D^T y) \]

Hard Thresholding recovers \( \hat{\gamma} \) if \( \|\gamma\|_0 < \frac{1}{2} \left( 1 + \frac{|\gamma_{\min}|}{|\gamma_{\max}|} \frac{1}{\mu(D)} \right) - \frac{1}{\mu(D)} \frac{\epsilon}{|\gamma_{\max}|} \) such that

- Recovery of the support
- \( \| \hat{\gamma} - \gamma \|_2 \leq \sqrt{\| \gamma \|_0} \left( \epsilon + \mu(D) (\|\gamma\|_0 - 1) |\gamma_{\max}| \right) \)

Stability of Soft Thresholding

\[ \hat{\gamma} = S_\beta (D^T y) \]

Soft Thresholding recovers \( \hat{\gamma} \) if \( \|\gamma\|_0 < \frac{1}{2} \left( 1 + \frac{|\gamma_{\min}|}{|\gamma_{\max}|} \frac{1}{\mu(D)} \right) - \frac{1}{\mu(D)} \frac{\epsilon}{|\gamma_{\max}|} \) such that

- Recovery of the support
- \( \| \hat{\gamma} - \gamma \|_2 \leq \sqrt{\| \gamma \|_0} \left( \epsilon + \mu(D) (\|\gamma\|_0 - 1) |\gamma_{\max}| + \beta \right) \)

\( \times \) Imperfect reconstruction in the noiseless case (\( \epsilon = 0 \))
What about the Dictionary $D$?

### Dictionaries for Sparse Representations

**Analytical dictionaries**  Transforms that sparsify data:

- Wavelets [Mallat et al], Curvelets [Candes et al], Shearlets [Kutyniok et al], ...

**Adaptable dictionary**

\[
\min_{\Gamma,D} \| Y - D\Gamma \|_F^2 \quad \text{s.t.} \quad \| \gamma_i \|_0 \leq k \quad \forall \ i
\]
**Dictionary Learning**

\[
\min_{\Gamma, D} \|Y - D\Gamma\|_F^2 \quad \text{s.t.} \quad \begin{align*}
\|\gamma_i\|_0 & \leq k, \quad \forall \ i, \\
\|d_j\|_2 & = 1, \quad \forall \ j
\end{align*}
\]

**General Approach: Block Coordinate Minimization**

- \(\Gamma^{t+1} \leftarrow \arg \min_{\Gamma} \|Y - D^t\Gamma\|_F^2\) \quad \text{s.t.} \quad \|\gamma_i\|_0 \leq k, \quad \forall \ i \quad \rightarrow \text{Sparse coding}\)
- \(D^{t+1} \leftarrow \arg \min_{D} \|Y - D\Gamma^t\|_F^2\) \quad \text{s.t.} \quad \|d_j\|_2 = 1, \quad \forall \ j \quad \rightarrow \text{Dictionary Update}\)

**Dictionary Learning Methods**

- Atom-wise approach with SVD - K-SVD \[\text{[Aharon et al, 2006]}\]
- Online Learning - ODL \[\text{[Mairal et al, 2009]}\]

...
Universal Dictionaries

What does a universal dictionary look like?

[Sulam et al, 2016]
Dictionary Learning in Image Processing

Formulation

\[
\min_{x, \gamma_i, D} \frac{\lambda}{2} \|y - x\|_2^2 + \sum_i \|D\gamma_i - R_i x\|_2^2 + \mu_i \|\gamma_i\|_0
\]

1. Extract all patches \(R_i y\) into the matrix \(Y\)
2. Fix \(x\) and solve \(\min_{\Gamma, D} \|Y - D\Gamma\|_F^2\) s.t. \(\|\gamma_i\|_0 \leq k\)
   Using K-SVD, ODL, ...
3. \(D\) and \(\gamma_i\) and solve for \(x\) – weighted averaging
Dictionary Learning in Image Processing

- (Gaussian) Denoising

[Mairal et. al., 2008]
Dictionary Learning in Image Processing

Inpainting formulation

\[
\min_{x, \gamma_i, D} \frac{\lambda}{2} \| y - M x \|^2 + \sum_i \| D \gamma_i - R_i x \|^2 + \mu_i \| \gamma_i \|_0
\]

[Mairal et. al., 2008]
Dictionary Learning in Image Processing

- Face Image Compression

Original
JPEG
JPEG-2000
K-SVD


[Bryt et. al., 2008]
Dictionary Learning in Image Processing

- Blind Deblurring

[Shao et. al., 2014]
Interlude - Massive Open Online Course!

