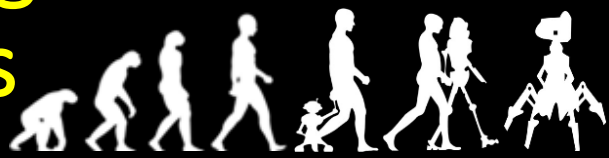
Signal Modeling:

From Convolutional Sparse Coding to Convolutional Neural Networks



Vardan Papyan

The Computer Science Department

Technion – Israel Institute of technology

Haifa 32000, Israel

Joint work with

Jeremias Sulam

Yaniv Romano

Prof. Michael Elad

Technion Israel Institute of Technology

erc

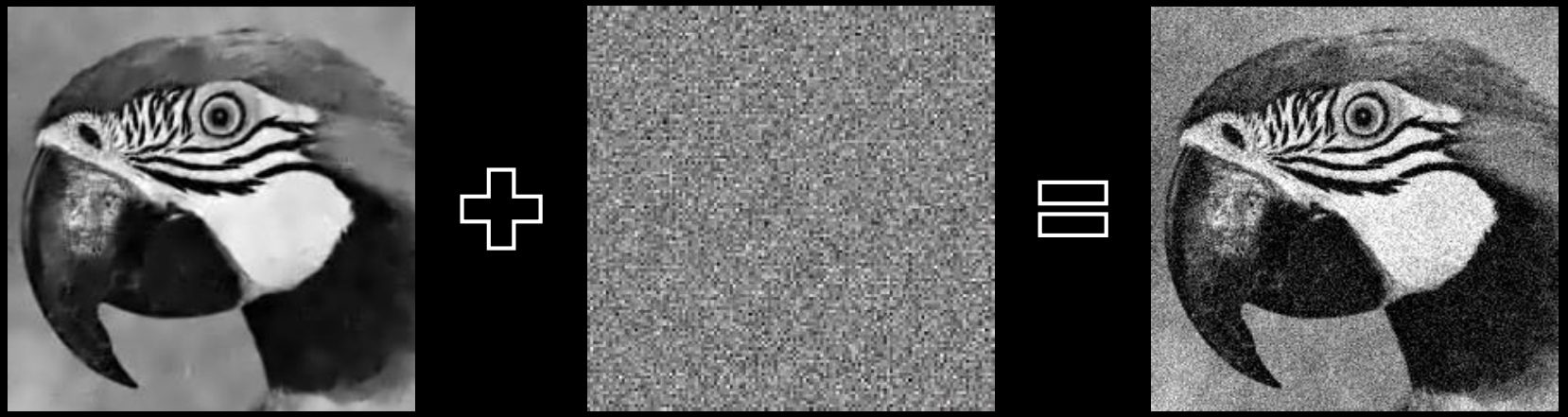
The research leading to these results has been received funding from the European union's Seventh Framework Program (FP/2007-2013) ERC grant Agreement ERC-SPARSE- 320649

Part I

Motivation and Background

Our Starting Point: Image Denoising

Original Image $\mathbf{X}$ + White Gaussian Noise $\mathbf{E}$ = Noisy Image $\mathbf{Y}$



Many image denoising algorithms can be cast as the minimization of an energy function of the form

$$\frac{1}{2}\|\mathbf{X}-\mathbf{Y}\|_2^2 + \mathrm{G}(\mathbf{X})$$

Relation to measurements + Prior or regularization

Leading Image Denoising Methods…

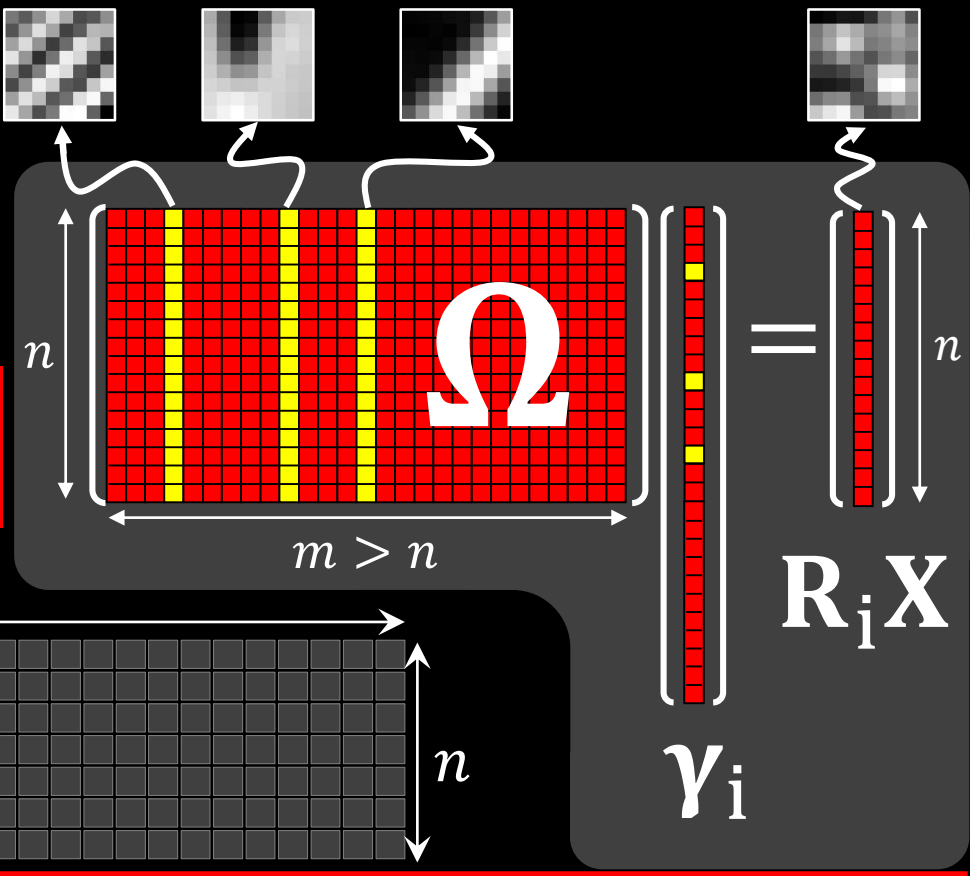
are built upon powerful patch-based local models:

- K-SVD: sparse representation modeling of image patches
  [Elad & Aharon, ‘06]
- BM3D: combines sparsity and self-similarity
  [Dabov, Foi, Katkovnik & Egiazarian ‘07]
- EPLL: uses GMM of the image patches
  [Zoran & Weiss ‘11]
- CSR: clustering and sparsity on patches
  [Dong, Li, Lei & Shi ‘11]
- MLP: multi-layer perceptron
  [Burger, Schuler & Harmeling ‘12]
- NCSR: non-local sparsity with centralized coefficients
  [Dong, Zhang, Shi & Li ‘13]
- WNNM: weighted nuclear norm of image patches
  [Gu, Zhang, Zuo & Feng ‘14]
- SSC–GSM: nonlocal sparsity with a GSM coefficient model
  [Dong, Shi, Ma & Li ‘15]

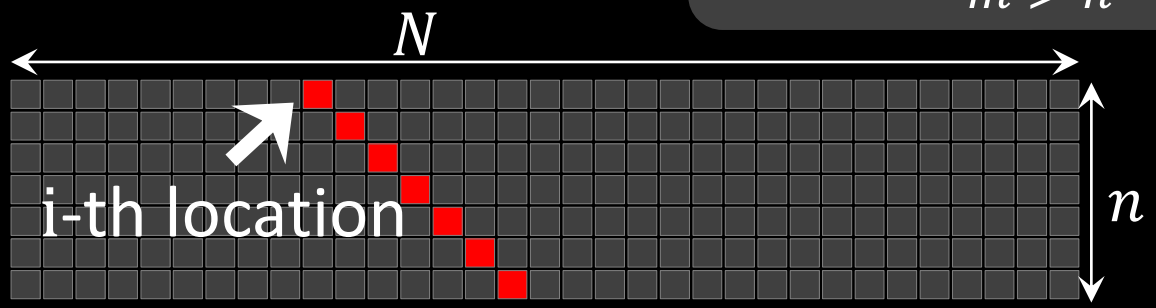
The Sparse-Land Model

- Assumes every patch a linear combination of a few columns, called atoms, from a matrix that is termed a dictionary.
- The operator $\mathbf{R}_i$ extracts the i-th n-dimensional patch from $\mathbf{X} \in \mathbb{R}^N$.
- Sparse coding:

$$(\mathbf{P}_0): \quad \min_{\boldsymbol{\gamma}_i} \|\boldsymbol{\gamma}_i\|_0 \ \text{s.t.} \ \mathbf{R}_i\mathbf{X} = \boldsymbol{\Omega}\boldsymbol{\gamma}_i$$



$\mathbf{R}_i =$



* $\mathbf{R}_i$ for 1D signals

Patch Denoising

Given a noisy patch $\mathbf{R}_i\mathbf{Y}$, solve

$$(\mathbf{P}_0^{\epsilon}):\quad \hat{\boldsymbol{\gamma}}_i = \underset{\boldsymbol{\gamma}_i}{\operatorname{argmin}} \ \|\boldsymbol{\gamma}_i\|_0 \ \text{ s.t. } \ \|\mathbf{R}_i\mathbf{Y} - \boldsymbol{\Omega}\boldsymbol{\gamma}_i\|_2 \le \epsilon$$

