

# FROM SHALLOW TO DEEP SPARSITY WITH CONVOLUTIONAL NETWORKS

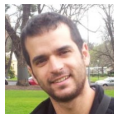
JEREMIAS SULAM

CoSIP INTENSE COURSE ON DEEP LEARNING

Joint work with



Vardan Papyan



Yaniv Romano



Michael Elad



Supported by ERC Grant

no. 320649

# This talk – A Führung<sub>(tour)</sub> of Sparse Modeling

## Sparse and Redundant Representations

Theory



Algorithms



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?

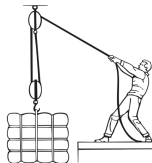
## ③ Convolutional Sparse Modeling

What happens to all the above if we now address the convolutional scenario?

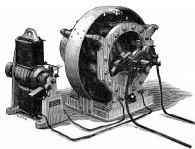
## ④ Multi-Layer Convolutional Sparse Modeling

Did someone say CNNs?

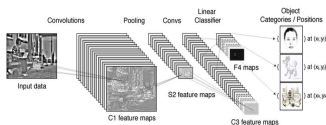
# Why do we need Models?



Newton



Maxwell

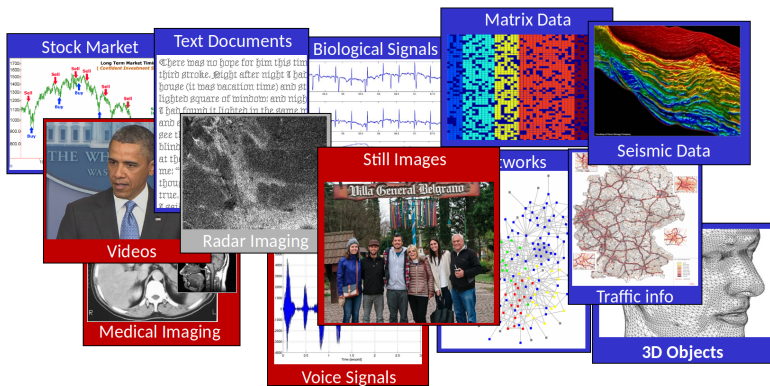


CoSIP ICDL  
Theories of Deep  
Learning?



“Nothing is more practical than a good theory” – Vladimir N. Vapnik

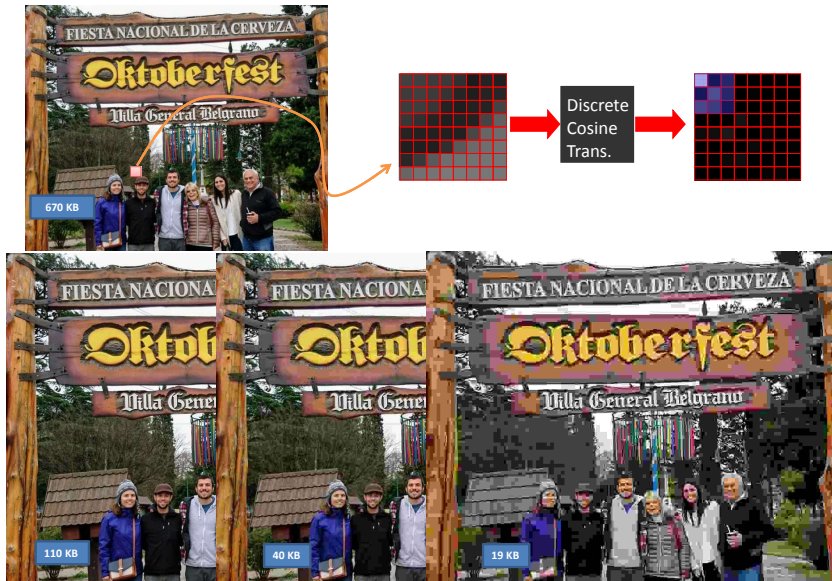
# Data Processing



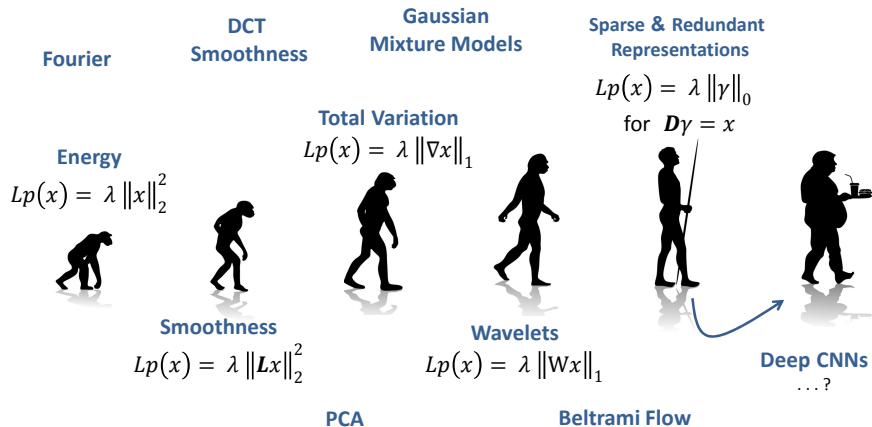
- All data has inherent **structure** than can be exploited
- This structure enables different **processing** tasks to be carried out

} **Signal Models**

# Example - JPEG



## Image Models



# Contents

- 1 Modeling
- 2 **Sparse Modeling**
- 3 Convolutional Sparse Modeling
- 4 Multi-Layer Convolutional Sparse Coding
- 5 Conclusion

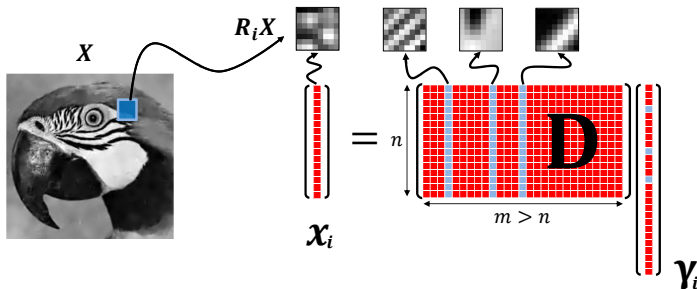
# Sparse Representations

*"Numquam ponenda est pluralitas sine necessitate "*

Occam's razor



# Sparse Representations



- 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!

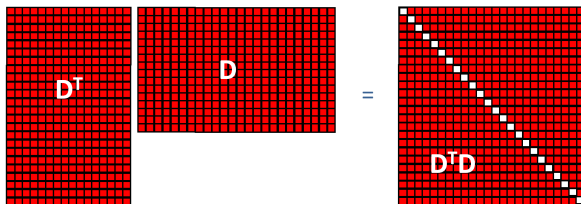
# Sparse Representations

## Characterization of the Dictionary

Mutual Coherence

$$\mu(\mathbf{D}) = \max_{i \neq j} |\mathbf{d}_i^T \mathbf{d}_j|$$

[Donoho & Elad, 2003]



## Uniqueness Guarantees

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

[Donoho & Elad, 2003]

# From Ideal to Noisy Signals

Assume now  $\mathbf{y} = \mathbf{D}\boldsymbol{\gamma} + \mathbf{v}$ , with  $\|\mathbf{v}\|_2 \leq \epsilon$

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

## Restricted Isometry Property - RIP

$\mathbf{D}$  is said to satisfy  $k$ -RIP with constant  $\delta_k$  if

$$(1 - \delta_k)\|\boldsymbol{\alpha}\|_2^2 \leq \|\mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq (1 + \delta_k)\|\boldsymbol{\alpha}\|_2^2$$

holds true for any  $\boldsymbol{\alpha}$  with  $\|\boldsymbol{\alpha}\|_0 = k$ .