- 2 courses
- > 1,700 students
- 104 countries
How come we have managed to treat **global** problems with only **local** modeling?

- Why treat all patches at the same scale?
  - Multi-Scale Approaches [Ophir et al, Sulam et al, Papyan et al]

- Why treat all patches independently?
  - Joint sparse coding [Ram et al, Romano et al, Mairal et al]

- Why just averaging at the end?
  - EPLL [Sulam et al, 2015], Boosting [Romano, 2015]

**Missing theoretical Backbone!**

For every \(i^{th}\) patch, \(R_i x = D\gamma_i, \|\gamma_i\| \ll k\)

- What is the underlying global model?
- Who are these signals?
- How should the pursuit be carried?
- How should the (global!) model be trained?
1 Modeling

2 Sparse Modeling

3 Convolutional Sparse Modeling

4 Multi-Layer Convolutional Sparse Coding

5 Conclusion
Convolutional Sparse Representations

\[ \mathbf{X} = \sum_{j=1}^{m} \mathbf{d}_j \ast \mathbf{\Gamma}_j \]
Convolutional Sparse Representations

\[ X = \sum_{j=1}^{m} d_j \ast \Gamma_j = D\Gamma \]
Convolutional Sparse Representations

### Why should we care?
- **Global model** with *shift-invariant local prior*
- Inherently **no disagreement** between overlapping patches
- Related to current practices (i.e., *patch averaging*)

\[
X = D\Gamma = \frac{1}{n} \sum_i R_i^T \Omega \gamma_i
\]

- **Growing Applications**: Pattern Detection [Mrup et al 08, Vidal et al 17], **Inpainting** [Heide, Heidrich & Wetzstein 15], **Super-resolution** [Gu, Zuo, Xie, Meng, Feng & Zhang 15], **CNNs**

### Formulation

\[
(P_1) : \min_{\Gamma} \frac{1}{2} \| Y - D\Gamma \|_2^2 + \lambda \| \Gamma \|_1
\]

Is this well founded?
Consider the following example

- Assume $m = 2$, $n = 64$.
- Then $\mu(D) \geq 0.063$
- Thus $\|\Gamma\|_0 < \frac{1}{2} \left(1 + \frac{1}{\mu(D)}\right) \approx 8$ i.e., $O(\sqrt{n})$

8 non-zeros globally! for an entire image! and of any size!
A localized formulation

$$\|\Gamma\|_{0,\infty}^s \triangleq \max_i \|\gamma_i\|_0$$

$$(P_{0,\infty}) : \min_{\Gamma} \|\Gamma\|_{0,\infty}^s \quad \text{s.t.} \quad D\Gamma = X$$

Is the solution to this problem unique? Can we retrieve it algorithmically?
Uniqueness via mutual coherence

\[(P_{0,\infty}): \min_{\Gamma} \|\Gamma\|_{0,\infty} \text{ s.t. } D\Gamma = X.\]

**Theorem**

*If a solution $\Gamma$ exists for the $P_{0,\infty}$ problem such that

$$\|\Gamma\|_{0,\infty} < \frac{1}{2} \left(1 + \frac{1}{\mu(D)}\right),$$

then this is necessarily the unique globally optimal solution.*

- This pose a **local constraint** for **global guarantees**, so they are far more optimistic compared to global constraints.

In the previous example ($m = 2$, $n = 64$), one can now allow 8 **non-zeros per stripe**; i.e., $\mathcal{O}(N)$. 
Recovery Guarantees

\[(P_{0,\infty}) : \min_\Gamma \|\Gamma\|_{0,\infty} \quad \text{s.t.} \quad D\Gamma = X.\]

**Theorem**

*If a solution $\Gamma$ exists for the $P_{0,\infty}$ problem such that

$$\|\Gamma\|_{0,\infty} < \frac{1}{2} \left(1 + \frac{1}{\mu(D)}\right),$$

then OMP and BP are guaranteed to find it.*

- Both OMP and BP are **global** pursuits: they do not assume local sparsity, though still succeed in solving the $P_{0,\infty}$ problem.