➡ Clean patch: $\boldsymbol{\Omega}\hat{\boldsymbol{\gamma}}_i$

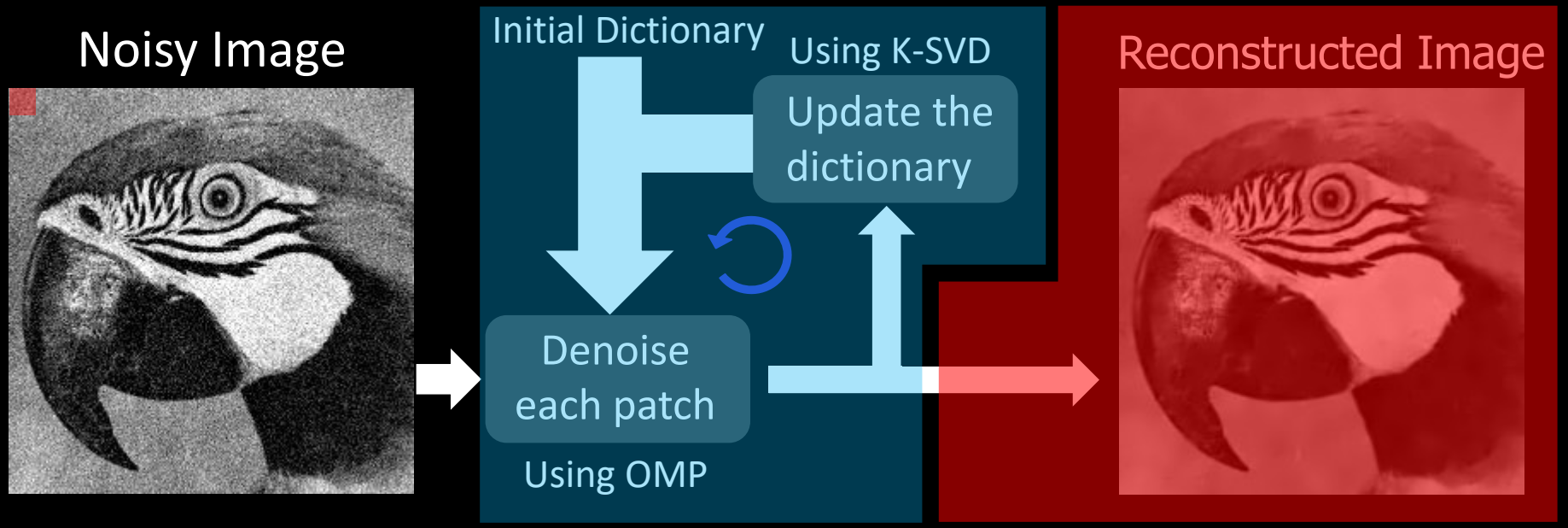
$(\mathbf{P}_0)$ and $(\mathbf{P}_0^{\epsilon})$ are hard to solve

Greedy methods such as Orthogonal Matching Pursuit (OMP) or Thresholding

Convex relaxations such as Basis Pursuit (BP)

$$(\mathbf{P}_1^{\epsilon}):\quad \min_{\boldsymbol{\gamma}_i} \ \|\boldsymbol{\gamma}_i\|_1 + \xi\|\mathbf{R}_i\mathbf{Y} - \boldsymbol{\Omega}\boldsymbol{\gamma}_i\|_2^2$$

Consider this Algorithm [Elad & Aharon, '06]



$$\hat{\mathbf{X}} = \underset{\mathbf{X},\{\boldsymbol{\gamma}_i\}_i,\boldsymbol{\Omega}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{X}-\mathbf{Y}\|_2^2 + \lambda \sum_i \|\boldsymbol{\gamma}_i\|_0 + \mu\|\mathbf{R}_i\mathbf{X}-\boldsymbol{\Omega}\boldsymbol{\gamma}_i\|_2^2$$

Relation to measurements

Prior or regularization

What is Missing?

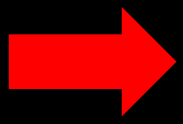
- Over the years, many researchers kept revisiting this algorithm and the line of thinking behind it, with a clear feeling that the final word has not been said, and that key features are still lacking.
- What is missing? Here is what our group thought of…
  - A multi-scale treatment [Ophir, Lustig & Elad '11] [Sulam, Ophir & Elad '14] [Papyan & Elad '15]
  - Exploiting self-similarities [Ram & Elad '13] [Romano, Protter & Elad '14]
  - Pushing to better agreement on the overlaps [Romano & Elad '13] [Romano & Elad '15]
  - Enforcing the local model on the final patches (EPLL) [Sulam & Elad '15]

Beyond all these, a key part that is missing is a **theoretical** backbone for the local model as a way to characterize the global unknown image

Missing Theoretical Backbone?

- The core global-local model assumption on $\mathbf{X}$:

$$\forall i \quad \mathbf{R}_i\mathbf{X} = \boldsymbol{\Omega}\boldsymbol{\gamma}_i \quad \text{where} \quad \|\boldsymbol{\gamma}_i\|_0 \leq k$$



Every patch in the unknown signal is expected to have a sparse representation w.r.t. the same dictionary $\boldsymbol{\Omega}$

- Questions to consider:
  I. Who are those signals belonging to this model? Do they exist?
  II. Under which conditions on $\boldsymbol{\Omega}$ would this model be feasible?
  III. How does one sample from this model?
  IV. How should we perform pursuit properly (and locally) under this model?
  V. How should we learn $\boldsymbol{\Omega}$ if this is indeed the model?

In this Talk

Limitations of patch averaging

Convolutional Sparse Coding (CSC) model

Theoretical study of CSC

Multi-Layer Convolutional Sparse Coding (ML-CSC)

Convolutional neural networks (CNN)

Fresh view of CNN through the eyes of sparsity

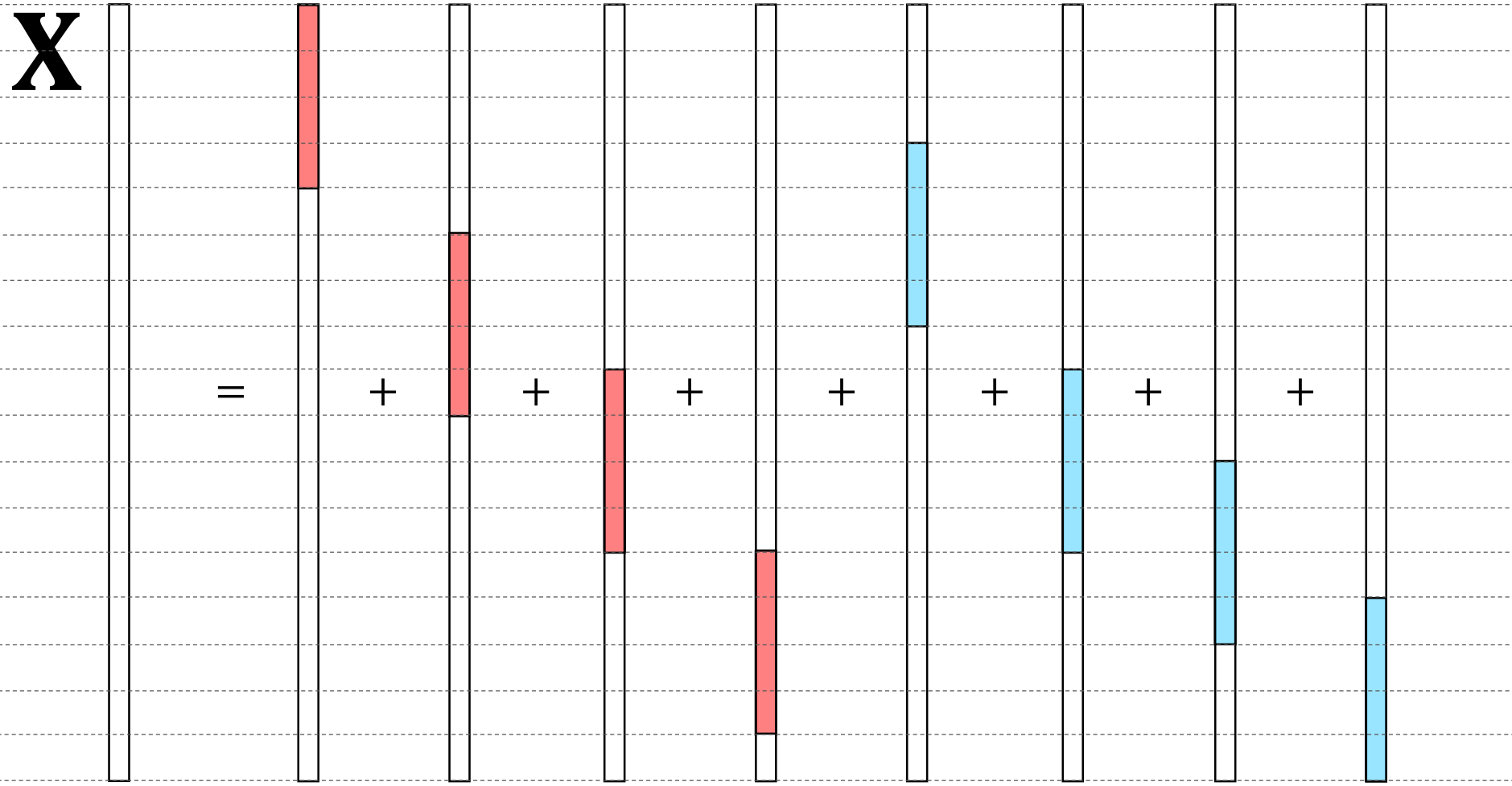
Part II
Convolutional Sparse Coding

Working Locally Thinking Globally-**Part I**:
Theoretical Guarantees for Convolutional Sparse Coding

Working Locally Thinking Globally-**Part II**:
Stability and Algorithms for Convolutional Sparse Coding