Have we lost hope in finding  $\boldsymbol{\gamma}$ ?

## Stability

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

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

since  $\delta_k \leq (k - 1)\mu(\mathbf{D})$

# Pursuit Algorithms

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

## • Greedy Algorithms

### • (Orthogonal) Matching Pursuit

Build support of  $\gamma$  progressively, one iteration at a time

- Hard Thresholding
- Iterative Hard Thresholding

$$\hat{\gamma}^{t+1} = \mathcal{H}_k \left( \hat{\gamma}^t - \eta \mathbf{D}^T (\mathbf{D} \hat{\gamma}^t - \mathbf{y}) \right)$$

## • Relaxation Approaches

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

- Convex optimization tools
- Soft Thresholding
- Iterative Soft Thresholding

... and many other variations.

# Pursuit Algorithms

These algorithms... do they work?

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

## Theorem: Stability of OMP

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

- Run for  $k$  iterations
- Find the correct support
- Stable solution

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

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

# Pursuit Algorithms

These algorithms... do they work?

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

## Theorem: Stability of BPDN

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

- Stable solution

$$\|\hat{\boldsymbol{\gamma}}_{\text{BP}} - \boldsymbol{\gamma}\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(\mathbf{D})(4\|\boldsymbol{\gamma}\|_0 - 1)}$$

✓ **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

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

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

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

## Stability of Soft Thresholding

$$\hat{\gamma} = \mathcal{S}_{\beta} (\mathbf{D}^T \mathbf{y})$$

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

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

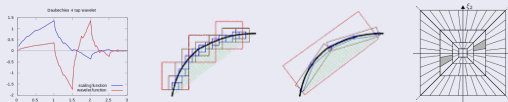
× **Imperfect reconstruction in the noiseless case** ( $\epsilon = 0$ )

# What about the Dictionary $D$ ?

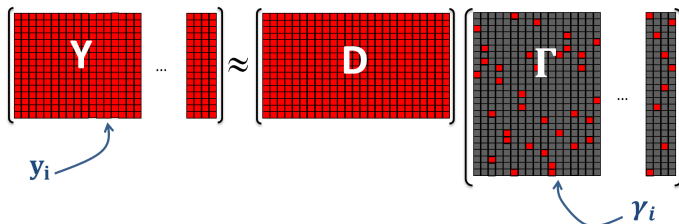
## Dictionaries for Sparse Representations

Analytical dictionaries Transforms that sparsify data:

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



Adaptable dictionary  $\min_{\Gamma, D} \|Y - D\Gamma\|_F^2 \text{ s.t. } \|\gamma_i\|_0 \leq k \quad \forall i$



# Dictionary Learning

$$\min_{\mathbf{\Gamma}, \mathbf{D}} \|\mathbf{Y} - \mathbf{D}\mathbf{\Gamma}\|_F^2 \text{ s.t. } \begin{cases} \|\gamma_i\|_0 \leq k, & \forall i, \\ \|\mathbf{d}_j\|_2 = 1, & \forall j \end{cases}$$

General Approach: Block Coordinate Minimization

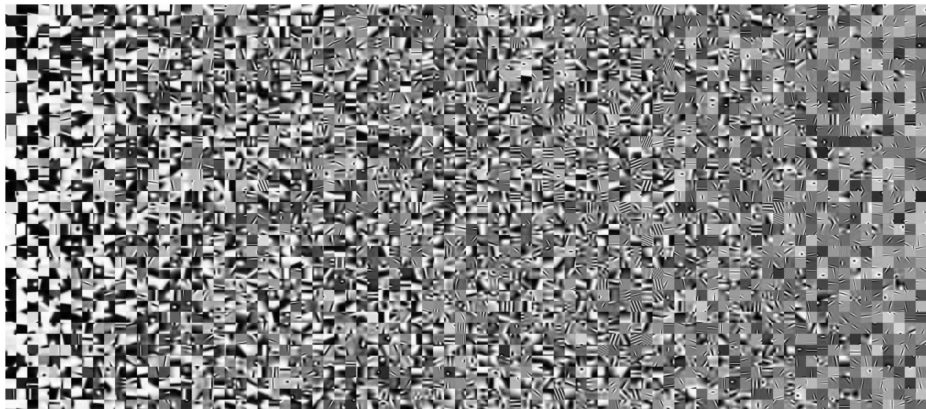
- $\mathbf{\Gamma}^{t+1} \leftarrow \arg \min_{\mathbf{\Gamma}} \|\mathbf{Y} - \mathbf{D}^t \mathbf{\Gamma}\|_F^2 \text{ s.t. } \|\gamma_i\|_0 \leq k \quad \forall i \rightarrow \text{Sparse coding}$
- $\mathbf{D}^{t+1} \leftarrow \arg \min_{\mathbf{D}} \|\mathbf{Y} - \mathbf{D} \mathbf{\Gamma}^t\|_F^2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

## Dictionary Learning Methods

- Least Squares solution - Method of Optimal Directions (MOD) [Engan et al, 2000]
- Atom-wise approach with SVD - K-SVD [Aharon et al, 2006]
- Online Learning - ODL [Mairal et al, 2009]
- ...

# Universal Dictionaries

What does a universal dictionary look like?

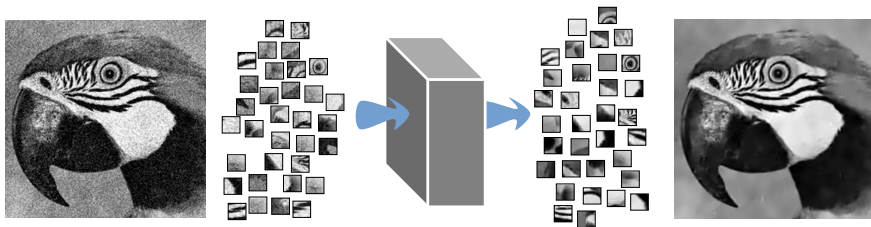


[Sulam et al, 2016]

# Dictionary Learning in Image Processing

## Formulation

$$\min_{\mathbf{x}, \gamma_i, \mathbf{D}} \frac{\lambda}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 + \sum_i \|\mathbf{D}\gamma_i - \mathbf{R}_i\mathbf{x}\|_2^2 + \mu_i \|\gamma_i\|_0$$