- **How about variants that would assume local sparsity?**

Convolutional Sparse Modeling

From ideal to noisy signals

\[ Y = D\Gamma + E, \quad \|E\|_2 \leq \epsilon \]

Modified pursuit

\[(P_0^\epsilon, \infty) : \min_{\Gamma} \|\Gamma\|_0^\epsilon, \infty \quad \text{s.t.} \quad \|Y - D\Gamma\|_2^2 \leq \epsilon^2.\]

Some practical questions:

- Is the solution stable?
- Is the solution obtained with OMP/BP close to the true one?
- Do we really need to solve a \textbf{global} pursuit?
Stability of the $P_{0,\infty}$ problem

**Stripe-RIP**

$D$ is said to satisfy $k$-SRIP (Stripe-RIP) with constant $\delta_k$ if

$$\forall \Delta \quad (1 - \delta_k)\|\Delta\|_2^s \leq \|D\Delta\|_2^2 \leq (1 + \delta_k)\|\Delta\|_2^2$$

holds true for any $\Delta$ with $\|\Delta\|_{0,\infty} = k$.

Say $\hat{\Gamma} = \arg\min_{\Gamma} \|\Gamma\|_{0,\infty}$ s.t. $\|Y - D\Gamma\|_2^2 \leq \epsilon^2$. How good of a solution is $\hat{\Gamma}$?

**Theorem**

*If the true representation $\Gamma$ satisfies $\|\Gamma\|_{0,\infty} = k < \frac{1}{2} \left( 1 + \frac{1}{\mu(D)} \right)$, then*

$$\|\Gamma - \hat{\Gamma}\|_2^2 \leq \frac{4\epsilon^2}{1 - \delta_{2k}} \leq \frac{4\epsilon^2}{1 - (2k - 1)\mu(D)}.$$  

(*since $\delta_k \leq (k - 1)\mu(D)$*)
Say we obtain an estimate $\hat{\Gamma}$ with OMP, how close is it to the underlying true vector?

**Theorem: Stability of OMP**

If $Y = D\Gamma + E$, $\epsilon_L = \|E\|_2,\infty^p = \max_i \|R_i E\|_2$, and

$$\|\Gamma\|_{0,\infty}^s < \frac{1}{2} \left( 1 + \frac{1}{\mu(D)} \right) - \frac{1}{\mu(D)} \cdot \frac{\epsilon_L}{|\Gamma_{min}|},$$

then, after $\|\Gamma\|_0$ iterations, OMP will

1. Find the correct support
2. $\|\hat{\Gamma}_{\text{OMP}} - \Gamma\|_2^2 \leq \frac{\epsilon^2}{1 - \mu(\|\Gamma\|_{0,\infty}^s - 1)}$
Stability of Pursuit Methods

Say we obtain an estimate $\hat{\Gamma}$ with **Basis Pursuit**, how close is it to the underlying true vector?

**Theorem: Stability of BP**

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

If $Y = D\Gamma + E$, and $\lambda = 4 \| E \|_p,\infty$, and $\| \Gamma \|_{0,\infty} < \frac{1}{3} \left( 1 + \frac{1}{\mu(D)} \right)$, then,

1. $\text{Supp}\{\Gamma_{BP}\} \subset \text{Supp}\{\Gamma\}$.
2. $\| \hat{\Gamma}_{BP} - \Gamma \|_\infty \leq 7.5 \| E \|_p,\infty = 7.5 \epsilon_L$.
3. All entries greater than $7.5 \epsilon_L$ will be found.
4. $\hat{\Gamma}_{BP}$ is unique.

- This provides a theoretical justification of recent – practical – works dealing with CSC
- [Bristow, Eriksson & Lucey 13], [Wohlberg 14], [Kong & Fowlkes 14], [Bristow & Lucey 14], [Heide, Heidrich & Wetzstein 15], [Sorel & Sroubek 16], [Vidal et al, 17]
Convolutional Pursuit via Local Processing

Traditional Methods
- Work on Fourier Domain to reduce complexity
- Don’t scale well to large images
- Don’t scale well to many channels

Follow a local analysis!

\[ X = D\Gamma = \sum_i R_i^T D_L \alpha_i \]

\( s_i \): slices
Convolutional Pursuit via Local Processing

\[
\begin{align*}
\min_{\Gamma} & \quad \frac{1}{2} \| Y - D\Gamma \|_2^2 + \lambda \| \Gamma \|_1 \\
\downarrow & \\
\min_{s_i, \alpha_i} & \quad \frac{1}{2} \| Y - \sum_{i} R_i^T s_i \|_2^2 + \lambda \sum_{i} \| \alpha_i \|_1 \quad \text{s.t.} \quad s_i = D_L \alpha_i \\
\downarrow & \\
\min_{s_i, \alpha_i, u_i} & \quad \frac{1}{2} \| Y - \sum_{i} R_i^T s_i \|_2^2 + \lambda \sum_{i} \| \alpha_i \|_1 + \frac{1}{\rho} \sum_{i} \| s_i - D_L \alpha_i + u_i \|_2^2
\end{align*}
\]
Convolutional Dictionary **Learning** based on Local Processing

patches

slices

**Algorithm**

- **Local Pursuit**
  \[
  \min_{\alpha_i} \frac{1}{2} \| s_i + u_i - D_L \alpha_i \|_2^2 + \lambda \| \alpha_i \|_1
  \]
  (LARS, OMP, FISTA @ GPU, ...) 