Vardan Papyan, Jeremias Sulam and Michael Elad

Convolutional Sparsity Assumes…

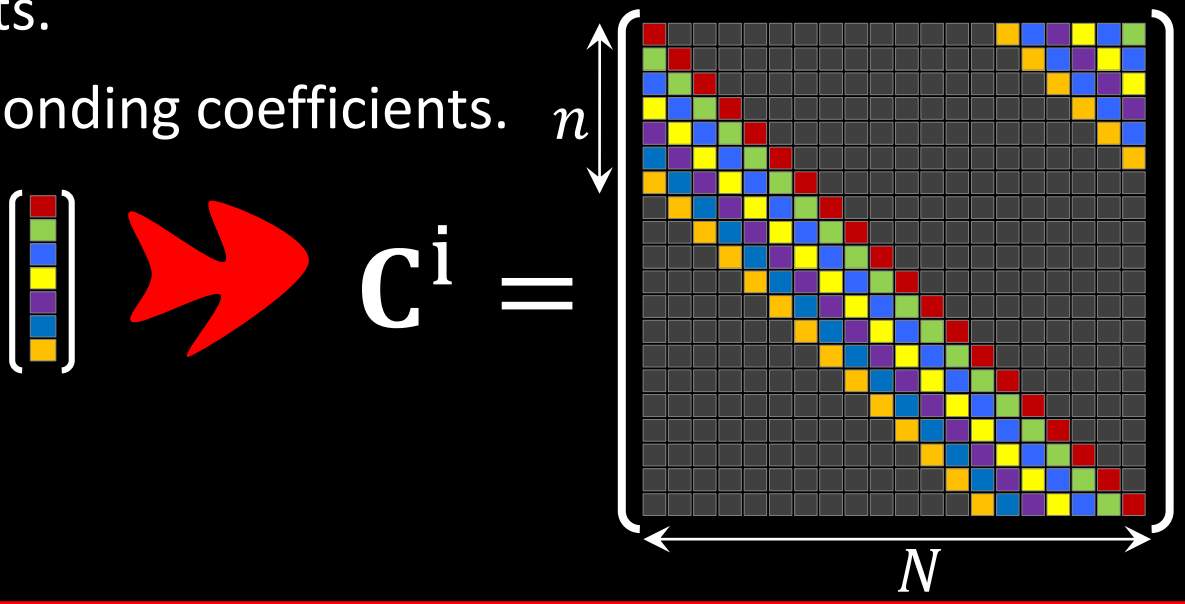


Convolutional Sparse Representation

- Formally, consider the following global sparsity-based model

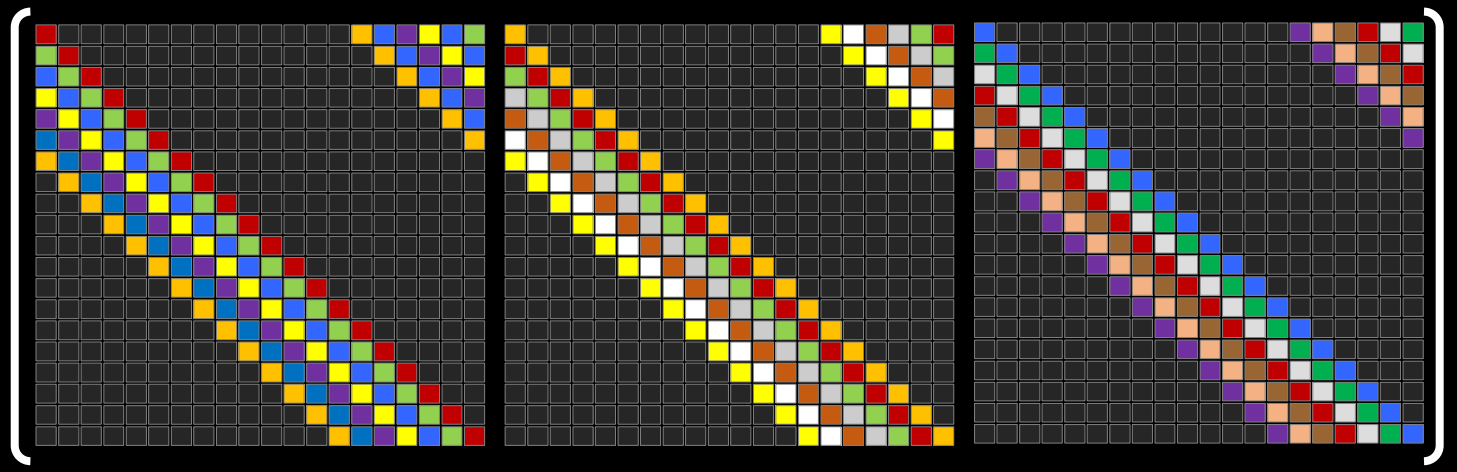
$$\mathbf{X} = \sum_{i=1}^{m} \mathbf{C}^i \boldsymbol{\Gamma}^i = \mathbf{D}\boldsymbol{\Gamma}$$

- $\mathbf{C}^i \in \mathbb{R}^{N \times N}$ is a banded and Circulant matrix containing a single atom with all of its shifts.
- $\boldsymbol{\Gamma}^i \in \mathbb{R}^N$ are its corresponding coefficients.

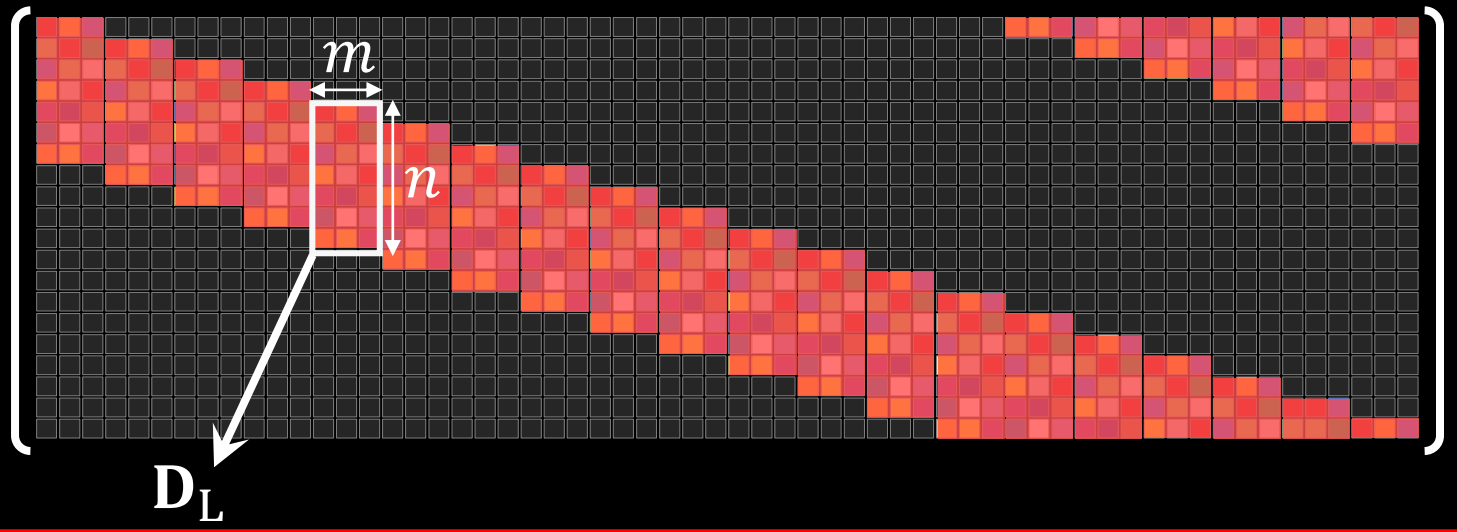


Two Interpretations

$[\mathbf{c}^1\ \mathbf{c}^2\ \mathbf{c}^3] =$



$\mathbf{D} =$



- This model has been used in the past [Lewicki & Sejnowski '99] [Hashimoto & Kurata, '00]
- Most works have focused on solving *efficiently* its associated pursuit, called **convolutional sparse coding**, using the BP algorithm.

$$(\mathbf{P}_1^{\epsilon}): \quad \min_{\boldsymbol{\Gamma}} \quad \|\boldsymbol{\Gamma}\|_1 + \xi\|\mathbf{Y} - \mathbf{D}\boldsymbol{\Gamma}\|_2^2$$

Convolutional dictionary

- Several applications were demonstrated:
  - Inpainting [Heide, Heidrich & Wetzstein '15]
  - Super-resolution [Gu, Zuo, Xie, Meng, Feng & Zhang '15]
  - Pattern detection in images and the analysis of instruments in music signals [Mørup, Schmidt & Hansen '08]
- However, little is known regrading its theoretical aspects.

Classical Sparse Theory (Noiseless)

Mutual Coherence: $\mu(\mathbf{D}) = \max_{i \neq j} |d_i^T d_j|$

[Donoho & Elad '03]

For a signal $\mathbf{X} = \mathbf{D\Gamma}$, if $\|\mathbf{\Gamma}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$ then this solution is necessarily the sparsest.

[Donoho & Elad '03]

The OMP and BP are guaranteed to recover the true sparse code assuming that $\|\mathbf{\Gamma}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$.

[Tropp '04]
[Donoho & Elad '03]

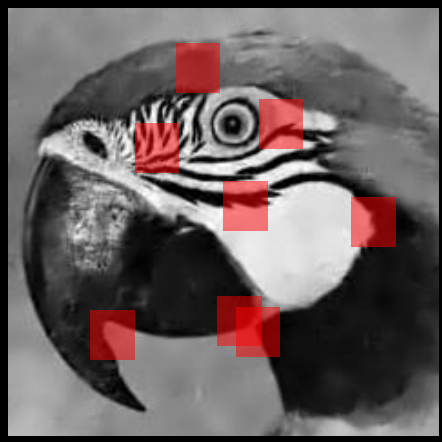
The Need for a Theoretical Study

- Assuming that $m = 2$ and $n = 64$ we have that $\mu(\mathbf{D}) \geq 0.063$.
- As a result, uniqueness and success of pursuits is guaranteed as long as

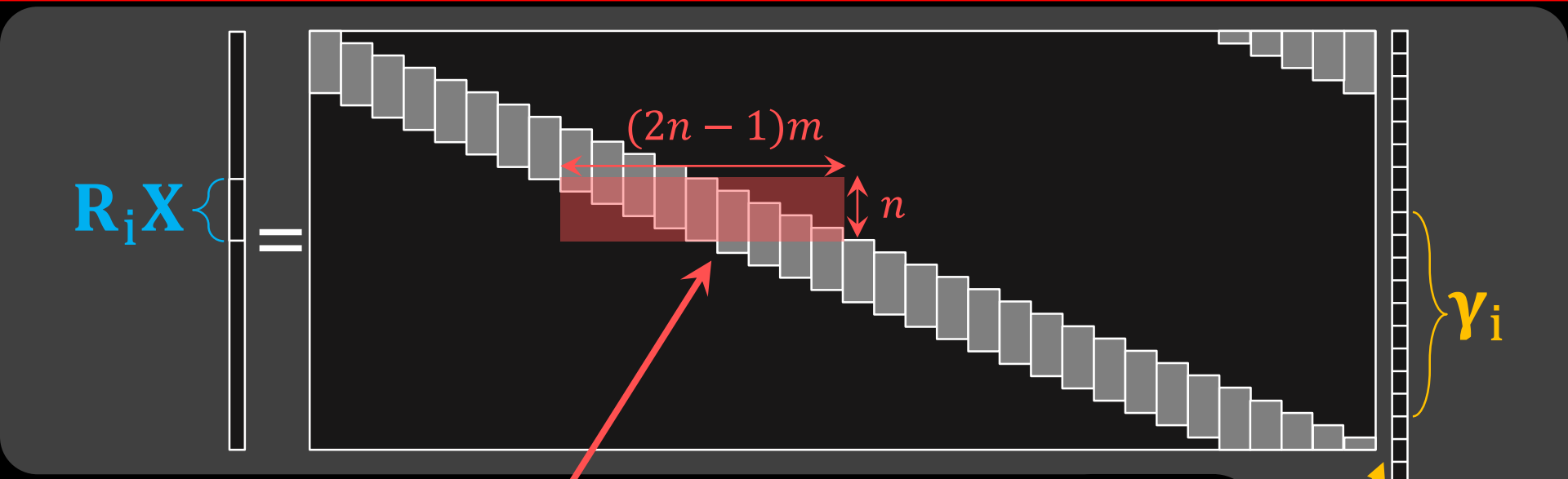
$$\|\boldsymbol{\Gamma}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right) \leq \frac{1}{2}\left(1 + \frac{1}{0.063}\right) \approx 8$$

- This is a very pessimistic result!

