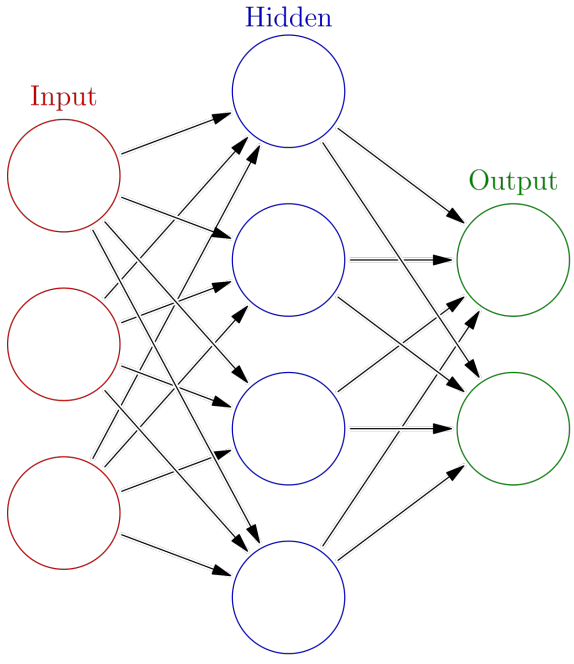
Deep and Convolutional Sparse Models

EN.580.709 - Fall 2019

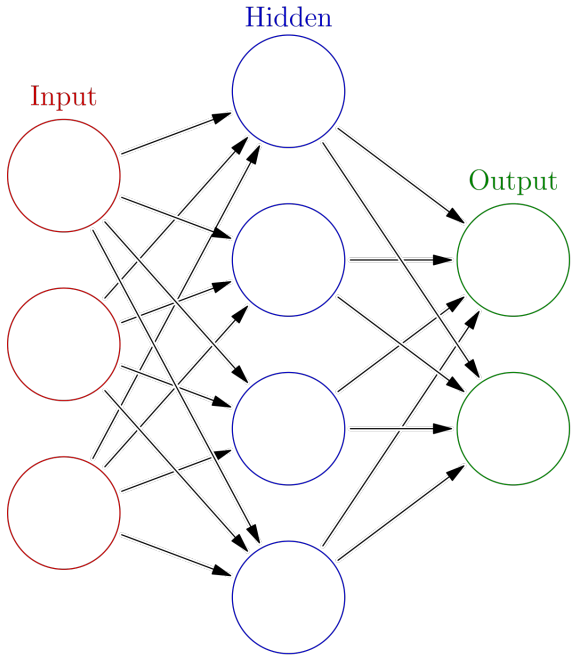
What are artificial neural networks?



- Input features

$$\mathbf{x} \in \mathbb{R}^n$$

What are artificial neural networks?