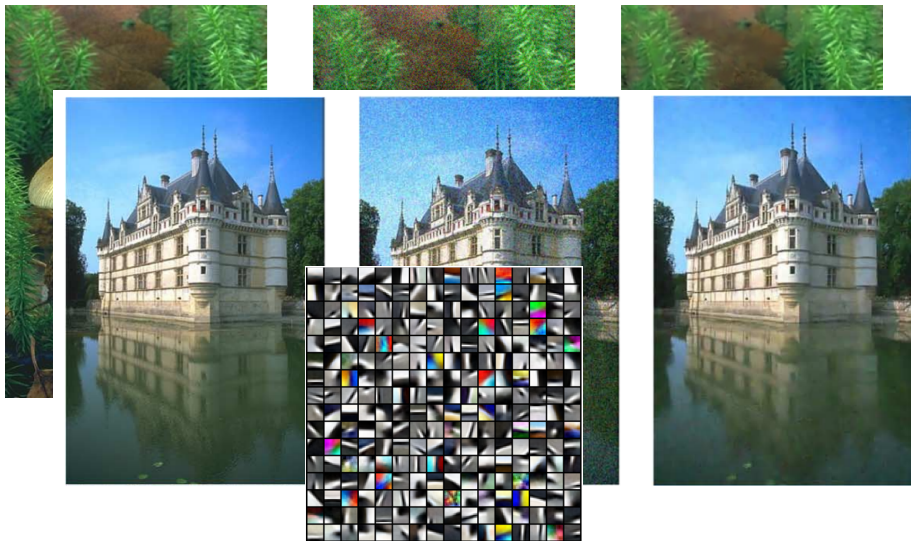


- ① Extract all patches  $\mathbf{R}_i\mathbf{y}$  into the matrix  $\mathbf{Y}$
- ② Fix  $\mathbf{x}$  and solve  $\min_{\Gamma, \mathbf{D}} \|\mathbf{Y} - \mathbf{D}\Gamma\|_F^2$  s.t.  $\|\gamma_i\|_0 \leq k$   
Using K-SVD, ODL, ...
- ③  $\mathbf{D}$  and  $\gamma_i$  and solve for  $\mathbf{x}$  – weighted averaging

# Dictionary Learning in Image Processing

- (Gaussian) Denoising

[Mairal et. al., 2008]

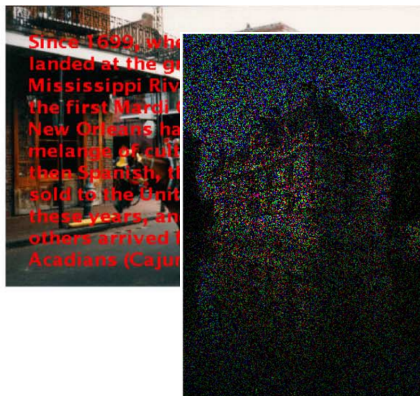


# Dictionary Learning in Image Processing

## Inpainting formulation

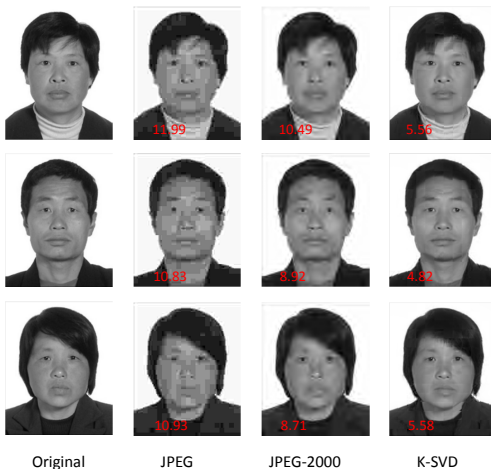
$$\min_{\mathbf{x}, \gamma_i, \mathbf{D}} \frac{\lambda}{2} \|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2^2 + \sum_i \|\mathbf{D}\gamma_i - \mathbf{R}_i\mathbf{x}\|_2^2 + \mu_i \|\gamma_i\|_0$$

[Mairal et. al., 2008]



# Dictionary Learning in Image Processing

- Face Image Compression



[Bryt et. al., 2008]

# Dictionary Learning in Image Processing

- Blind Deblurring



[Shao et. al., 2014]

# Interlude - Massive Open Online Course!



[Courses](#) ▾ [Programs](#) ▾ [Schools & Partners](#) [About](#) ▾

Search:



Sign In

[Register](#)



## 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



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](#)

- 2 courses
- > 1,700 students
- 104 countries

### Instructors



Yaniv Romano



Michael Elad

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, Pappas 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,  $\mathbf{R}_i \mathbf{x} = \mathbf{D} \boldsymbol{\gamma}_i, \quad \|\boldsymbol{\gamma}_i\| \ll k$

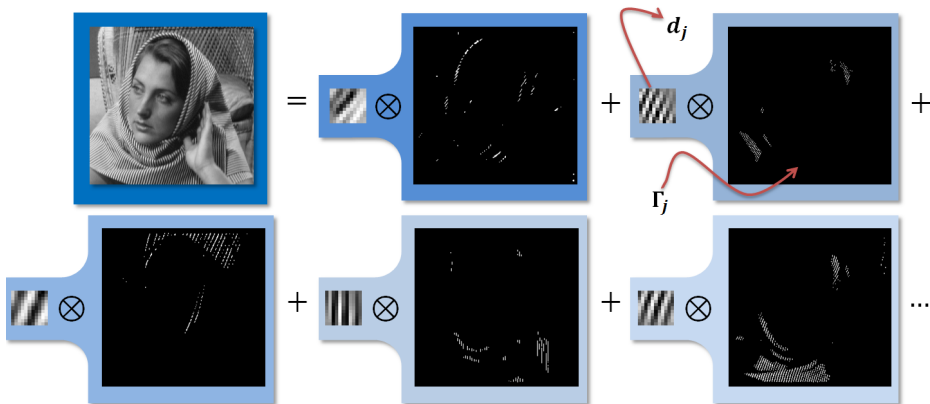
- What is the underlying global model?
  - Who are these signals?
  - How should the pursuit be carried?
- How should the (global!) model be trained?

# Contents

- 1 Modeling
- 2 Sparse Modeling
- 3 Convolutional Sparse Modeling**
- 4 Multi-Layer Convolutional Sparse Coding
- 5 Conclusion

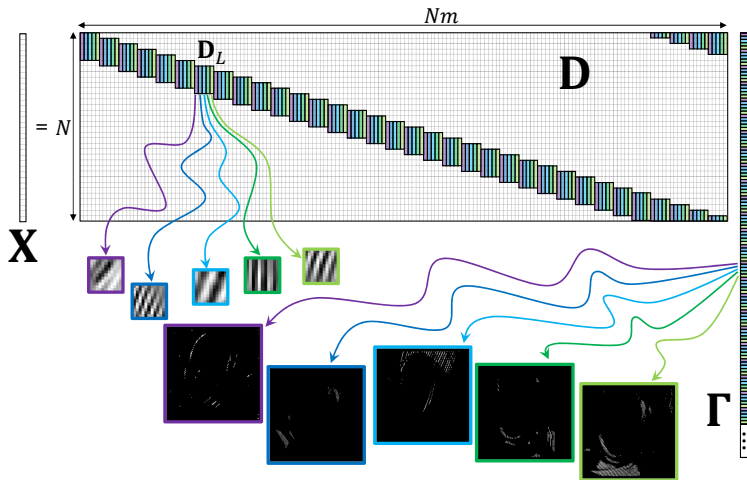
## Convolutional Sparse Representations

$$\mathbf{X} = \sum_{j=1}^m \mathbf{d}_j * \Gamma_j$$



## Convolutional Sparse Representations

$$\mathbf{X} = \sum_{j=1}^m \mathbf{d}_j * \Gamma_j = \mathbf{D}\Gamma$$



# 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*)

$$\mathbf{X} = \mathbf{D}\mathbf{\Gamma} = \frac{1}{n} \sum_i \mathbf{R}_i^T \mathbf{\Omega} \gamma_i$$