Less than 8 non-zeros GLOBALLY!!!



The Local Representation



stripe-dictionary

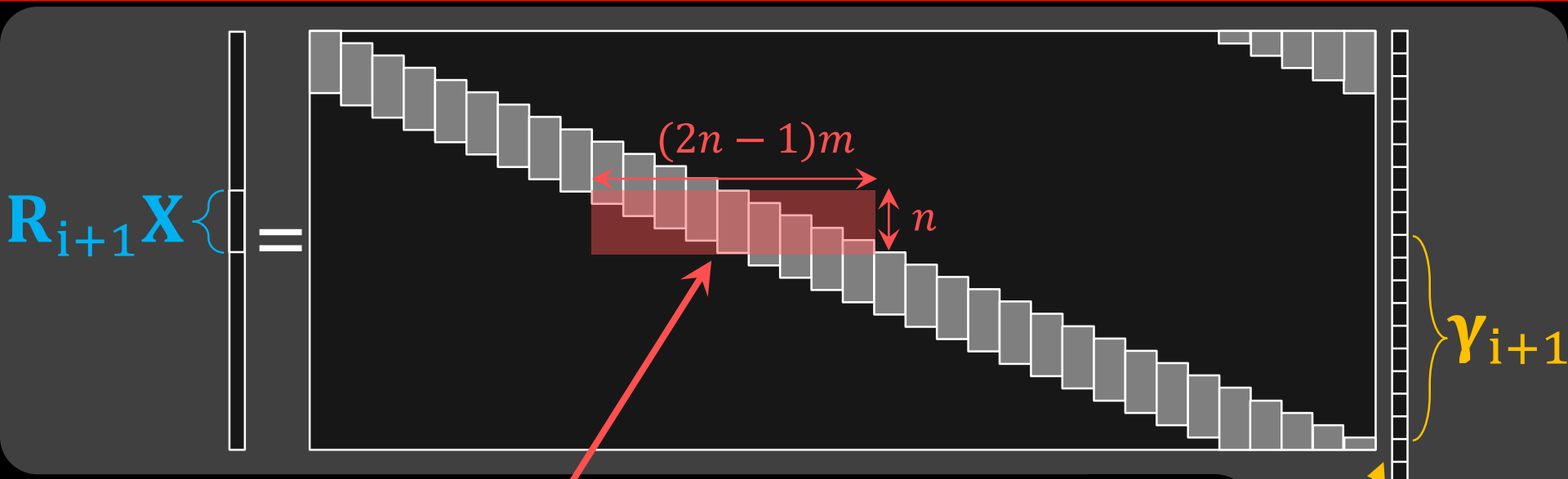
stripe vector

$$\mathbf{X} = \mathbf{D}\boldsymbol{\Gamma}$$

$$\mathbf{R}_i\mathbf{X} = \boldsymbol{\Omega}\boldsymbol{\gamma}_i$$

Adjacent representations overlap, as they skip by m items as we sweep through the patches of $\mathbf{X}$

The Local Representation



stripe-dictionary

stripe vector

$$\mathbf{X} = \mathbf{D}\boldsymbol{\Gamma}$$

$$\mathbf{R}_i\mathbf{X} = \boldsymbol{\Omega}\boldsymbol{\gamma}_i$$

Adjacent representations overlap, as they skip by m items as we sweep through the patches of $\mathbf{X}$

Inherent Positive Properties

✓A clear global model with a shift invariant local prior.

- Every patch has a sparse representation w.r.t. to a local dictionary $\boldsymbol{\Omega}$.

✓No disagreement on the patch overlaps.

✓Related to the current common practice of patch averaging.

- The signal can be written as

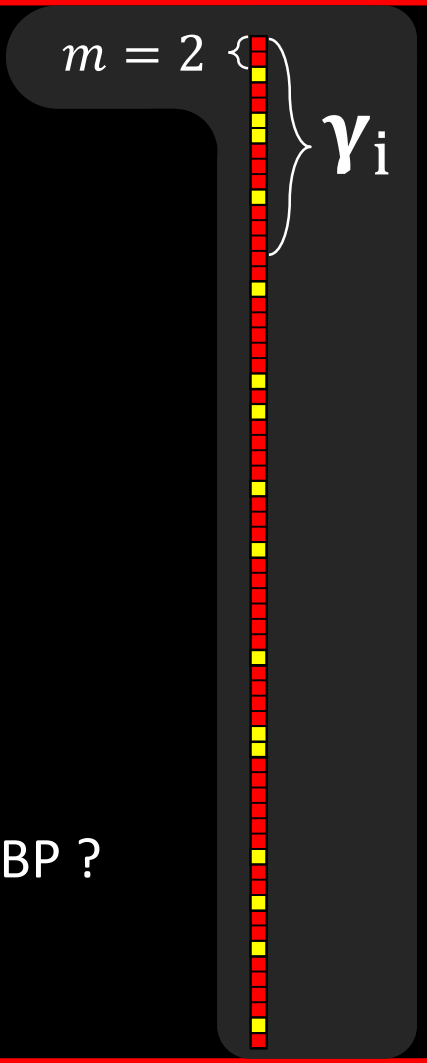
$$\mathbf{X} = \mathbf{D}\boldsymbol{\Gamma} = \frac{1}{n}\sum_{\mathrm{i}} \mathbf{R}_{\mathrm{i}}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\gamma}_{\mathrm{i}}$$

- $\mathbf{R}_{\mathrm{i}}^{\mathrm{T}}$ puts the patch $\boldsymbol{\Omega}\boldsymbol{\gamma}_{\mathrm{i}}$ in the i-th location in the N-dimensional vector.
- *The patch averaging scheme solves the sparse coding problem independently for every patch while convolutional sparse coding seeks for the representations of all the patches together*.

The $\ell_{0,\infty}$ Norm and the $\mathbf{P}_{0,\infty}$ Problem

$$\|\boldsymbol{\Gamma}\|_{0,\infty}^{s} = \max_{i} \|\boldsymbol{\gamma}_i\|_0$$

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



A global sparse vector is likely if it can represent every patch in the signal sparsely.

The Main Questions We Aim to Address:

I. Uniqueness of the solution to this problem ?

II. Guaranteed recovery of the solution via global OMP/BP ?

Uniqueness via Mutual Coherence

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

Theorem: If a solution $\boldsymbol{\Gamma}$ is found for $\left(\mathbf{P}_{0,\infty}\right)$ such that:

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

Then this is necessarily the unique **globally** optimal solution to this problem.

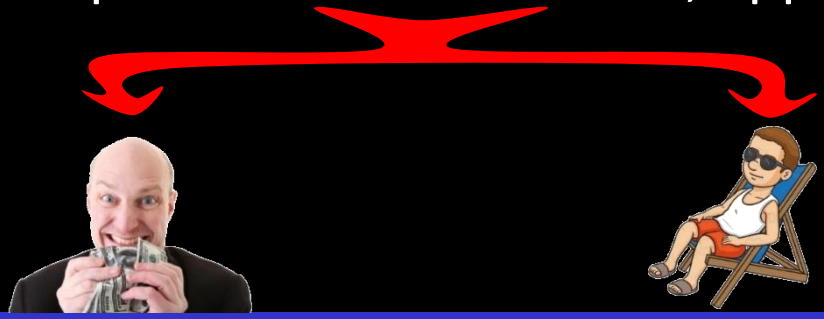
We should be excited about this result and later ones because they pose a **local** constraint for a **global** guarantee, and as such, they are far more optimistic compared to the **global** guarantees

8 non-zeros per stripe can result in $0.06 \cdot N$ non-zeros globally

Recovery Guarantees

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

Lets solve this problem via OMP or BP, applied globally



Theorem: If a solution $\boldsymbol{\Gamma}$ of $(\mathbf{P}_{0,\infty})$ satisfies:

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

Then **global** OMP and BP are guaranteed to find it.

Both OMP and BP do not assume **local** sparsity but still succeed.
One could propose algorithms that rely on this assumption

From Ideal to Noisy Signals

- So far, we have assumed an ideal signal $\mathbf{X} = \mathbf{D\Gamma}$.
- However, in practice we usually have $\mathbf{Y} = \mathbf{D\Gamma} + \mathbf{E}$ where $\mathbf{E}$ is due to noise or model deviations.
- To handle this, we redefine our problem as:

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

- **The Main Questions We Aim to Address:**
  I. Stability of the solution to this problem ?
  II. Stability of the solution obtained via global OMP/BP ?
  III. The same recovery done via local operations ?