- Input features

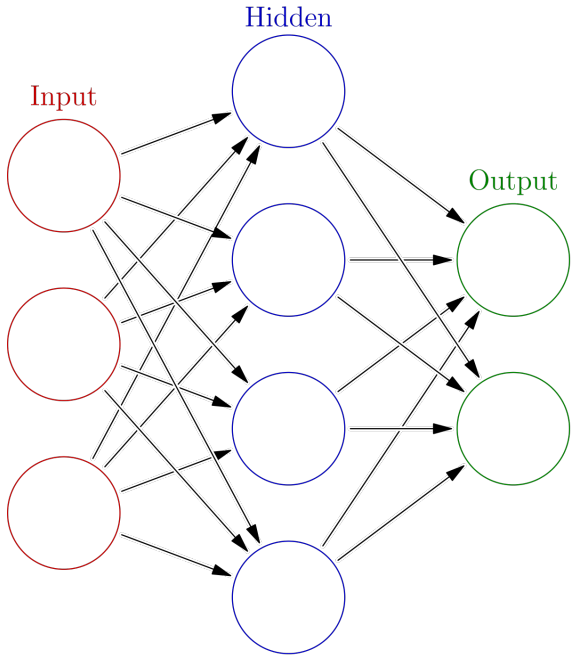
$$\mathbf{x} \in \mathbb{R}^n$$

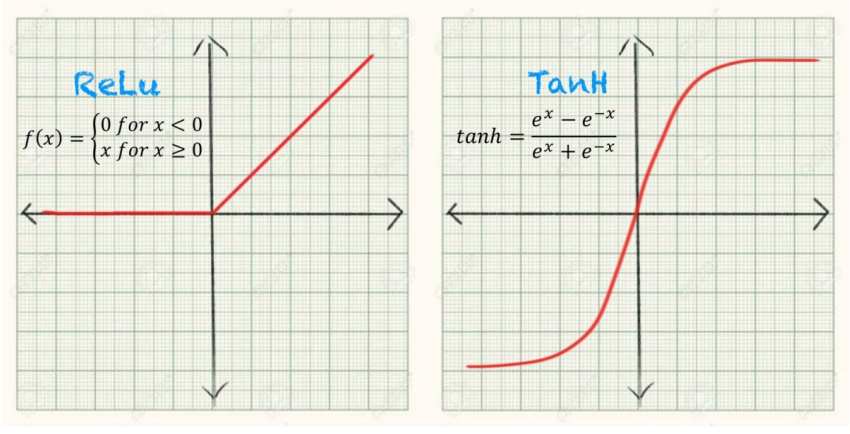
- Hidden neuron

$$h_j = \sigma\left(\sum_{i=1}^{n} w_{i,j} x_i + b_j\right)$$

$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) \in \mathbb{R}^m$$

What are artificial neural networks?





- Input features

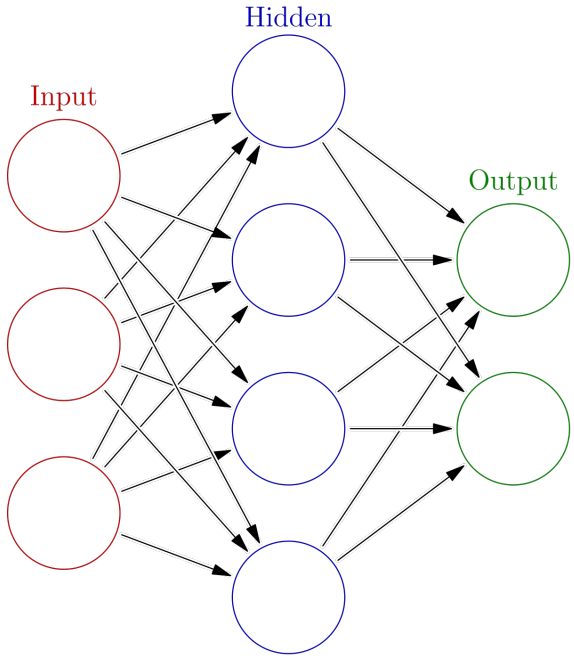
$$\mathbf{x} \in \mathbb{R}^n$$

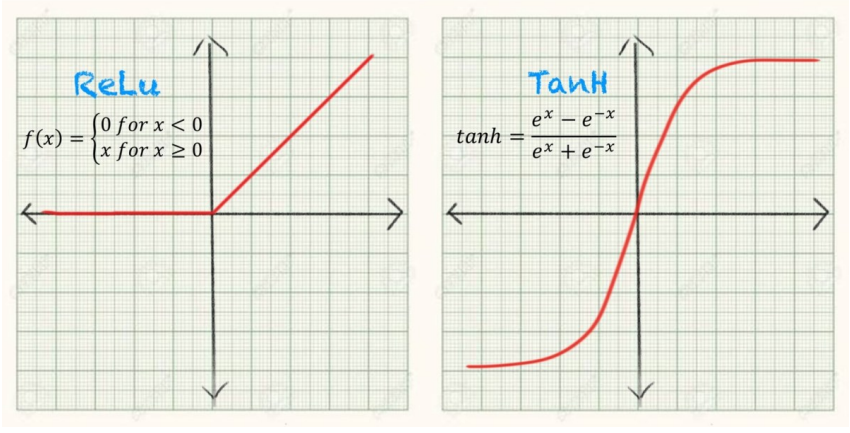
- Hidden neuron

$$h_j = \sigma\left(\sum_{i=1}^{n} w_{i,j} x_i + b_j\right)$$

$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) \in \mathbb{R}^m$$

What are artificial neural networks?