- 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

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

**Is this well founded?**

# Sparse Representations Theory

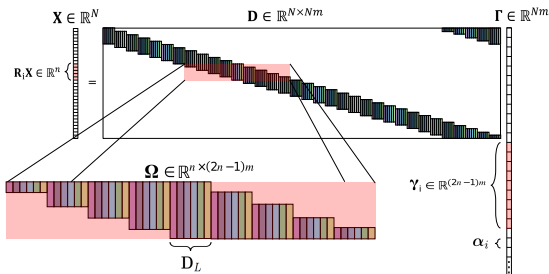
## Consider the following example

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



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

## From Global to Local



$$= \sum \mathbf{A}_i \otimes \mathbf{B}_i$$

The diagram shows a stack of grayscale images representing the components  $\mathbf{A}_i$  and  $\mathbf{B}_i$  in the sparse representation. A red arrow points to a specific component  $\mathbf{B}_i$  labeled  $\gamma_i$ .

## A localized formulation

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

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

Is the solution to this problem **unique**? Can we retrieve it **algorithmically**?

## Uniqueness via mutual coherence

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

## Theorem

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

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

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

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

## Theorem

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

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

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?

B. Wohlberg, *Convolutional Sparse Coding With Overlapping Group Norms*, ArXiv, 2017

# From ideal to noisy signals

$$\mathbf{Y} = \mathbf{D}\mathbf{\Gamma} + \mathbf{E}, \quad \|\mathbf{E}\|_2 \leq \epsilon$$

## Modified pursuit

$$(\mathbf{P}_{0,\infty}^\epsilon) : \quad \min_{\mathbf{\Gamma}} \quad \|\mathbf{\Gamma}\|_{0,\infty}^s \quad \text{s.t.} \quad \|\mathbf{Y} - \mathbf{D}\mathbf{\Gamma}\|_2^2 \leq \epsilon^2.$$

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 **global** pursuit?

# Stability of the $P_{0,\infty}^\epsilon$ problem

## Stripe-RIP

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

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

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

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

## Theorem

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

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

(since  $\delta_k \leq (k - 1)\mu(\mathbf{D})$ )

# Stability of Pursuit Methods

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

## Theorem: Stability of OMP

If  $\mathbf{Y} = \mathbf{D}\Gamma + \mathbf{E}$ ,  $\epsilon_L = \|\mathbf{E}\|_{2,\infty}^p = \max_i \|\mathbf{R}_i \mathbf{E}\|_2$ , and

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

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

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

If  $\mathbf{Y} = \mathbf{D}\Gamma + \mathbf{E}$ , and  $\lambda = 4\|\mathbf{E}\|_{2,\infty}^p$ , and  $\|\Gamma\|_{0,\infty}^s < \frac{1}{3} \left(1 + \frac{1}{\mu(\mathbf{D})}\right)$ , then,

- ①  $Supp\{\Gamma_{BP}\} \subset Supp\{\Gamma\}$ .
- ②  $\|\hat{\Gamma}_{BP} - \Gamma\|_{\infty} \leq 7.5\|\mathbf{E}\|_{2,\infty}^p = 7.5 \epsilon_L$ .
- 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]

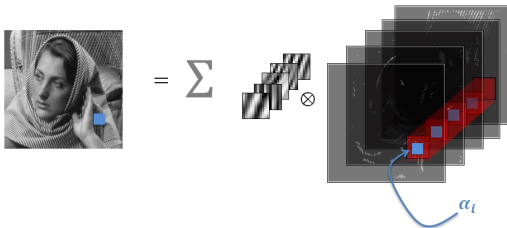
# Convolutional Pursuit via Local Processing

## 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!

$$\mathbf{X} = \mathbf{D}\mathbf{\Gamma} = \sum_i \mathbf{R}_i^T \underbrace{\mathbf{D}_L \alpha_i}_{\mathbf{s}_i: \text{ slices}}$$



## Convolutional Pursuit via Local Processing

$$\min_{\mathbf{\Gamma}} \frac{1}{2} \|\mathbf{Y} - \mathbf{D}\mathbf{\Gamma}\|_2^2 + \lambda \|\mathbf{\Gamma}\|_1$$

$$\downarrow$$

$$\min_{\mathbf{s}_i, \boldsymbol{\alpha}_i} \frac{1}{2} \|\mathbf{Y} - \sum_i \mathbf{R}_i^T \mathbf{s}_i\|_2^2 + \lambda \sum_i \|\boldsymbol{\alpha}_i\|_1 \quad \text{s.t.} \quad \mathbf{s}_i = \mathbf{D}_L \boldsymbol{\alpha}_i$$

$$\downarrow$$

$$\min_{\mathbf{s}_i, \boldsymbol{\alpha}_i, \mathbf{u}_i} \frac{1}{2} \|\mathbf{Y} - \sum_i \mathbf{R}_i^T \mathbf{s}_i\|_2^2 + \lambda \sum_i \|\boldsymbol{\alpha}_i\|_1 + \frac{1}{\rho} \sum_i \|\mathbf{s}_i - \mathbf{D}_L \boldsymbol{\alpha}_i + \mathbf{u}_i\|_2^2$$

# Convolutional Dictionary **Learning** based on Local Processing



## 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 Laplarian

$$s_i \leftarrow p_i - \frac{1}{\rho+n} R_i \hat{X}$$

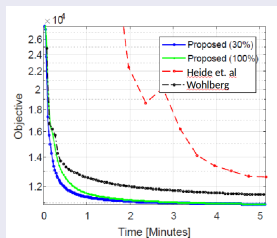
### Dual Update

$$u_i \leftarrow u_i + s_i - D_L \alpha_i$$

### 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

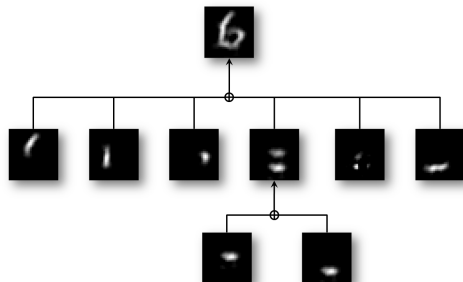
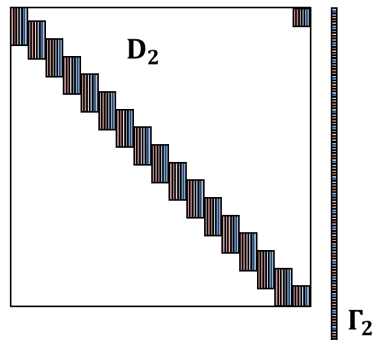
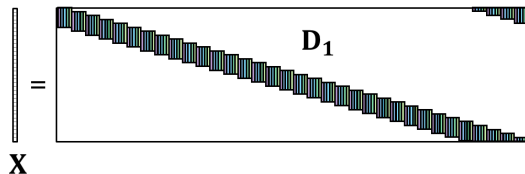
- Composition of convolutional filters
- Adaptive to data