[Candes & Tao '05]

Stability of $\left(\mathbf{P}_{0,\infty}^{\epsilon}\right)$ via Stripe-RIP

$$\left(\mathbf{P}_{0,\infty}^{\epsilon}\right): \quad \min_{\boldsymbol{\Gamma}} \ \|\boldsymbol{\Gamma}\|_{0,\infty}^{s} \ \text{s.t.} \ \|\mathbf{Y}-\mathbf{D}\boldsymbol{\Gamma}\|_2 \le \epsilon \quad \Rightarrow \quad \hat{\boldsymbol{\Gamma}}$$

Definition: $\mathbf{D}$ is said to satisfy Stripe-RIP with constant δ_k if:

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

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

Theorem: If the true representation $\boldsymbol{\Gamma}$ satisfies

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

Then a solution $\hat{\boldsymbol{\Gamma}}$ for $\left(\mathbf{P}_{0,\infty}^{\epsilon}\right)$ must be close to it

$$\left\|\hat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}\right\|_2^2 \le \frac{4\epsilon^2}{1-\delta_{2k}} \le \frac{4\epsilon^2}{1-(2k-1)\mu(\mathbf{D})}$$

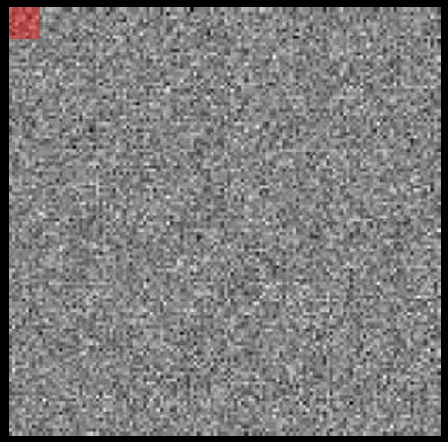
$$\delta_k \le (k-1)\mu(\mathbf{D})$$

Technion
Israel Institute of Technology

Local Noise Assumption

- Thus far, our analysis relied on the local sparsity of the underlying solution $\boldsymbol{\Gamma}$, which was enforced through the $\ell_{0,\infty}$ norm.
- In what follows, we present stability guarantees for both OMP and BP that will also depend on the local energy in the noise vector $\mathbf{E}$.
- This will be enforced via the $\ell_{2,\infty}$ norm, defined as:

$$\|\mathbf{E}\|_{2,\infty}^{\mathrm{p}} = \max_{\mathrm{i}} \|\mathbf{R}_{\mathrm{i}}\mathbf{E}\|_2$$



Stability of OMP

Theorem: If $\mathbf{Y} = \mathbf{D\Gamma} + \mathbf{E}$ where

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

Then OMP run for $\|\mathbf{\Gamma}\|_0$ iterations will

1. Find the correct support
2. $\|\mathbf{\Gamma}_{\mathrm{OMP}} - \mathbf{\Gamma}\|_2^2 \leq \frac{\|\mathbf{E}\|_2^2}{1-\left(\|\mathbf{\Gamma}\|_{0,\infty}^{s}-1\right)\mu(\mathbf{D})}$

Stability of Lagrangian BP

$$(\mathbf{P}_1^{\epsilon}): \quad \mathbf{\Gamma}_{\mathrm{BP}} = \min_{\mathbf{\Gamma}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D\Gamma}\|_2^2 + \xi\|\mathbf{\Gamma}\|_1$$

Theorem: For $\mathbf{Y} = \mathbf{D\Gamma} + \mathbf{E}$, if $\xi = 4\|\mathbf{E}\|_{2,\infty}^{\mathrm{p}}$ and

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

Then we are guaranteed that

1. The support of $\mathbf{\Gamma}_{\mathrm{BP}}$ is contained in that of Γ
2. $\|\mathbf{\Gamma}_{\mathrm{BP}} - \mathbf{\Gamma}\|_{\infty} \leq 7.5\|\mathbf{E}\|_{2,\infty}^{\mathrm{p}}$
3. Every entry greater than $7.5\|\mathbf{E}\|_{2,\infty}^{\mathrm{p}}$ will be found
4. $\mathbf{\Gamma}_{\mathrm{BP}}$ is unique.

Proof relies on the work of [Tropp '06]

Stability of Lagrangian BP

$$(\mathbf{P}_1^{\epsilon}): \quad \mathbf{\Gamma}_{\mathrm{BP}} = \min_{\mathbf{\Gamma}} \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{\Gamma}\|_2^2 + \xi\|\mathbf{\Gamma}\|_1$$

Theorem: For $\mathbf{Y} = \mathbf{D}\mathbf{\Gamma} + \mathbf{E}$, if $\xi = 4\|\mathbf{E}\|_{2,\infty}^{\mathrm{p}}$ and

$$\|\mathbf{\Gamma}\|_{0,\infty}^{\mathrm{s}} < \frac{1}{3}\left(1 + \frac{}{}\right.$$

Then we are guaranteed that

1. The support of $\mathbf{\Gamma}_{\mathrm{BP}}$ is contain
2. $\|\mathbf{\Gamma}_{\mathrm{BP}} - \mathbf{\Gamma}\|_{\infty} \leq 7.5\|\mathbf{E}\|_{2,\infty}^{\mathrm{p}}$
3. Every entry greater than 7.5
4. $\mathbf{\Gamma}_{\mathrm{BP}}$ is unique.

Theoretical foundation for recent works tackling the convolutional sparse coding problem via BP

[Bristow, Eriksson & Lucey '13]
[Wohlberg '14]
[Kong & Fowlkes '14]
[Bristow & Lucey '14]
[Heide, Heidrich & Wetzstein '15]
[Šorel & Šroubek '16]

Technion
Israel Institute of Technology

Proof relies on the work of [Tropp '06]

Global Pursuit via Local Processing

$$(\mathbf{P}_1^{\epsilon}): \quad \mathbf{\Gamma}_{\mathrm{BP}} = \min_{\mathbf{\Gamma}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{\Gamma}\|_2^2 + \xi\|\mathbf{\Gamma}\|_1$$

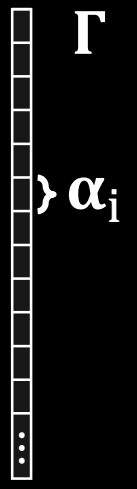
- Thus far, we have seen that while the CSC is a global model, its theoretical guarantees rely on local properties. Yet this global-local relation can also be exploited for practical purposes. Next, we show how one can solve the global BP problem using only local operations.
- Iterative Soft Thresholding [Blumensath & Davies '08]:

$$\mathbf{\Gamma}^{\mathrm{t}} = \mathcal{S}_{\xi/c}\left(\underbrace{\mathbf{\Gamma}^{\mathrm{t}-1} + \frac{1}{c}\mathbf{D}^{\mathrm{T}}(\mathbf{Y}-\mathbf{D}\mathbf{\Gamma}^{\mathrm{t}-1})}_{\text{Gradient step}}\right)$$

Projection onto $\mathbf{L}_1$ ball

- This can be equally written as:

$$\forall i \quad \underbrace{\boldsymbol{\alpha}_{\mathrm{i}}^{\mathrm{t}}}_{\text{local sparse code}} = \mathcal{S}_{\xi/c}\Big(\boldsymbol{\alpha}_{\mathrm{i}}^{\mathrm{t}-1} + \underbrace{\mathbf{D}_{\mathrm{L}}^{\mathrm{T}}}_{\text{local dictionary}} \underbrace{\mathbf{R}_{\mathrm{i}}(\mathbf{Y}-\overbrace{\mathbf{D}\mathbf{\Gamma}^{\mathrm{t}-1}}^{\text{global aggregation}})}_{\text{local residual}}\Big)$$

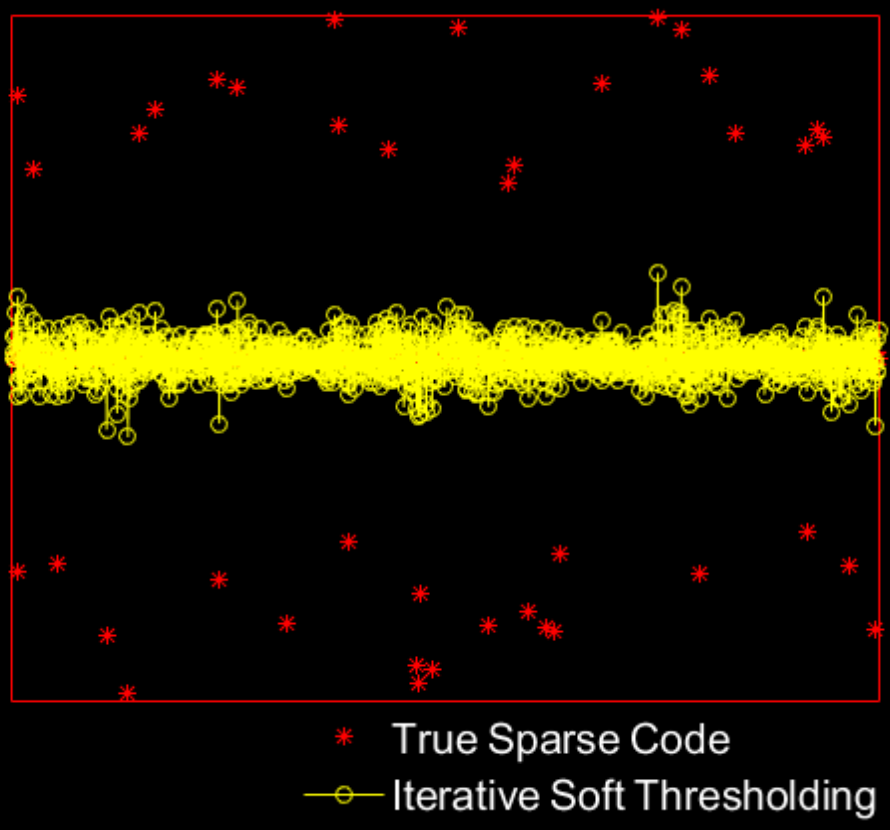


* $c > 0.5\,\lambda_{\max}(\mathbf{D}^{\mathrm{T}}\mathbf{D})$

Simulation

Details:

- Signal length: $N = 300$
- Patch size: $n = 25$
- Unique atoms: $p = 5$
- Global sparsity: $k = 40$
- Number of iterations: 400
- Lagrangian: $\xi = 4\|\mathbf{E}\|_{2,\infty}^{\mathrm{p}}$



Partial Summary of CSC

- What we have seen so far is a new way to analyze the global CSC model using local sparsity constrains. We proved:


Uniqueness of the solution to the noiseless problem.


Stability of the solution to the noisy problem.


Guarantee of success and stability of both OMP and BP.

- We obtained guarantees and algorithms that operate locally while claiming global optimality.