- Input features

$$\mathbf{x} \in \mathbb{R}^n$$

- Hidden neuron

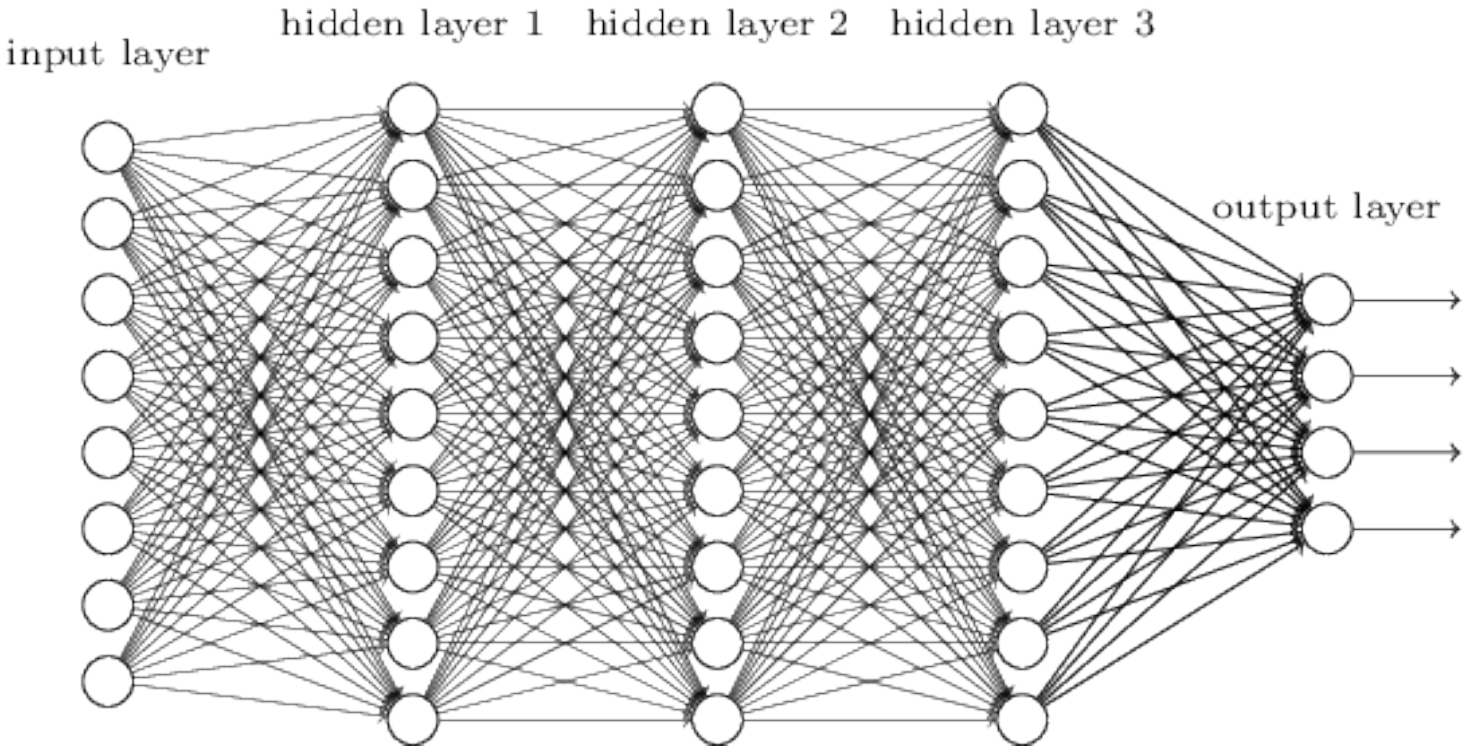
$$h_j = \sigma\left(\sum_{i=1}^{n} w_{i,j} x_i + b_j\right)$$