Multi-Layer  $\Updownarrow$  Convolutional Sparse Coding

## Convolutional Sparse Coding

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

## Multi-Layer CSC



## ML-CSC Definition

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

$$\begin{aligned}\mathbf{X} &= \mathbf{D}_1 \boldsymbol{\Gamma}_1, & \|\boldsymbol{\Gamma}_1\|_{0,\infty}^s &\leq \lambda_1, \\ \boldsymbol{\Gamma}_1 &= \mathbf{D}_2 \boldsymbol{\Gamma}_2, & \|\boldsymbol{\Gamma}_2\|_{0,\infty}^s &\leq \lambda_2, \\ &\vdots \\ \boldsymbol{\Gamma}_{K-1} &= \mathbf{D}_K \boldsymbol{\Gamma}_K, & \|\boldsymbol{\Gamma}_K\|_{0,\infty}^s &\leq \lambda_K.\end{aligned}$$

- $\mathcal{M}_\lambda$  the set of signals satisfying the ML-CSC assumption.
- If  $\mathbf{X}(\boldsymbol{\Gamma}_i) \in \mathcal{M}_\lambda$ , then

$$\mathbf{X}(\boldsymbol{\Gamma}_i) = \mathbf{D}_1 \mathbf{D}_2 \dots \mathbf{D}_K \boldsymbol{\Gamma}_K = \underbrace{\mathbf{D}^{(K)}}_{\text{Effective Dictionary}} \boldsymbol{\Gamma}_K$$

# A New Problem Formulation

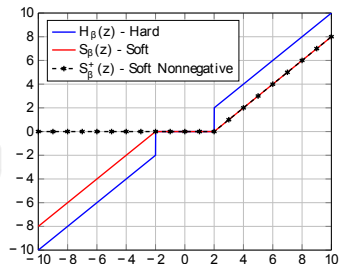
Say we get  $\mathbf{Y} = \mathbf{X}(\boldsymbol{\Gamma}_i) + \mathbf{E}$ , how to (deep) sparse code?

## Deep Coding Problem

$$\begin{aligned}
 (\text{DCP}_{\lambda}^{\mathcal{E}}) : \quad & \text{find } \{\boldsymbol{\Gamma}_i\}_{i=1}^K \quad \text{s.t.} \quad \|\mathbf{Y} - \mathbf{D}_1 \boldsymbol{\Gamma}_1\|_2^2 \leq \mathcal{E}_0, & \|\boldsymbol{\Gamma}_1\|_{0,\infty}^s \leq \lambda_1 \\
 & \|\boldsymbol{\Gamma}_1 - \mathbf{D}_2 \boldsymbol{\Gamma}_2\|_2^2 \leq \mathcal{E}_1, & \|\boldsymbol{\Gamma}_2\|_{0,\infty}^s \leq \lambda_2 \\
 & \vdots & \vdots \\
 & \|\boldsymbol{\Gamma}_{K-1} - \mathbf{D}_K \boldsymbol{\Gamma}_K\|_2^2 \leq \mathcal{E}_{K-1}, & \|\boldsymbol{\Gamma}_K\|_{0,\infty}^s \leq \lambda_K,
 \end{aligned}$$

Given  $\mathbf{Y} = \mathbf{D}_1 \boldsymbol{\Gamma}_1 + \mathbf{E}$ , how to find  $\boldsymbol{\Gamma}_1$ ?

Simplest alternative:  $\hat{\boldsymbol{\Gamma}}_1 = \mathcal{P}_{\beta_1}(\mathbf{D}_1^T \mathbf{Y})$



Solving the  $\text{DCP}_\lambda^\mathcal{E}$ 

## Layered Thresholding (LT) algorithm

$$\hat{\mathbf{r}}_2 = \mathcal{P}_{\beta_2}(\mathbf{D}_2^T \hat{\mathbf{r}}_1 = \mathcal{P}_{\beta_1}(\mathbf{D}_1^T \mathbf{Y}))$$

Written differently,

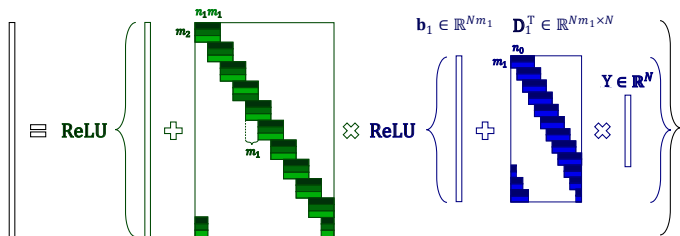
$$\hat{\mathbf{r}}_2 = \text{ReLU}(\mathbf{D}_2^T \text{ReLU}(\mathbf{D}_1^T \mathbf{Y} + \mathbf{b}_1) + \mathbf{b}_2)$$

Forward Pass of CNN

$$\hat{\mathbf{r}}_2 \in \mathbb{R}^{Nm_2}$$

$$\mathbf{b}_2 \in \mathbb{R}^{Nm_2}$$

$$\mathbf{D}_2^T \in \mathbb{R}^{Nm_2 \times Nm_1}$$



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

# Theoretical Claims for the $\text{DCP}_\lambda^\mathcal{E}$

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

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

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

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

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

$$\|\hat{\mathbf{\Gamma}}_i - \mathbf{\Gamma}_i\|_{2,\infty}^p \leq \sqrt{\|\mathbf{\Gamma}_i\|_{0,\infty}^p} (\epsilon_L^{i-1} + \mu(\mathbf{D}_i) (\|\mathbf{\Gamma}_i\|_{0,\infty}^s - 1) |\mathbf{\Gamma}_i^{\max}| + \beta_i)$$

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

- Even in the noiseless case, it is incapable of recovering the solution to the  $\text{DCP}_\lambda$ .
- Its success depends on the ratio  $|\mathbf{\Gamma}_i^{\min}|/|\mathbf{\Gamma}_i^{\max}|$

# Multi-Layer Basis Pursuit

$$(\text{DCP}_{\lambda}^{\mathcal{E}}) : \quad \text{find} \quad \{\mathbf{\Gamma}_i\}_{i=1}^K \quad \text{s.t.} \quad \begin{aligned} \|\mathbf{Y} - \mathbf{D}_1 \mathbf{\Gamma}_1\|_2^2 &\leq \mathcal{E}_0, & \|\mathbf{\Gamma}_1\|_{0,\infty}^{\mathbf{s}} &\leq \lambda_1 \\ \|\mathbf{\Gamma}_1 - \mathbf{D}_2 \mathbf{\Gamma}_2\|_2^2 &\leq \mathcal{E}_1, & \|\mathbf{\Gamma}_2\|_{0,\infty}^{\mathbf{s}} &\leq \lambda_2 \end{aligned}$$