Part III

Going Deeper

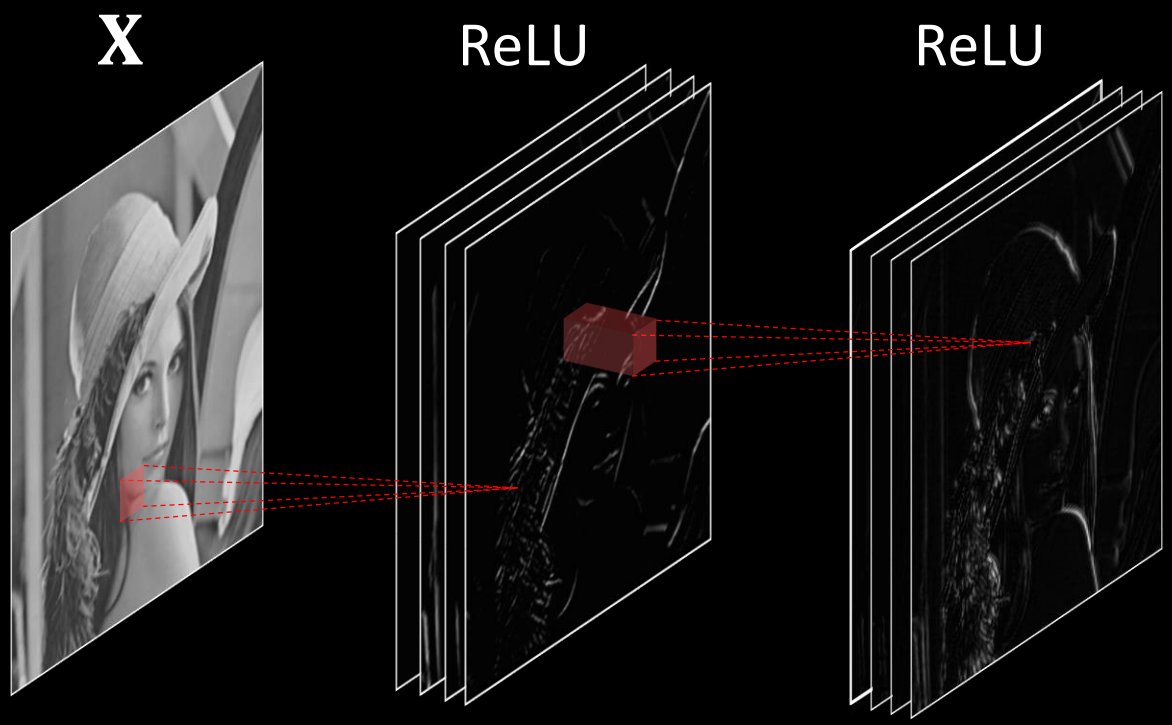
Convolutional Neural Networks Analyzed via Convolutional Sparse Coding

Vardan Papyan, Yaniv Romano and Michael Elad

CSC and CNN

- There seems to be a relation between CSC and CNN:
  - Both have a convolutional structure.
  - Both use a data driven approach for training their model.
  - The most popular non-linearity employed in CNN, called ReLU, is known to be connected to sparsity.
- In this part, we aim to make this connection clear in order to provide a theoretical understanding of CNN through the eyes of sparsity.
- But first, a short review of CNN…

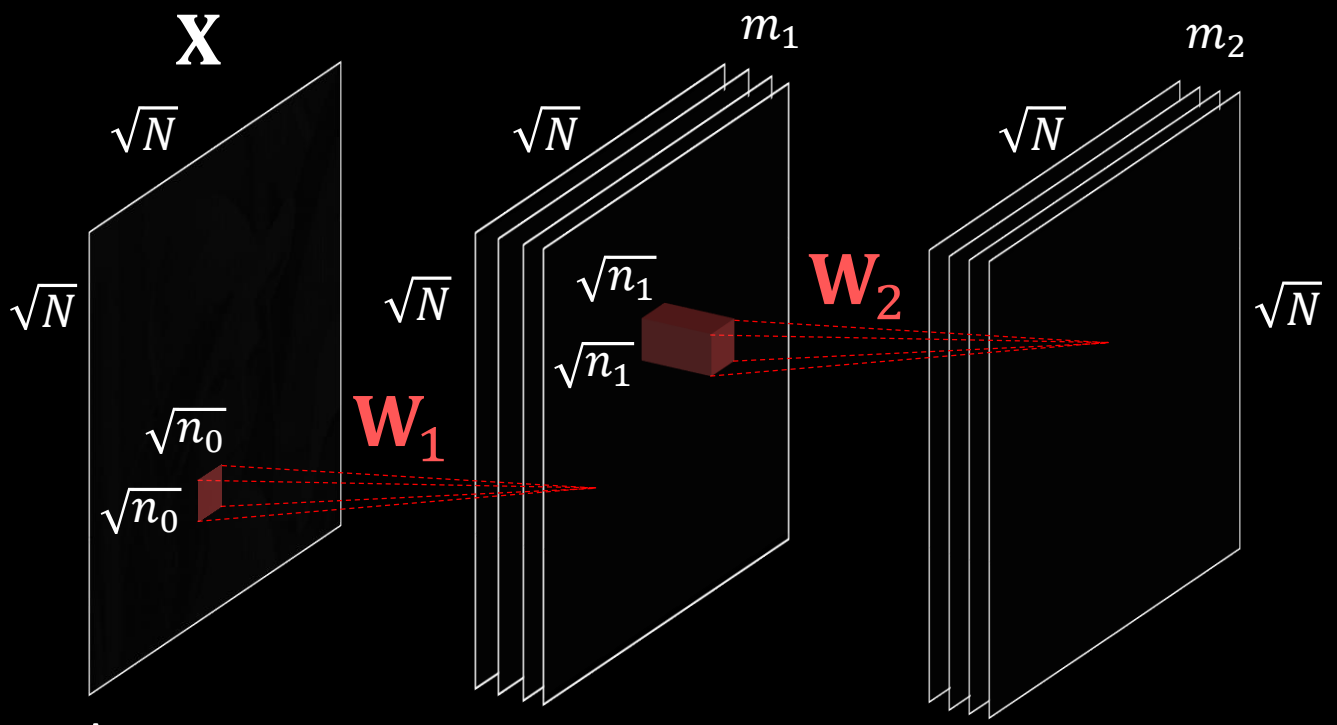
CNN



[LeCun, Bottou, Bengio and Haffner ‘98]
[Krizhevsky, Sutskever & Hinton ‘12]
[Simonyan & Zisserman ‘14]
[He, Zhang, Ren & Sun ‘15]

$$\mathrm{ReLU}(z) = \max(0, z)$$

CNN



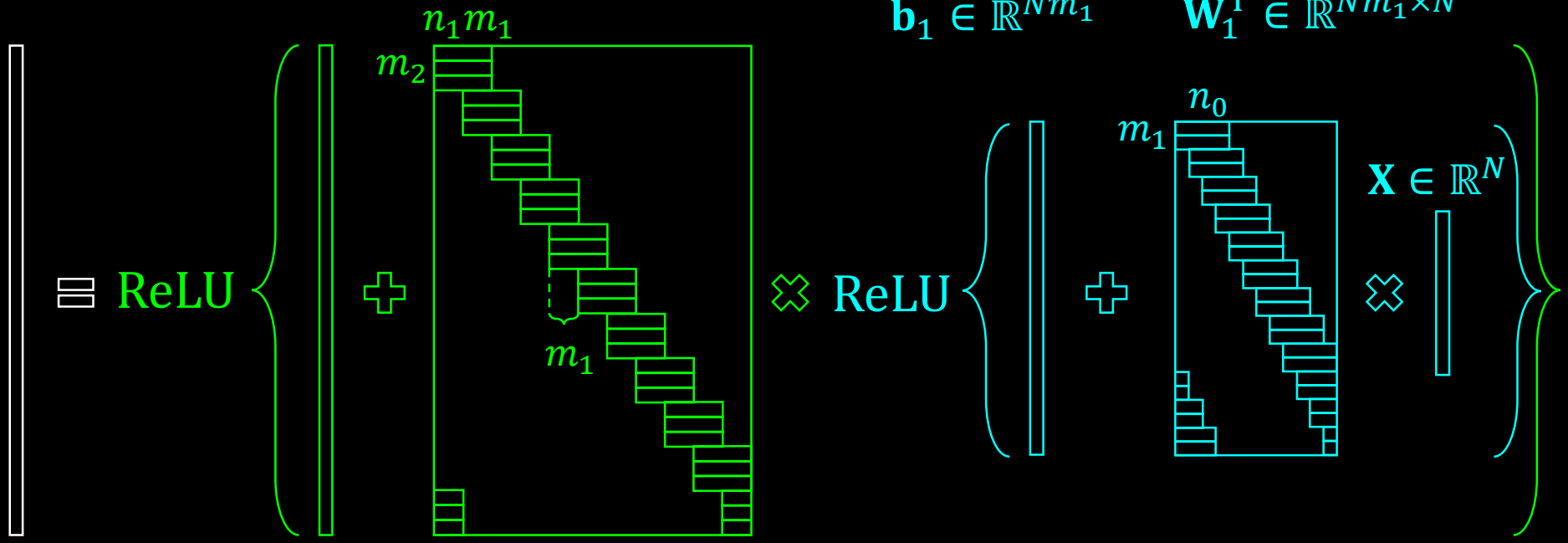
No pooling stage:

- Can be replaced by a convolutional layer with increased stride without loss in performance [Springenberg, Dosovitskiy, Brox & Riedmiller '14]
- The current state-of-the-art in image recognition does not use it [He, Zhang, Ren & Sun '15]

Mathematically…

$$f(\mathbf{X}, \{\mathbf{W}_i\}, \{\mathbf{b}_i\}) = \mathrm{ReLU}\left(\mathbf{b}_2 + \mathbf{W}_2^{\mathrm{T}}\,\mathrm{ReLU}\left(\mathbf{b}_1 + \mathbf{W}_1^{\mathrm{T}}\mathbf{X}\right)\right)$$

$\mathbf{Z}_2 \in \mathbb{R}^{Nm_2}$ $\mathbf{b}_2 \in \mathbb{R}^{Nm_2}$ $\mathbf{W}_2^{\mathrm{T}} \in \mathbb{R}^{Nm_2 \times Nm_1}$ $\mathbf{b}_1 \in \mathbb{R}^{Nm_1}$ $\mathbf{W}_1^{\mathrm{T}} \in \mathbb{R}^{Nm_1 \times N}$ $\mathbf{X} \in \mathbb{R}^{N}$



Training Stage of CNN

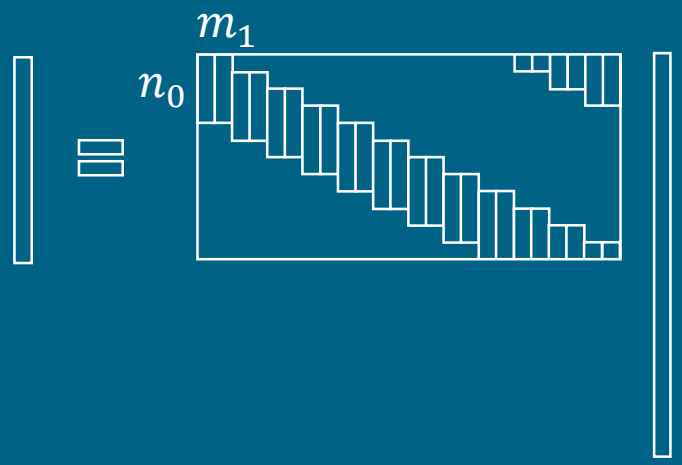
- Consider the task of classification, for example.
- Given a set of signals $\{\mathbf{X}_j\}_j$ and their corresponding labels $\{h(\mathbf{X}_j)\}_j$, the CNN learns an end-to-end mapping.