$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) \in \mathbb{R}^m$$

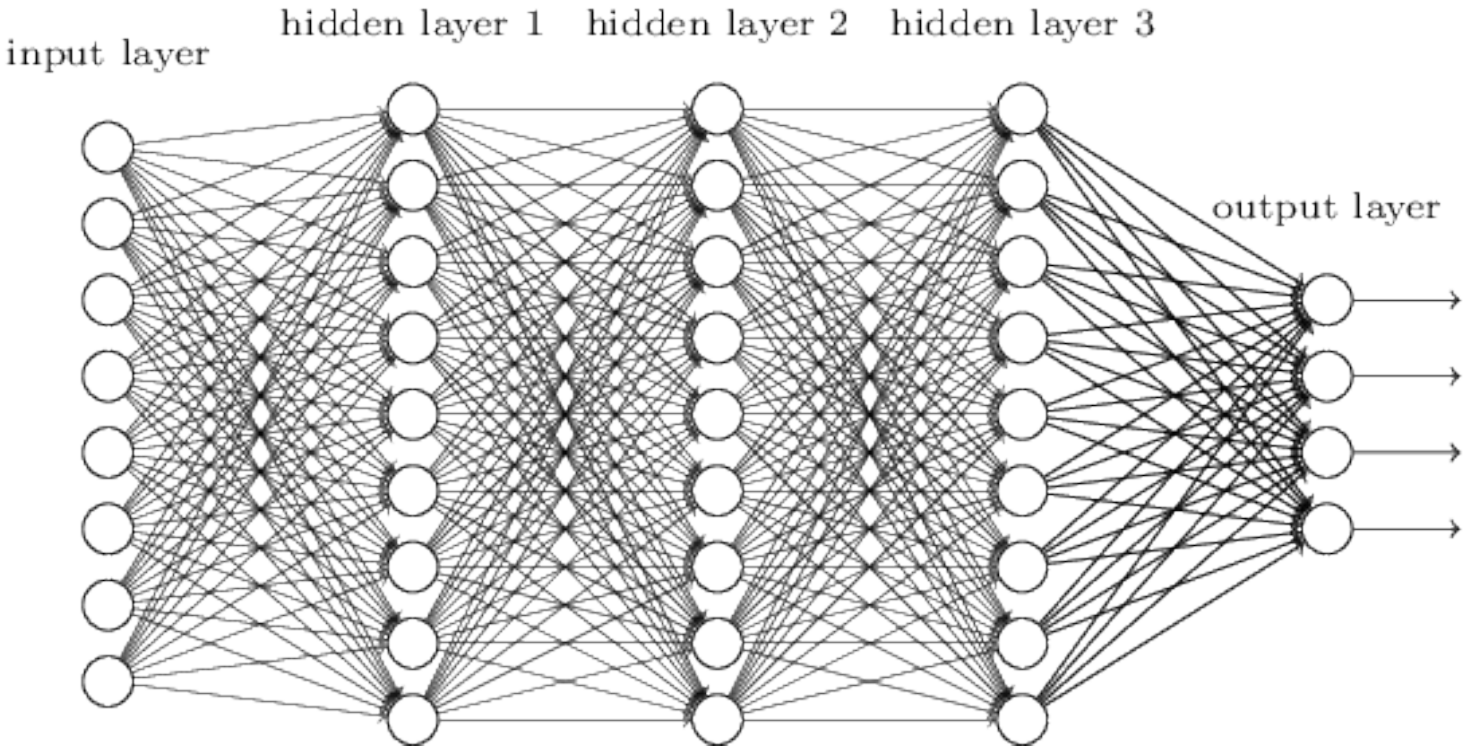
- Output

$$\mathbf{y} = \mathbf{W}_2\mathbf{h} \in \mathbb{R}^{m_2}$$

What are deep neural networks?