- **Slice Estimate**
  \[ p_i \leftarrow \frac{1}{\rho} R_i Y + D_l \alpha_i - u_i \]

- **Slice Aggregation**
  \[ \hat{X} \leftarrow \sum_i R_i^T p_i \]

- **Local Laplacian**
  \[ s_i \leftarrow p_i - \frac{1}{\rho + n} R_i \hat{X} \]

- **Dual Update**
  \[ u_i \leftarrow u_i + s_i - D_L \alpha_i \]

- **Dictionary Update**
  \[
  \min_D \sum_i \| s_i + u_i - D_L \alpha_i \|_2^2
  \]
  (K-SVD, ODL, Trainlets, ...)
Partial Summary of CSC

- Global guarantees under local sparsity constraints
- The claims are far more flexible than traditional ones
- Guarantees for pursuit methods in recovering the solution (or their stability)
- The global pursuit can be decomposed into local operations
Contents

1 Modeling

2 Sparse Modeling

3 Convolutional Sparse Modeling

4 Multi-Layer Convolutional Sparse Coding

5 Conclusion
CSC and CNN

Convolutional Neural Networks
- Composition of convolutional filters
- Adaptive to data

Convolutional Sparse Coding
- Single layer of CSC
- Dictionaries are adapted to data
- Underlying sparse model
- Theoretical analysis of related algorithms

Multi-Layer ↔ Convolutional Sparse Coding
Multi-Layer CSC

\[ \mathbf{X} = \mathbf{D}_1 \]

\[ \mathbf{D}_2 \]

\[ \mathbf{\Gamma}_2 \]
ML-CSC Definition

Given a set of convolutional dictionaries \( \{D_i\}_{i=1}^{L} \), a signal \( X \in \mathbb{R}^N \) admits a representation in terms of the ML-CSC model if

\[
X = D_1 \Gamma_1, \quad \|\Gamma_1\|_{0,\infty}^s \leq \lambda_1,
\]

\[
\Gamma_1 = D_2 \Gamma_2, \quad \|\Gamma_2\|_{0,\infty}^s \leq \lambda_2,
\]

\[
\quad \vdots
\]

\[
\Gamma_{K-1} = D_K \Gamma_K, \quad \|\Gamma_K\|_{0,\infty}^s \leq \lambda_K.
\]

- \( M_{\lambda} \) the set of signals satisfying the ML-CSC assumption.
- If \( X(\Gamma_i) \in M_{\lambda} \), then

\[
X(\Gamma_i) = D_1 D_2 \ldots D_K \Gamma_K = D^{(K)} \Gamma_K
\]

Effective Dictionary
A New Problem Formulation

Say we get $Y = X(\Gamma_i) + E$, how to (deep) sparse code?

Deep Coding Problem

\[
(DCP_\lambda^\xi) \colon \text{find } \{\Gamma_i\}_{i=1}^K \text{ s.t. } \begin{align*}
\|Y - D_1 \Gamma_1\|_2^2 &\leq \xi_0, \\
\|\Gamma_1 - D_2 \Gamma_2\|_2^2 &\leq \xi_1, \\
\|\Gamma_2 - D_3 \Gamma_3\|_2^2 &\leq \xi_2, \\
&\vdots \\
\|\Gamma_{K-1} - D_K \Gamma_K\|_2^2 &\leq \xi_{K-1}, \\
\|\Gamma_K\|_{0,\infty} &\leq \lambda_K,
\end{align*}
\]

Given $Y = D_1 \Gamma_1 + E$, how to find $\Gamma_1$?