$$\min_{\{\mathbf{W}_i\},\{\mathbf{b}_i\},\mathbf{U}} \sum_j \ell\left(h(\mathbf{X}_j), \mathbf{U}, f(\mathbf{X}, \{\mathbf{W}_i\}, \{\mathbf{b}_i\})\right)$$

True label Classifier Output of last layer

Back to CSC

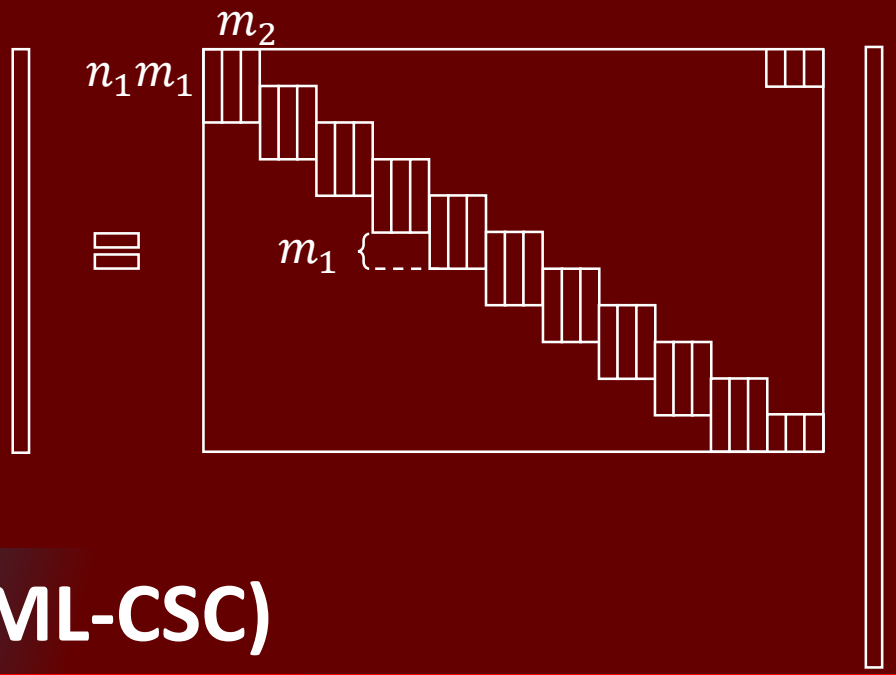
$\mathbf{X} \in \mathbb{R}^{N}$ $\quad \mathbf{D}_1 \in \mathbb{R}^{N \times N m_1}$ $\quad \boldsymbol{\Gamma}_1 \in \mathbb{R}^{N m_1}$



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

Similarly, the representations themselves could also be assumed to have such a structure.

$\boldsymbol{\Gamma}_1 \in \mathbb{R}^{N m_1}$ $\quad \mathbf{D}_2 \in \mathbb{R}^{N m_1 \times N m_2}$ $\quad \boldsymbol{\Gamma}_2 \in \mathbb{R}^{N m_2}$



Multi-Layer CSC (ML-CSC)

Deep Coding and Learning Problems

$(\mathbf{DCP}_\lambda)$: Find a set of representations satisfying

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

$$(\mathbf{DLP}_\lambda): \quad \min_{\{\mathbf{D_i}\}_{i=1}^{K},\mathbf{U}} \sum_{j} \ell\left(h\left(\mathbf{X_j}\right), \mathbf{U}, \mathbf{DCP}^{\star}\left(\mathbf{X_j}, \{\mathbf{D_i}\}\right)\right)$$

True label — Classifier — Deepest representation obtained by solving the DCP

Task driven dictionary learning
[Mairal, Bach & Ponce '12]

Deep Coding and Learning Problems

$\left(\mathbf{DCP}_{\lambda}^{\mathcal{E}}\right)$: Find a set of representations satisfying

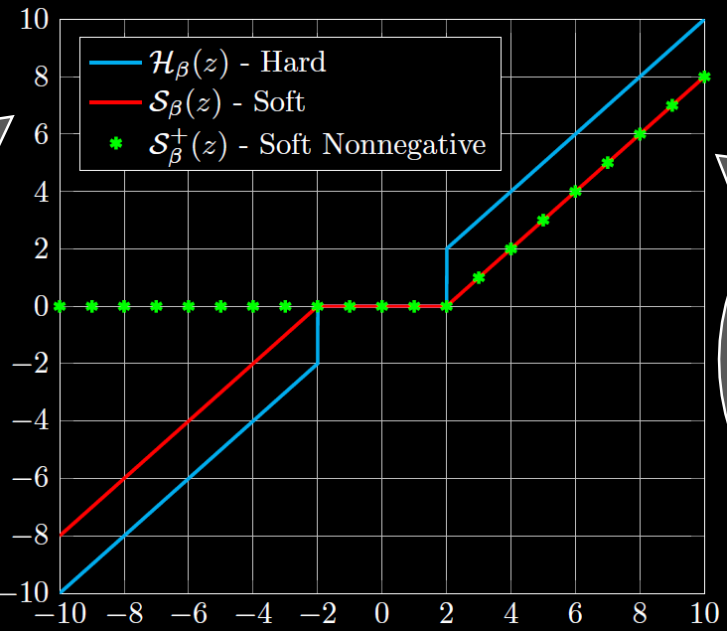
$$\begin{aligned} \|\mathbf{Y}-\mathbf{D}_1\boldsymbol{\Gamma}_1\|_2 &\le \mathcal{E}_0 & \|\boldsymbol{\Gamma}_1\|_{0,\infty}^{s} &\le \lambda_1 \\ \|\boldsymbol{\Gamma}_1-\mathbf{D}_2\boldsymbol{\Gamma}_2\|_2 &\le \mathcal{E}_1 & \|\boldsymbol{\Gamma}_2\|_{0,\infty}^{s} &\le \lambda_2 \\ &\vdots & &\vdots \\ \|\boldsymbol{\Gamma}_{K-1}-\mathbf{D}_K\boldsymbol{\Gamma}_K\|_2 &\le \mathcal{E}_{K-1} & \|\boldsymbol{\Gamma}_K\|_{0,\infty}^{s} &\le \lambda_K \end{aligned}$$

$$\left(\mathbf{DLP}_{\lambda}^{\mathcal{E}}\right): \quad \min_{\{\mathbf{D}_i\}_{i=1}^{K},\mathbf{U}} \sum_{j} \ell\left(\mathrm{h}\left(\mathbf{X}_j\right), \mathbf{U}, \mathbf{DCP}^{\star}\left(\mathbf{Y}_j, \{\mathbf{D}_i\}\right)\right)$$

True label — Classifier — Deepest representation obtained by solving the DCP

The simplest pursuit in the sparse representation is the thresholding algorithm. Given an input signal $\mathbf{X}$, this operates by:

$$\mathbf{D}^{\mathrm{T}}\mathbf{X} \Rightarrow \mathcal{P}_{\beta}\left(\mathbf{D}^{\mathrm{T}}\mathbf{X}\right)$$



Restricting the coefficients to be nonnegative does not restrict the expressiveness of the model

ReLU = Soft Nonnegative Thresholding [Fawzi, Davies & Frossard '15]

Consider this for Solving the DCP

- Layered thresholding (LT):

Estimate $\boldsymbol{\Gamma}_1$ via the thresholding algorithm

$$\widehat{\boldsymbol{\Gamma}}_2 = \mathcal{P}_{\beta_2}\left(\mathbf{D}_2^{\mathrm{T}}\,\mathcal{P}_{\beta_1}\left(\mathbf{D}_1^{\mathrm{T}}\mathbf{X}\right)\right)$$

Estimate $\boldsymbol{\Gamma}_2$ via the thresholding algorithm

$(\mathbf{DCP}_\lambda)$: Find a set of representations satisfying

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

- Forward pass of CNN:

$$f(\mathbf{X}) = \mathrm{ReLU}\left(\mathbf{b}_2 + \mathbf{W}_2^{\mathrm{T}}\,\mathrm{ReLU}\left(\mathbf{b}_1 + \mathbf{W}_1^{\mathrm{T}}\mathbf{X}\right)\right)$$

The layered (soft nonnegative) thresholding and the forward pass algorithm are equal !!!

Consider this for Solving the DLP

- DLP:

$$\min_{\{\mathbf{D}_i\}_{i=1}^{K},\mathbf{U}} \sum_j \ell\left(\mathrm{h}(\mathbf{X}_j), \mathbf{U}, \mathbf{DCP}^{\star}\left(\mathbf{X}_j, \{\mathbf{D}_i\}\right)\right)$$