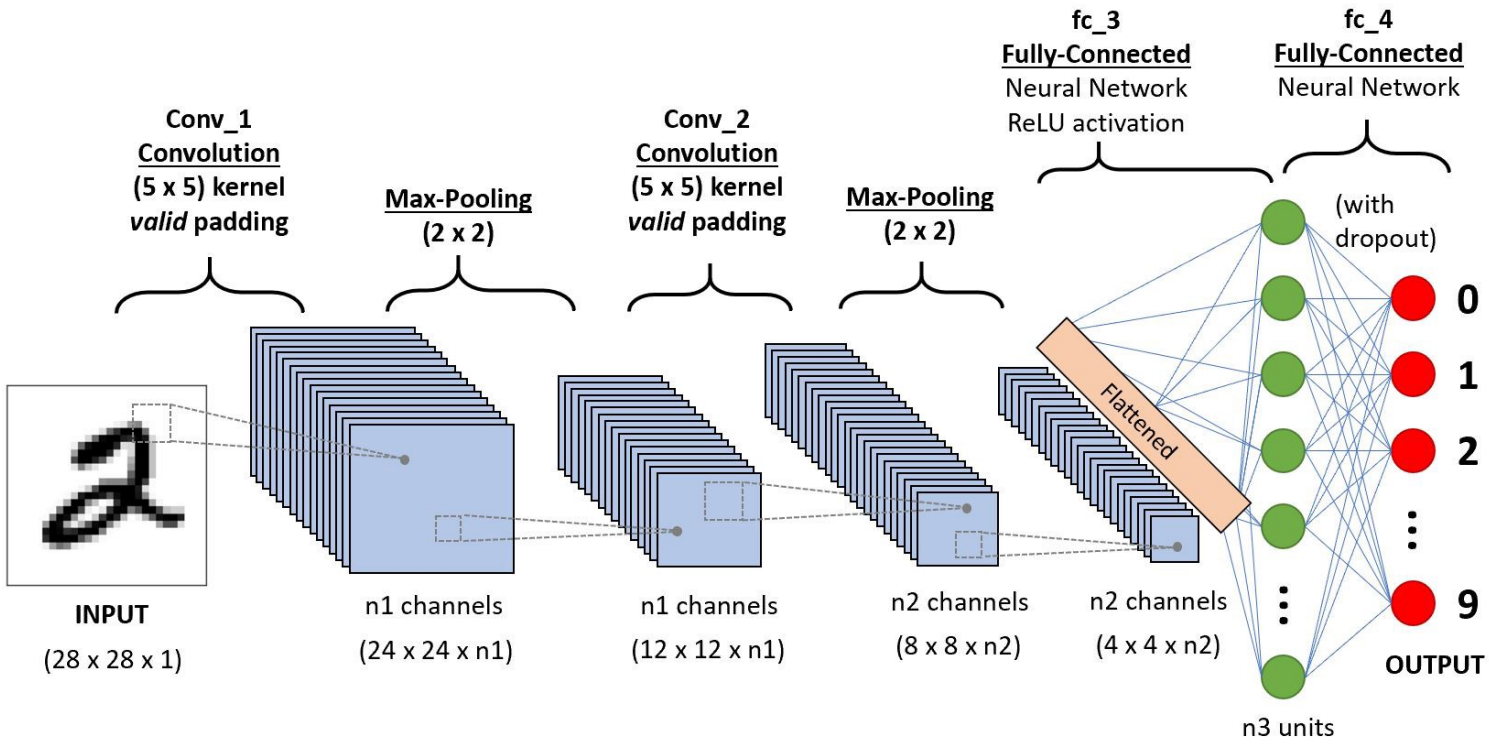


What are deep neural networks?