Simplest alternative: $\hat{\Gamma}_1 = P_{\beta_1}(D_1^T Y)$
Solving the DCP_\lambda^\varepsilon

Layered Thresholding (LT) algorithm

\[ \hat{\Gamma}_2 = P_{\beta_2}(D_2^T \hat{\Gamma}_1) = P_{\beta_1}(D_1^T Y) \]

Written differently,

\[ \hat{\Gamma}_2 = \text{ReLU}(D_2^T \text{ReLU}(D_1^T Y + b_1) + b_2) \]

Forward Pass of CNN

The forward pass is a pursuit seeking for the sparse representations under the ML-CSC model
## Theoretical Claims for the DCP $\mathcal{E}_\lambda$

### Stability of the solution of DCP $\mathcal{E}_\lambda$

If a set of solutions $\{\Gamma_i\}_{i=1}^K$ satisfy $\|\Gamma_i\|_{0,\infty} \leq \frac{1}{2} \left( 1 + \frac{1}{\mu(D_i)} \right)$, then

$$
\|\hat{\Gamma}_i - \Gamma_i\|_2^2 \leq \frac{4\mathcal{E}_{i-1}^2}{1 - (2\|\Gamma_i\|_{s,\infty} - 1)\mu(D_i)}
$$

### Stability of the Multi-Layer Thresholding (a.k.a forward pass)

If a set of solutions $\{\Gamma_i\}_{i=1}^K$ satisfy $\|\Gamma_i\|_{0,\infty} \leq \frac{1}{2} \left( 1 + \frac{1}{\mu(D_i)} \frac{|\Gamma_i^{min}|}{|\Gamma_i^{max}|} \right) - \frac{1}{\mu(D_i)} \frac{\mathcal{E}_{i-1}^L}{|\Gamma_i^{max}|}$, then the forward pass will identify the correct support, and

$$
\|\hat{\Gamma}_i - \Gamma_i\|_2^2 \leq \sqrt{\|\Gamma_i\|_{0,\infty}^p \left( \mathcal{E}_{i-1}^L + \mu(D_i) \left( \|\Gamma_i\|_{0,\infty}^s - 1 \right) |\Gamma_i^{max}| + \beta_i \right)}
$$

Cisse et al, **Parseval Networks**, 2017: $R_i(D_i) = \frac{\beta}{2} \|D_i^T D_i - I\|_2^2$

- Even in the noiseless case, it is incapable of recovering the solution to the DCP $\lambda$.
- Its success depends on the ratio $|\Gamma_i^{min}| / |\Gamma_i^{max}|$
Multi-Layer Convolutional Sparse Coding

Multi-Layer Basis Pursuit

\[(\text{DCP}_\lambda^{\mathcal{E}}) : \text{find } \{\Gamma_i\}_{i=1}^K \text{ s.t. } \|Y - D_1 \Gamma_1\|_2^2 \leq \mathcal{E}_0, \quad \|\Gamma_1\|_{0,\infty}^s \leq \lambda_1 \]
\[\|\Gamma_1 - D_2 \Gamma_2\|_2^2 \leq \mathcal{E}_1, \quad \|\Gamma_2\|_{0,\infty}^s \leq \lambda_2\]

Layered BP

\[\hat{\Gamma}_i = \arg \min_{\Gamma_i} \frac{1}{2} \left\| \hat{\Gamma}_{i-1} - D_i \Gamma_i \right\|_2^2 + \zeta_i \left\| \Gamma_i \right\|_1\]

Stability

If \( \{\Gamma_i\}_{i=1}^K \) satisfy \( \|\Gamma_i\|_{0,\infty} \leq \frac{1}{3} \left( 1 + \frac{1}{\mu(D_i)} \right) \), then

- \( \text{Supp}\{\hat{\Gamma}_i\} \subseteq \text{Supp}\{\Gamma_i\} \)
- \( \|\hat{\Gamma}_i - \Gamma_i\|_{2,\infty}^p \leq 7.5^i \|E\|_{2,\infty}^p \prod_{j=1}^i \sqrt{\|\Gamma_j\|_{0,\infty}^p} \)
- Every sufficiently large entry will be recovered

✓ Exact recovery in noiseless case
✓ Independent of the signal contrast
× Bound increase with depth
Multi-Layer Convolutional Sparse Coding

Multi-Layer Basis Pursuit

\[
\text{Solve} \quad \min_{\Gamma_1} \quad \frac{1}{2} \| Y - D_1 \Gamma_1 \|_2^2 + \lambda_1 \| \Gamma_1 \|_1 \to \hat{\Gamma}_1
\]

\[
\text{Solve} \quad \min_{\Gamma_2} \quad \frac{1}{2} \| \hat{\Gamma}_1 - D_2 \Gamma_2 \|_2^2 + \lambda_2 \| \Gamma_2 \|_1 \to \hat{\Gamma}_2
\]

[Sun et al, *Supervised Deep Sparse Coding Networks*, '17]
Multi-Layer Basis Pursuit

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

with
\[
\Gamma_1^k \leftarrow S_{\lambda_1/c_1} \left( \Gamma_1^{k-1} + \frac{1}{c_1} D_1^T (Y - D_1 \Gamma_1^{k-1}) \right)
\]
Looking into the Networks

- The forward pass is a pursuit seeking for the sparse representations under the ML-CSC model.