## Layered BP

$$\hat{\mathbf{\Gamma}}_i = \arg \min_{\mathbf{\Gamma}_i} \quad \frac{1}{2} \|\hat{\mathbf{\Gamma}}_{i-1} - \mathbf{D}_i \mathbf{\Gamma}_i\|_2^2 + \zeta_i \|\mathbf{\Gamma}_i\|_1$$

## Stability

If  $\{\mathbf{\Gamma}_i\}_{i=1}^K$  satisfy  $\|\mathbf{\Gamma}_i\|_{0,\infty} \leq \frac{1}{3} \left(1 + \frac{1}{\mu(\mathbf{D}_i)}\right)$ , then

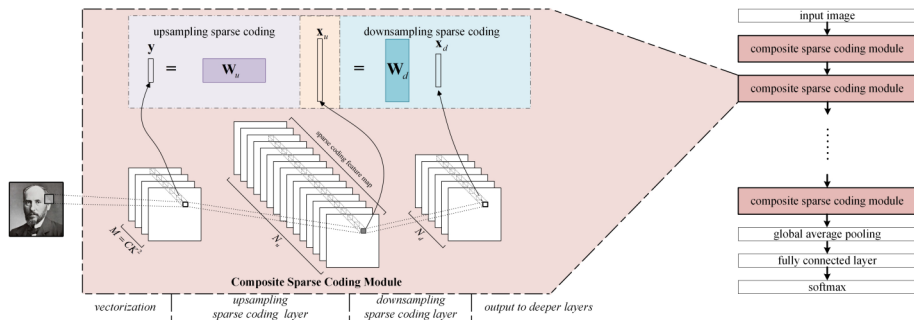
- $\text{Supp}\{\hat{\mathbf{\Gamma}}_i\} \subseteq \text{Supp}\{\mathbf{\Gamma}_i\}$
- $\|\hat{\mathbf{\Gamma}}_i - \mathbf{\Gamma}_i\|_{2,\infty}^p \leq 7.5^i \|\mathbf{E}\|_{2,\infty}^p \prod_{j=1}^i \sqrt{\|\mathbf{\Gamma}_j\|_{0,\infty}^p}$
- Every sufficiently large entry will be recovered

- ✓ Exact recovery in noiseless case
- ✓ Independent of the signal contrast
- ✗ Bound increase with depth

## Multi-Layer Basis Pursuit

$$\text{Solve } \min_{\mathbf{\Gamma}_1} \frac{1}{2} \|\mathbf{Y} - \mathbf{D}_1 \mathbf{\Gamma}_1\|_2^2 + \lambda_1 \|\mathbf{\Gamma}_1\|_1 \rightarrow \hat{\mathbf{r}}_1$$

$$\text{Solve } \min_{\mathbf{\Gamma}_2} \frac{1}{2} \|\hat{\mathbf{r}}_1 - \mathbf{D}_2 \mathbf{\Gamma}_2\|_2^2 + \lambda_2 \|\mathbf{\Gamma}_2\|_1 \rightarrow \hat{\mathbf{r}}_2$$

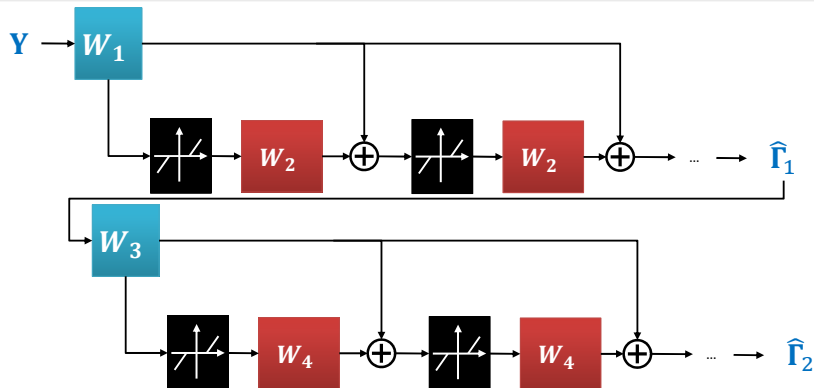


[Sun et al, *Supervised Deep Sparse Coding Networks*, '17]

## Multi-Layer Basis Pursuit

$$\text{Solve } \min_{\Gamma_1} \frac{1}{2} \|\mathbf{Y} - \mathbf{D}_1 \Gamma_1\|_2^2 + \lambda_1 \|\Gamma_1\|_1$$

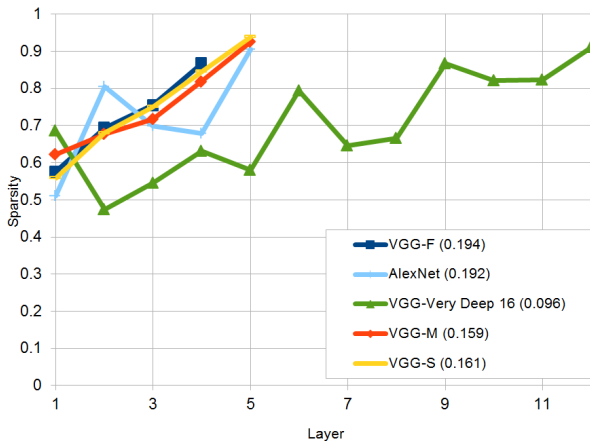
$$\text{with } \Gamma_1^k \leftarrow \mathcal{S}_{\lambda_1/c_1} \left( \Gamma_1^{k-1} + \frac{1}{c_1} \mathbf{D}_1^T (\mathbf{Y} - \mathbf{D}_1 \Gamma_1^{k-1}) \right)$$



[LISTA Networks, Gregor &amp; LeCun]

# Looking into the Networks

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



## 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  $\mathbf{Y} = \mathbf{X}(\mathbf{\Gamma}_i) + \mathbf{E}$ ,  $\mathbf{X} \in \mathcal{M}_\lambda$ .

## ML-CSC Projection ( $\mathcal{P}_{\mathcal{M}_\lambda}$ )

Given  $\mathbf{Y}$  and convolutional dictionaries  $\{\mathbf{D}_i\}_{i=1}^K$ ,

$$(\mathcal{P}_{\mathcal{M}_\lambda}) : \min_{\{\mathbf{\Gamma}_i\}} \|\mathbf{Y} - \mathbf{X}(\mathbf{\Gamma}_i)\|_2 \quad \text{s.t.} \quad \mathbf{X}(\mathbf{\Gamma}_i) \in \mathcal{M}_\lambda.$$

- Unlike the  $\text{DCP}_\lambda^\mathcal{E}$ , the solution  $\hat{\mathbf{X}} \in \mathcal{M}_\lambda$ :

$$\hat{\mathbf{X}} = \mathbf{D}_1 \hat{\mathbf{\Gamma}}_1 = \mathbf{D}_1 \mathbf{D}_2 \hat{\mathbf{\Gamma}}_2 = \cdots = \mathbf{D}^{(i)} \hat{\mathbf{\Gamma}}_i$$

- A solution to the  $\text{DCP}_\lambda^\mathcal{E}$ , provides  $\hat{\mathbf{\Gamma}}_{i-1} \neq \mathbf{D}_i \hat{\mathbf{\Gamma}}_i$

# Stability of the $\mathcal{P}_{\mathcal{M}_\lambda}$ problem

## Theorem

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

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

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

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

# Pursuit Algorithms

- How to solve  $\mathcal{P}_{\mathcal{M}_\lambda}$ ?
- How to seek for  $\{\hat{\Gamma}_i\}$  while assuring  $\mathbf{X}(\Gamma_i) \in \mathcal{M}_\lambda$ ?