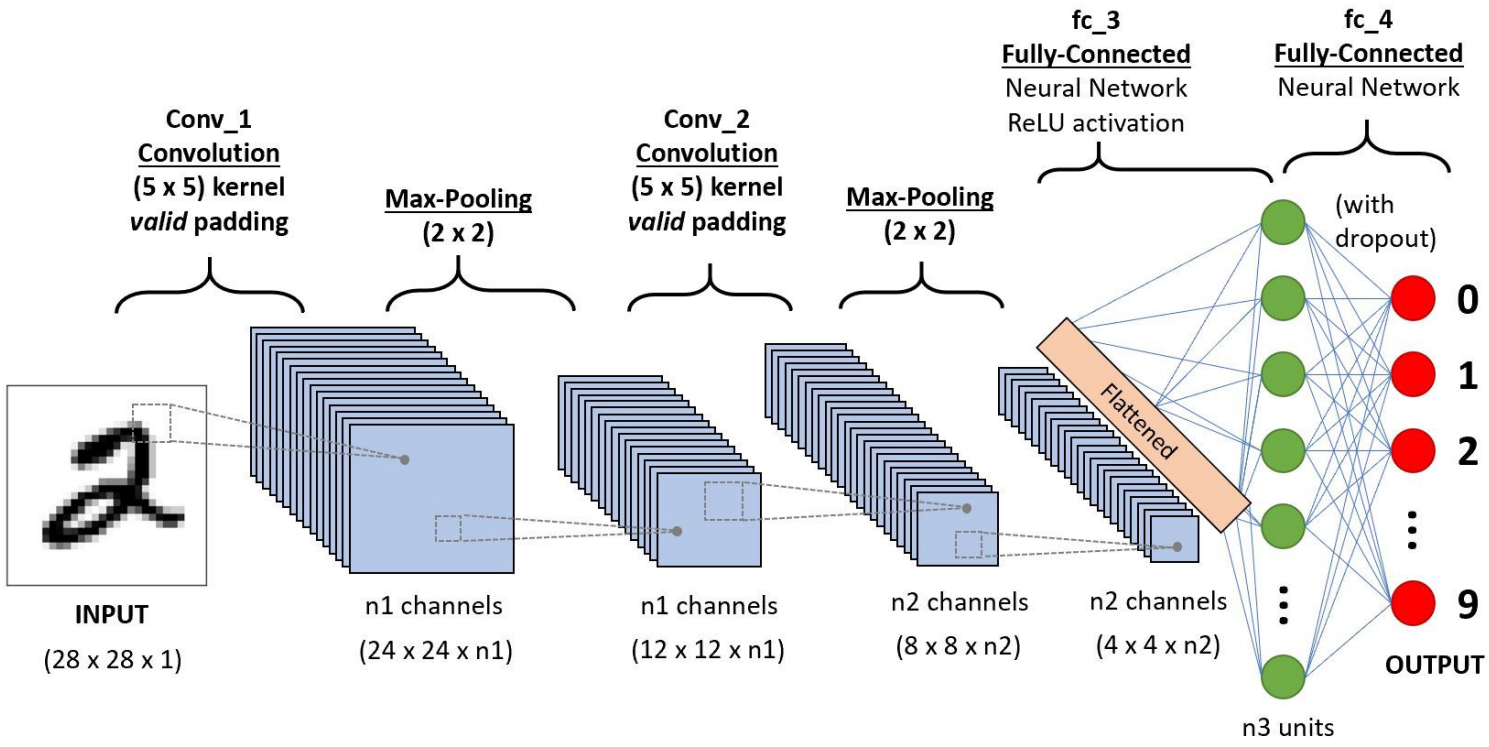
$$\mathbf{y} = \Phi(\mathbf{x}) = \mathbf{W}_L \sigma(\mathbf{W}_{L-1} \sigma(\dots \sigma(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) \dots) + \mathbf{b}_{L-1})$$

What are deep convolutional neural networks?