![Graph showing sparsity across layers for different networks.](image)
Checkpoint Recap

✓ The forward pass in an CNN is a pursuit for signals following the multi-layer CSC!
✓ Theoretical claims for the Multi-layer Thresholding algorithm
✓ Layered BP presented as alternative with stronger guarantees

- How can we project signals onto the ML-CSC model?
- Is the model empty?
- How should the convolutional filters be trained?
- How is the learning of the ML-CSC model related to traditional CNN algorithms?
- How does it perform?
A Projection Approach

Say \( Y = X(\Gamma_i) + E, \quad X \in M_\lambda. \)

ML-CSC Projection \((P_{M_\lambda})\)

Given \( Y \) and convolutional dictionaries \( \{D_i\}_{i=1}^K \),

\[
(P_{M_\lambda}) : \quad \min_{\{\Gamma_i\}} \| Y - X(\Gamma_i) \|_2 \quad \text{s.t.} \quad X(\Gamma_i) \in M_\lambda.
\]

- Unlike the DCP \( E_\lambda \), the solution \( \hat{X} \in M_\lambda: \)

\[
\hat{X} = D_1 \hat{\Gamma}_1 = D_1 D_2 \hat{\Gamma}_2 = \cdots = D^{(i)} \hat{\Gamma}_i
\]

- A solution to the DCP \( E_\lambda \), provides \( \hat{\Gamma}_{i-1} \neq D_i \hat{\Gamma}_i \)
Stability of the $\mathcal{P}_{\mathcal{M}_\lambda}$ problem

**Theorem**

$X(\Gamma_i) \in \mathcal{M}_\lambda$ is observed through $Y = X(\Gamma_i) + E$, $\|E\|_2 \leq \mathcal{E}_0$, and $\|\Gamma_i\|_{0,\infty} = \lambda_i < \frac{1}{2} \left(1 + \frac{1}{\mu \left(D^{(i)}\right)}\right)$, for $1 \leq i \leq K$,

Then, the solution $\{\hat{\Gamma}_i\}_{i=1}^K$ to the $\mathcal{P}_{\mathcal{M}_\lambda}$ problem satisfies

$$\|\Gamma_i - \hat{\Gamma}_i\|_2^2 \leq \frac{4\mathcal{E}_0^2}{1 - (2\|\Gamma_i\|_{0,\infty} - 1)\mu \left(D^{(i)}\right)}$$

✓ Bound is not cumulative across layers
✓ Dependence on $\mu \left(D^{(L)}\right)$ - not necessarily a bad thing!
Pursuit Algorithms

- How to solve $P_{M_\lambda}$?
- How to seek for $\{\hat{\Gamma}_i\}$ while assuring $X(\Gamma_i) \in M_\lambda$?

**ML-CSC Pursuit**

- Global Pursuit:
  $$\hat{\Gamma}_K \leftarrow \arg\min_{\Gamma} \|Y - D^{(K)}\Gamma\|_2^2 \text{ s.t. } \|\Gamma\|_{0,\infty} \leq k;$$

- Finding the inner representations:
  $$\text{for } j = K, \ldots, 1 \text{ do}$$
  $$\hat{\Gamma}_{j-1} \leftarrow D_j \hat{\Gamma}_j$$
  $$\text{end}$$
Theorem: Stability of the Pursuit - $\ell_1$ case

\[ Y = X(\Gamma_i) + E, \quad X \in M_\lambda, \quad \|E\|_{2,\infty} \leq \epsilon_0. \quad \|\Gamma_i\|_{0,\infty} = \lambda_i < \frac{1}{2} \left(1 + \frac{1}{\mu(D_i)}\right), \]

\[ i = 1, \ldots, K - 1 \text{ and } \|\Gamma_K\|_{0,\infty} = \lambda_i \leq \frac{1}{3} \left(1 + \frac{1}{\mu(D(K))}\right). \quad \{\Gamma_i\} \text{ satisfy the N.V.S. for } D_i. \]

Let

\[
\hat{\Gamma}_K \leftarrow \arg \min_{\Gamma} \|Y + D^{(K)}\Gamma\|_2^2 + \zeta_L \|\Gamma\|_1
\]

\[
\hat{\Gamma}_{i-1} \leftarrow D_i \hat{\Gamma}_i, \quad i = K, \ldots, 1
\]