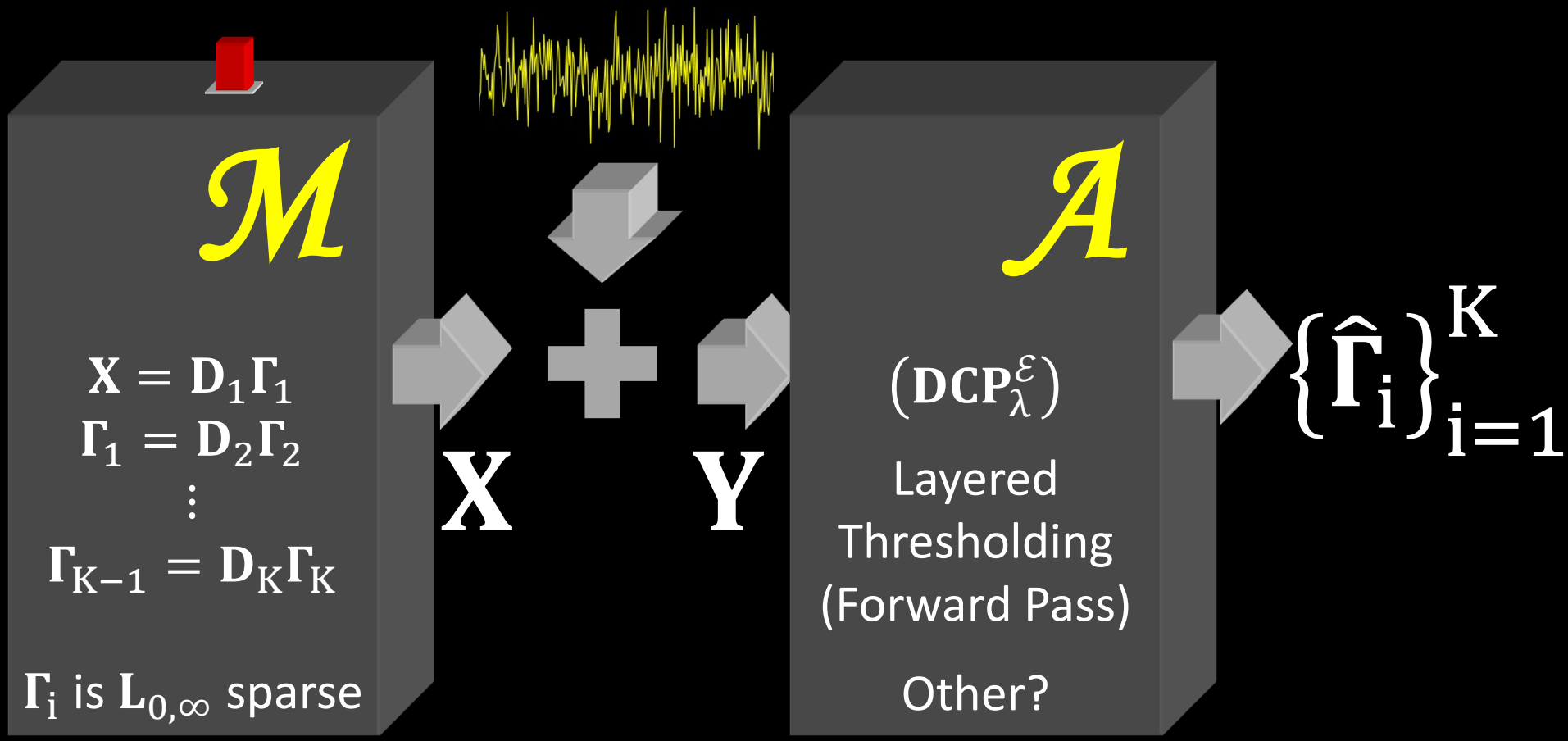
Estimate via the layered thresholding algorithm

- CNN training:

$$\min_{\{\mathbf{W}_i\},\{\mathbf{b}_i\},\mathrm{U}} \sum_j \ell\left(\mathrm{h}(\mathbf{X}_j), \mathbf{U}, f(\mathbf{X}, \{\mathbf{W}_i\}, \{\mathbf{b}_i\})\right)$$

The problem solved by the training stage of CNN and the DLP are equal assuming that the DCP is approximated via the layered thresholding algorithm

Theoretical Questions



Uniqueness of $(\mathbf{DCP}_\lambda)$

$(\mathbf{DCP}_\lambda)$: Find a set of representations satisfying

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

Is this set unique?

Theorem: If a set of solutions $\{\boldsymbol{\Gamma}_i\}_{i=1}^{K}$ is found for $(\mathbf{DCP}_\lambda)$ such that:

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

Then these are necessarily the unique solution to this problem.

The feature maps CNN aims to recover are unique

[Papyan, Sulam & Elad '16]

Stability of $\left(\mathbf{DCP}_{\lambda}^{\mathcal{E}}\right)$

Theorem: If the true representations $\{\boldsymbol{\Gamma}_i\}_{i=1}^{K}$ satisfy

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

And the error thresholds for $(\mathbf{DCP}_{\lambda}^{\mathcal{E}})$ are

$$\mathcal{E}_0^2 = \|\mathbf{E}\|_2^2, \qquad \mathcal{E}_i^2 = \frac{4\mathcal{E}_{i-1}^2}{1-\left(2\|\boldsymbol{\Gamma}_i\|_{0,\infty}^{s}-1\right)\mu(\mathbf{D}_i)}$$

Then the set of solutions $\{\widehat{\boldsymbol{\Gamma}}_i\}_{i=1}^{K}$ obtained by solving this problem must be close to the true ones

$$\left\|\widehat{\boldsymbol{\Gamma}}_i - \boldsymbol{\Gamma}_i\right\|_2^2 \leq \mathcal{E}_i^2$$

The problem CNN aims to solve is stable under certain conditions

[Papyan, Sulam & Elad '16]

Stability of LT

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

Then the layered soft thresholding will*

1. Find the correct supports
2. $\left\|\Gamma_i^{LT} - \Gamma_i\right\|_{2,\infty}^{p} \leq \varepsilon_L^i$

We have defined $\varepsilon_L^0 = \|E\|_{2,\infty}^{p}$ and

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

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

* For correctly chosen thresholds

Limitations of the Forward Pass

- The stability analysis reveals several inherent limitations of the forward pass algorithm:
  - Even in the noiseless case, it is incapable of recovering the solution of the DCP problem.
  - Its success depends on the ratio $\left|\boldsymbol{\Gamma}_i^{\min}\right|/|\boldsymbol{\Gamma}_i^{\max}|$. This is a direct consequence of the forward pass algorithm relying on the simple thresholding operator.
  - The distance between the true sparse vector and the estimated one increases exponentially as function of the layer depth.
- In the next and final part we propose a new algorithm attempting to solve some of these problems.

Part IV
What Next?

Convolutional Neural Networks Analyzed via Convolutional Sparse Coding

Vardan Papyan, Yaniv Romano and Michael Elad

Layered Basis Pursuit (Noiseless)

- **Our Goal:** $(\mathbf{DCP}_\lambda)$: Find a set of representations satisfying

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

- Layered thresholding: $\hat{\boldsymbol{\Gamma}}_2 = \mathcal{P}_{\beta_2}\left(\mathbf{D}_2^T\,\mathcal{P}_{\beta_1}\left(\mathbf{D}_1^T\mathbf{X}\right)\right)$
- Thresholding is the simplest pursuit known in the field of sparsity.
- Layered BP: $\boldsymbol{\Gamma}_1^{LBP} = \min_{\boldsymbol{\Gamma}_1} \|\boldsymbol{\Gamma}_1\|_1 \;\; \text{s.t.} \;\; \mathbf{X} = \mathbf{D}_1\boldsymbol{\Gamma}_1$

$$\boldsymbol{\Gamma}_2^{LBP} = \min_{\boldsymbol{\Gamma}_2} \|\boldsymbol{\Gamma}_2\|_1 \;\; \text{s.t.} \;\; \boldsymbol{\Gamma}_1^{LBP} = \mathbf{D}_2\boldsymbol{\Gamma}_2$$

Deconvolutional networks
[Zeiler, Krishnan, Taylor & Fergus '10]

Guarantee for Success of Layered BP

- **Our Goal:** $(\mathbf{DCP}_\lambda)$: Find a set of representations satisfying

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

Theorem: If a set of solutions $\{\boldsymbol{\Gamma}_i\}_{i=1}^K$ of $(\mathbf{DCP}_\lambda)$ satisfy

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

Then the layered BP is guaranteed to find them.

- ✓The layered BP can retrieve the underlying representations in the noiseless case, a task in which the forward pass fails.
- ✓Its success does not depend on the ratio $\left|\boldsymbol{\Gamma}_i^{\min}\right|/|\boldsymbol{\Gamma}_i^{\max}|$.

[Papyan, Sulam & Elad '16]

Stability of Layered BP

Theorem: Assuming that

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

Then we are guaranteed that*

1. The support of $\mathbf{\Gamma}_i^{LBP}$ is contained in that of $\mathbf{\Gamma}_i$
2. $\left\|\mathbf{\Gamma}_i^{LBP} - \mathbf{\Gamma}_i\right\|_{2,\infty} \leq \varepsilon_L^i$
3. Every entry in $\mathbf{\Gamma}_i$ greater than $\varepsilon_L^i / \sqrt{\|\mathbf{\Gamma}_i\|_{0,\infty}^{p}}$ will be found

$$\varepsilon_L^i = 7.5^i \|\mathbf{E}\|_{2,\infty} \prod_{j=1}^{i} \sqrt{\|\mathbf{\Gamma}_j\|_{0,\infty}^{p}}$$

* For correctly chosen $\{\xi_i\}_{i=1}^{K}$

[Papyan, Sulam & Elad '16]

Layered Iterative Thresholding

Layered BP

$$\boldsymbol{\Gamma}_1^{\mathrm{LBP}} = \min_{\boldsymbol{\Gamma}_1} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}_1\boldsymbol{\Gamma}_1\|_2^2 + \xi_1\|\boldsymbol{\Gamma}_1\|_1$$

$$\boldsymbol{\Gamma}_2^{\mathrm{LBP}} = \min_{\boldsymbol{\Gamma}_2} \frac{1}{2}\left\|\boldsymbol{\Gamma}_1^{\mathrm{LBP}}-\mathbf{D}_2\boldsymbol{\Gamma}_2\right\|_2^2 + \xi_2\|\boldsymbol{\Gamma}_2\|_1$$

Can be seen as a recurrent neural network [Gregor & LeCun '10]

Layered IT

$$\boldsymbol{\Gamma}_1^t = \mathcal{S}_{\xi_1/c_1}\left(\boldsymbol{\Gamma}_1^{t-1} + \frac{1}{c_1}\mathbf{D}_1^T(\mathbf{Y}-\mathbf{D}_1\boldsymbol{\Gamma}_1^{t-1})\right)$$

$$\boldsymbol{\Gamma}_2^t = \mathcal{S}_{\xi_2/c_2}\left(\boldsymbol{\Gamma}_2^{t-1} + \frac{1}{c_2}\mathbf{D}_2^T(\widehat{\boldsymbol{\Gamma}}_1-\mathbf{D}_2\boldsymbol{\Gamma}_2^{t-1})\right)$$

* $c_i > 0.5\,\lambda_{\max}(\mathbf{D}_i^T\mathbf{D}_i)$

Conclusion

We described the limitations of patch based processing as a motivation for the CSC model.

We then presented a theoretical study of this model both in a noiseless and a noisy setting.

A multi-layer extension for it, tightly connected to CNN, was proposed and similarly analyzed.

Finally, an alternative to the forward pass algorithm was presented.

Future Work: leveraging theoretical insights into practical implications

Questions?