## ML-CSC Pursuit

- Global Pursuit:

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

- Finding the inner representations:

```

for  $j = K, \dots, 1$  do
  |  $\hat{\Gamma}_{j-1} \leftarrow \mathbf{D}_j \hat{\Gamma}_j$ 
end
  
```

# Stability of Pursuit Algorithms

## Theorem: Stability of the Pursuit - $\ell_1$ case

$\mathbf{Y} = \mathbf{X}(\mathbf{\Gamma}_i) + \mathbf{E}$ ,  $\mathbf{X} \in \mathcal{M}_\lambda$ ,  $\|\mathbf{E}\|_{2,\infty} \leq \epsilon_0$ .  $\|\mathbf{\Gamma}_i\|_{0,\infty} = \lambda_i < \frac{1}{2} \left(1 + \frac{1}{\mu(\mathbf{D}_i)}\right)$ ,  
 $i = 1, \dots, K-1$  and  $\|\mathbf{\Gamma}_K\|_{0,\infty} = \lambda_i \leq \frac{1}{3} \left(1 + \frac{1}{\mu(\mathbf{D}^{(K)})}\right)$ .  $\{\mathbf{\Gamma}_i\}$  satisfy the N.V.S. for  $\mathbf{D}_i$ .

Let

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

$$\hat{\mathbf{\Gamma}}_{i-1} \leftarrow \mathbf{D}_i \hat{\mathbf{\Gamma}}_i, \quad i = K, \dots, 1$$

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

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

**Non Vanishing Support property**  $\mathbf{\Gamma}$  will not cause atoms to be combined such that their supports cancel each other.

# Stability of Pursuit Algorithms

## Theorem: Stability of the Pursuit - $\ell_0$ case

Suppose  $\mathbf{Y} = \mathbf{X}(\mathbf{\Gamma}_i) + \mathbf{E}$ ,  $\|\mathbf{Y} - \mathbf{X}\|_2 \leq \mathcal{E}_0$ , and  $\epsilon_0 = \|\mathbf{E}\|_{2,\infty}^P$ . Let  $\mathbf{\Gamma}_i$  satisfy the N.V.S. property for the respective dictionaries  $\mathbf{D}_i$ , with  $\|\mathbf{\Gamma}_i\|_{0,\infty}^s = \lambda_i < \frac{1}{2} \left(1 + \frac{1}{\mu(\mathbf{D}_i)}\right)$ , for  $1 \leq i \leq K$ , and  $\|\mathbf{\Gamma}_K\|_{0,\infty}^s < \frac{1}{2} \left(1 + \frac{1}{\mu(\mathbf{D}^{(K)})}\right) - \frac{1}{\mu(\mathbf{D}^{(K)})} \cdot \frac{\epsilon_0}{|\mathbf{\Gamma}_K^{min}|}$ , and

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

$$\hat{\mathbf{\Gamma}}_i \leftarrow \mathbf{D}_{i+1}\hat{\mathbf{\Gamma}}_{i+1}, \quad i = K, \dots, 1$$

Then

- ①  $Supp(\hat{\mathbf{\Gamma}}_i) \subseteq Supp(\mathbf{\Gamma}_i)$ ,
- ②  $\|\hat{\mathbf{\Gamma}}_i - \mathbf{\Gamma}_i\|_2^2 \leq \frac{\mathcal{E}_0^2}{1 - \mu(\mathbf{D}^{(K)}) (\|\mathbf{\Gamma}_K\|_{0,\infty}^s - 1)} \left(\frac{3}{2}\right)^{K-i}$ .

# What about the Dictionaries?

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

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

Example:

- $\mathbf{D}_i$  are Random Dictionaries, i.e.,  $\mathbf{d}_K^j = \mathbf{R}_j^T \mathbf{v}$ ,  $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma_i^2 \mathbf{I})$
- One can construct  $\Gamma_K$  with  $\|\Gamma_K\|_{0,\infty} \leq 2$  such that  $\Pr(\Gamma_{i-1}^K = 0) = 0 \rightarrow \text{dense!}$

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

If one seeks for  $\{\Gamma_i\}$ , one must seek also for  $\{\mathbf{D}_i\}$  that would allow for that decomposition.



## How to Learn?

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

Problematic:

- The constraints on  $\mathbf{\Gamma}_i$  are coupled
- $\mathbf{\Gamma}_i$  depends on  $\{\mathbf{D}_j\}_{j=i+1}^K$ .

## Sparsity Proxies

$$\mathbf{\Gamma}_{K-1} = \mathbf{D}_K \mathbf{\Gamma}_K. \quad \Rightarrow \|\mathbf{\Gamma}_{K-1}\|_{0,\infty}^s \leq c_K \|\mathbf{D}_K\|_0 \|\mathbf{\Gamma}_K\|_{0,\infty}^s$$

$$\|\mathbf{\Gamma}_i\|_{0,\infty}^s \leq c \prod_{j=i+1}^K \|\mathbf{D}_j\|_0 \|\mathbf{\Gamma}_K\|_{0,\infty}^s.$$

# MultiLayer Convolutional Dictionary Learning

## Problem formulation

$$\min_{\{\mathbf{\Gamma}_K^t\}, \{\mathbf{D}_i\}} \sum_{t=1}^T \|\mathbf{Y}^t - \mathbf{D}_1 \mathbf{D}_2 \dots \mathbf{D}_K \mathbf{\Gamma}_K^t\|_2^2 + \sum_{i=2}^K \zeta_i \|\mathbf{D}_i\|_0 \quad \text{s.t.} \quad \|\mathbf{\Gamma}_K^t\|_{0,\infty}^s \leq \lambda_K$$

## Algorithm

**Data:** Training samples  $\{\mathbf{Y}_i\}$ , initial convolutional dictionaries  $\mathbf{D}_i^0$

**for**  $t = 1, \dots, T$  **do**

    Draw  $\mathbf{Y}^t$  at random;

    Sparse Coding:  $\hat{\mathbf{\Gamma}}_K \leftarrow \arg \min_{\mathbf{\Gamma}} \|\mathbf{Y}^t - \mathbf{D}^{(K)} \mathbf{\Gamma}\|_2$  s.t.  $\|\mathbf{\Gamma}\|_{0,\infty}^s \leq \lambda_K$  (IHT/FISTA);

    Update Dictionaries:

**for**  $k = K, \dots, 1$  **do**

$\mathbf{D}_k \leftarrow \arg \min_{\mathbf{D}_k} \|\mathbf{Y}^t - \mathbf{D}_1 \dots \mathbf{D}_k \dots \mathbf{D}_K \mathbf{\Gamma}_K\|_2 + \zeta_k \|\mathbf{D}_k\|_0$  (PGD);