Then, for every $i^{th}$ layer,

- $\text{Supp}(\hat{\Gamma}_i) \subseteq \text{Supp}(\Gamma_i)$

- $\|\hat{\Gamma}_i - \Gamma_i\|_{2,\infty}^p \leq \epsilon_K \prod_{j=i+1}^L \sqrt{\frac{3c_j}{2}}, \quad \rightarrow \text{Tightest for the deepest layer!}$

**Non Vanishing Support property** $\Gamma$ will not cause atoms to be combined such that their supports cancel each other.
Theorem: Stability of the Pursuit - $\ell_0$ case

Suppose $Y = X(\Gamma_i) + E$, $\|Y - X\|_2 \leq \mathcal{E}_0$, and $\epsilon_0 = \|E\|_{2,\infty}$. Let $\Gamma_i$ satisfy the N.V.S. property for the respective dictionaries $D_i$, with $\|\Gamma_i\|_{s,\infty} = \lambda_i < \frac{1}{2} \left(1 + \frac{1}{\mu(D_i)}\right)$, for $1 \leq i \leq K$, and $\|\Gamma_K\|_{s,\infty} < \frac{1}{2} \left(1 + \frac{1}{\mu(D(K))}\right) - \frac{1}{\mu(D(K))} \cdot \frac{\epsilon_0}{|\Gamma_{\min}|}$, and

$$\hat{\Gamma}_K \leftarrow \arg \min_{\Gamma} \|Y - D^{(K)}\Gamma\|_2^2 \quad \text{s.t.} \quad \|\Gamma\|_{0,\infty} \leq \lambda_K \quad \text{(with OMP)}$$

$$\hat{\Gamma}_i \leftarrow D_{i+1} \hat{\Gamma}_{i+1}, \quad i = K, \ldots, 1$$

Then

1. $\text{Supp}(\hat{\Gamma}_i) \subseteq \text{Supp}(\Gamma_i)$,

2. $\|\hat{\Gamma}_i - \Gamma_i\|_2^2 \leq \frac{\epsilon_0^2}{1 - \mu(D(K)) (\|\Gamma_K\|_{s,\infty} - 1) \left(\frac{3}{2}\right)^{K-i}}$. 
What about the Dictionaries?

The existence of $X \in \mathcal{M}_\lambda$ depends on proper dictionaries $D_i$.

- Why should $\hat{\Gamma}_{i-1} = D_i \hat{\Gamma}_i$ be sparse?
- Is the model empty?

Example:

i) $D_i$ are Random Dictionaries, i.e., $d_{iK}^j = R_j^T v$, $v \sim \mathcal{N}(0,\sigma_i^2 I)$

ii) One can construct $\Gamma_K$ with $\|\Gamma_K\|_{0,\infty} \leq 2$ such that $\Pr(\Gamma_{i-1}^K = 0) = 0 \rightarrow dense!$

i.e, if $D$ is random, $\exists \Gamma$ such that $D\Gamma$ is sparse. In this case, the model is empty!

If one seeks for $\{\Gamma_i\}$, one must seek also for $\{D_i\}$ that would allow for that decomposition.
How to Learn?

$$\min_{\{\Gamma_i^t\}, \{D_i\}} \sum_{t=1}^{T} \|Y^t - X^t(\Gamma_i^t, D_i)\|_2^2 \quad \text{s.t.} \quad \begin{cases} X^t \in \mathcal{M}_\lambda, \\ \|d_i^j\|_2 = 1, \forall i, j \end{cases}$$

Problematic:
- The constraints on $\Gamma_i$ are coupled
- $\Gamma_i$ depends on $\{D_j\}_{j=i+1}^K$.

Sparsity Proxies

$$\Gamma_{K-1} = D_K \Gamma_K. \quad \Rightarrow \|\Gamma_{K-1}\|_{0,\infty}^s \leq c_K \|D_K\|_0 \|\Gamma_K\|_{0,\infty}^s$$

$$\|\Gamma_i\|_{0,\infty}^s \leq c \prod_{j=i+1}^{K} \|D_j\|_0 \|\Gamma_K\|_{0,\infty}^s.$$
Problem formulation

\[ \min_{\{\Gamma_t^K\}, \{D_i\}} \sum_{t=1}^T \|Y_t - D_1 D_2 \cdots D_K \Gamma_t^K \|_2^2 + \sum_{i=2}^K \zeta_i \|D_i\|_0 \quad \text{s.t.} \quad \|\Gamma_t^K\|_{s,\infty} \leq \lambda_K \]