**end**

**end**

# Related work

## Dictionary Learning

**Chasing-Butterflies** :  $\min \|\mathbf{Y} - \prod_{j=1}^{L+1} \mathbf{S}_j\|_2^2, \mathbf{S}_j \text{ sparse}$  [L 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

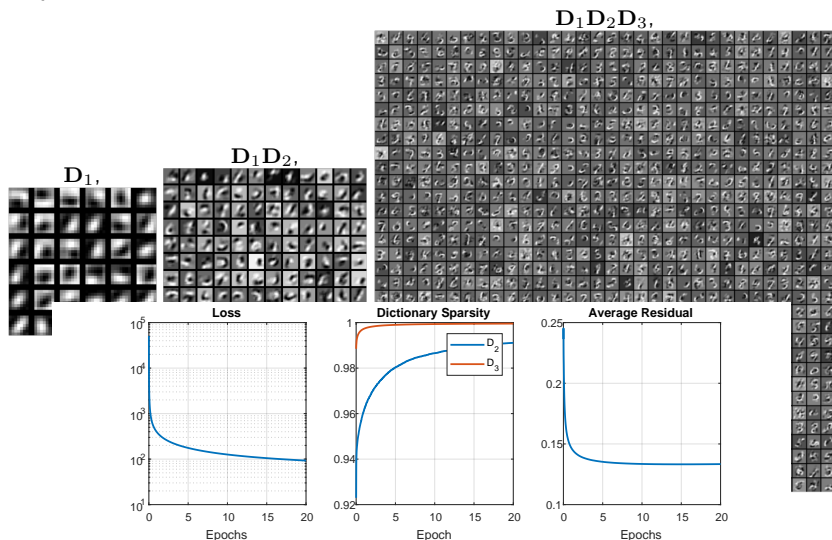
**Sparse AutoEncoders** : imposing sparse-enforcing loss in hidden layer [Ng, 2011]

**K-Sparse AutoEncoders** :  $\min_{\mathbf{W}, \mathbf{b}, \mathbf{b}'} \|\mathbf{Y} - (\mathbf{W} \mathbf{H}_k (\mathbf{W}^T \mathbf{X} + \mathbf{b}) + \mathbf{b}')\|_2$  [Makhzani, 2014]

**Winner-Take-All AutoEncoders** : “Spatial” sparsity + “life-time” sparsity [Makhzani, 2015]

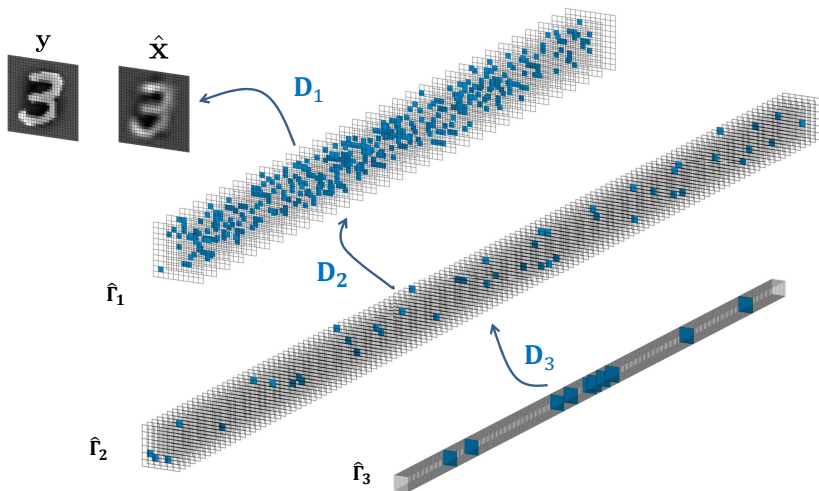
# Learning an MNIST model

Multi-Layer Convolutional Dictionaries:



# Learning an MNIST model

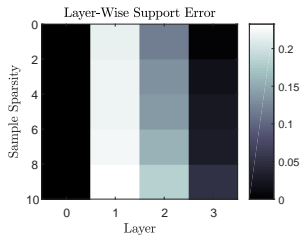
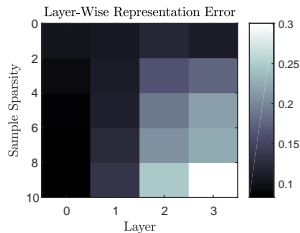
Multi-Layer Convolutional Decomposition:



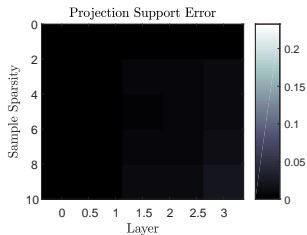
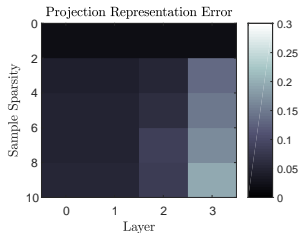
# Learning an MNIST model

## Sparse Recovery (Synthetic Data):

Layered-BP

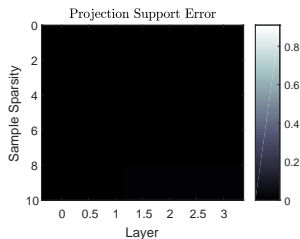
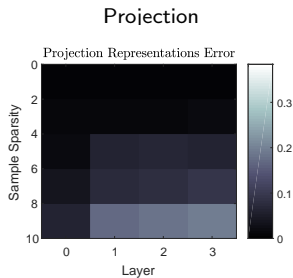
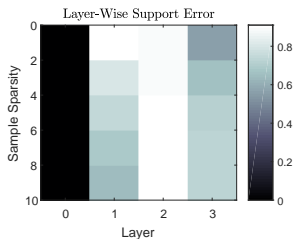
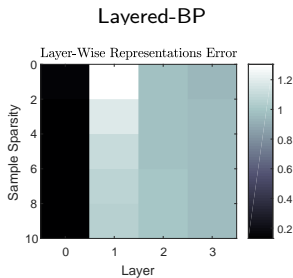


Projection

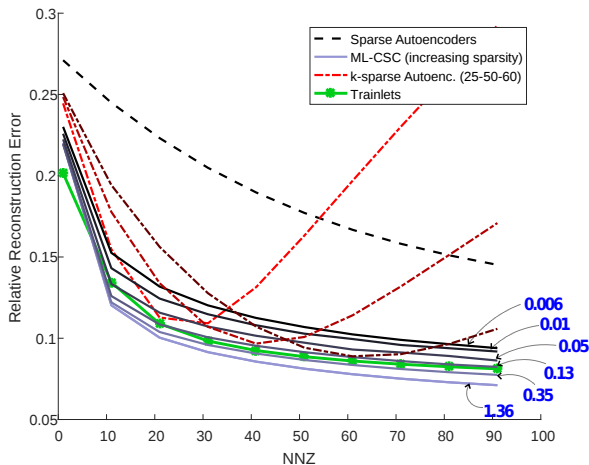


# Learning an MNIST model

## Sparse Recovery (MNIST Data):



## M-term Approximation



# Classification

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

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

# Contents

- 1 Modeling
- 2 Sparse Modeling
- 3 Convolutional Sparse Modeling
- 4 Multi-Layer Convolutional Sparse Coding
- 5 Conclusion

# 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**