Algorithm

**Data:** Training samples \( \{Y_i\} \), initial convolutional dictionaries \( D_i^0 \)

**for** \( t = 1, \ldots, T \) **do**

1. Draw \( Y^t \) at random;
2. **Sparse Coding:** \( \hat{\Gamma}_K \leftarrow \arg\min_{\Gamma} \|Y^t - D^{(K)} \Gamma\|_2 \quad \text{s.t.} \quad \|\Gamma\|_{s,\infty} \leq \lambda_K \) (IHT/FISTA);
3. **Update Dictionaries:**
   **for** \( k = K, \ldots, 1 \) **do**
   1. \( D_k \leftarrow \arg\min_{D_k} \|Y^t - D_1 \cdots D_k \cdots D_K \Gamma_K\|_2 + \zeta_k \|D_k\|_0 \) (PGD);
   **end**
**end**
Related work

Dictionary Learning

**Chasing-Butterflies**: \[
\min\|Y - \prod_{j=1}^{L+1} S_j\|_2^2, \quad S_j \text{ sparse} \quad [\text{LeMagoarou et al, 2015}]
\]

**Fast-Transforms Learning**: cascades of convolutions with sparse kernels [Chabiron et al, 2015]

**Trainlets**: Sparse combinations of shift-invariant wavelet atoms (which can be expressed as sparse convolutions!) [Sulam et al, 2016]

Auto-encoders

**Sparse AutoEncoders**: imposing sparse-enforcing loss in hidden layer [Ng, 2011]

**K-Sparse AutoEncoders**: \[
\min_{W,b,b'} \|Y - (WH_k(W^TX + b) + b')\|_2 \quad [\text{Makhzani, 2014}]
\]

**Winner-Take-All AutoEncoders**: “Spatial” sparsity + “life-time” sparsity [Makhzani, 2015]
Multi-Layer Convolutional Sparse Coding

Learning an MNIST model

Multi-Layer Convolutional Dictionaries:

\[ D_1 D_2 D_3, \]

Loss

Dictionary Sparsity

Average Residual
Learning an MNIST model

Multi-Layer Convolutional Decomposition:

\[ y \quad \hat{x} \]

\[ \hat{r}_1 \quad D_1 \]

\[ \hat{r}_2 \quad D_2 \]

\[ \hat{r}_3 \quad D_3 \]
**Sparse Recovery (Synthetic Data):**

**Layered-BP**

- Layer-Wise Representation Error
- Layer-Wise Support Error

**Projection**

- Projection Representation Error
- Projection Support Error
Sparse Recovery (MNIST Data):

Layered-BP:

Layer-Wise Representations Error

Projection:

Projection Representations Error

Layer-Wise Support Error

Projection Support Error
M-term Approximation

- Sparse Autoencoders
- ML-CSC (increasing sparsity)
- k-sparse Autoenc. (25-50-60)
- Trainlets

Relative Reconstruction Error vs. NNZ

- 0.006
- 0.01
- 0.05
- 0.13
- 0.35
- 1.36

From shallow to deep sparsity
Unsupervised Setting: After training a representation model, we compute features with it for each training example, and learn a linear classifier on them.

<table>
<thead>
<tr>
<th>Method</th>
<th>Classification Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stacked Denoising Autoencoder (3 layers)</td>
<td>1.28%</td>
</tr>
<tr>
<td>k-Sparse Autoencoder (1K units)</td>
<td>1.35%</td>
</tr>
<tr>
<td>Shallow WTA Autoencoder (2K units)</td>
<td>1.20%</td>
</tr>
<tr>
<td>Stacked WTA Autoencoder (2K units)</td>
<td>1.11%</td>
</tr>
<tr>
<td>ML-CSC (1K units) - 2nd Layer Rep.</td>
<td>1.30%</td>
</tr>
<tr>
<td>ML-CSC (2K units) - 2nd &amp; 3rd Layer Rep.</td>
<td>1.15%</td>
</tr>
</tbody>
</table>
Ongoing work

- Unsupervised Classification ...
  Cifar Dictionaries
Ongoing work

- Unsupervised Classification ...
- Supervised Training ...
- Generalization to average performance bounds ...

**Take Home Messages**

- Model assumptions enables us to propose algorithms serving signals in this model
- More importantly, it enables to develop theoretical guarantees for these algorithms
- In particular, the ML-CSC provides a formal framework for the study of CNN, architectures and algorithms