Convolutional Feature Maps

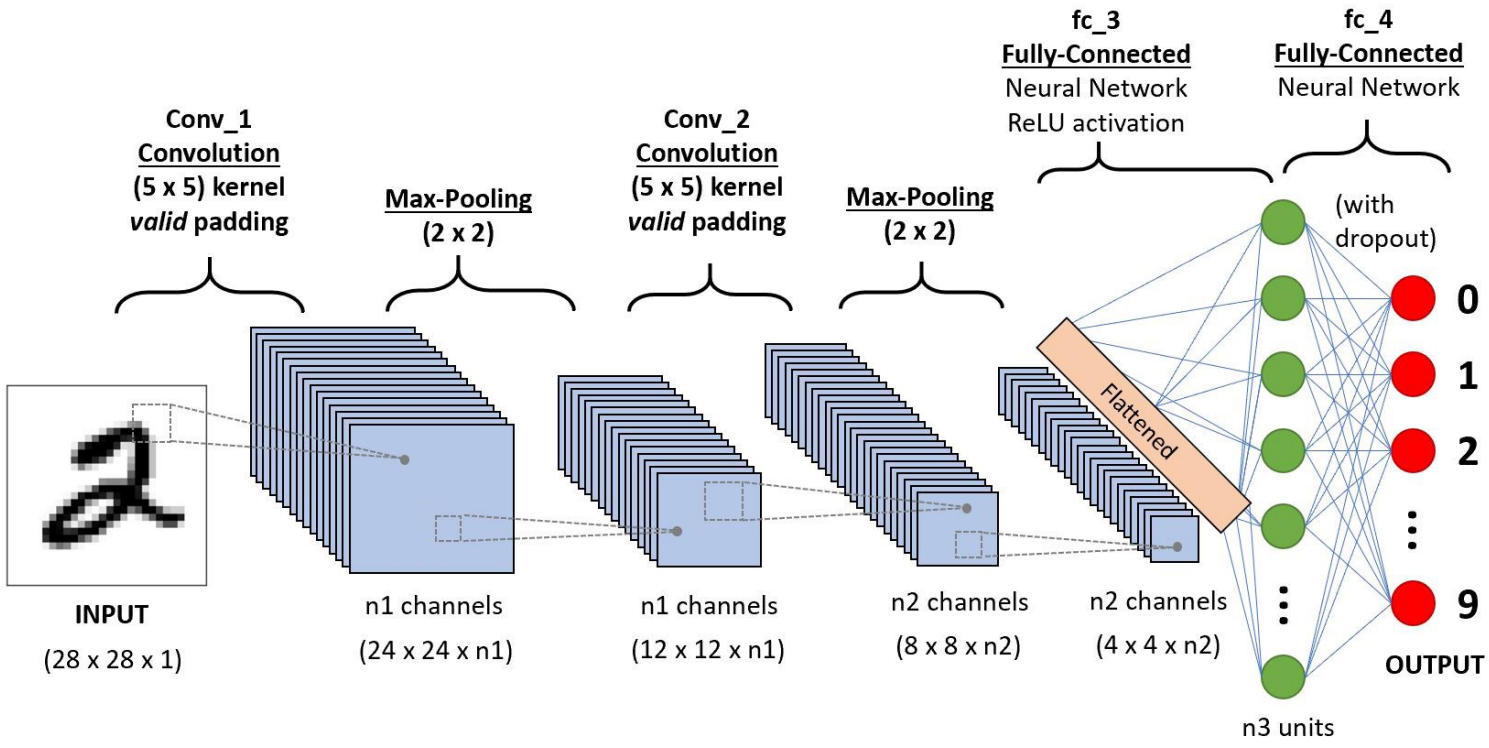
What are deep convolutional neural networks?



Convolutional Feature Maps

$$\mathbf{h}_i = \sigma(\mathbf{w}_1 * \mathbf{x} + b_i)$$

What are deep convolutional neural networks?

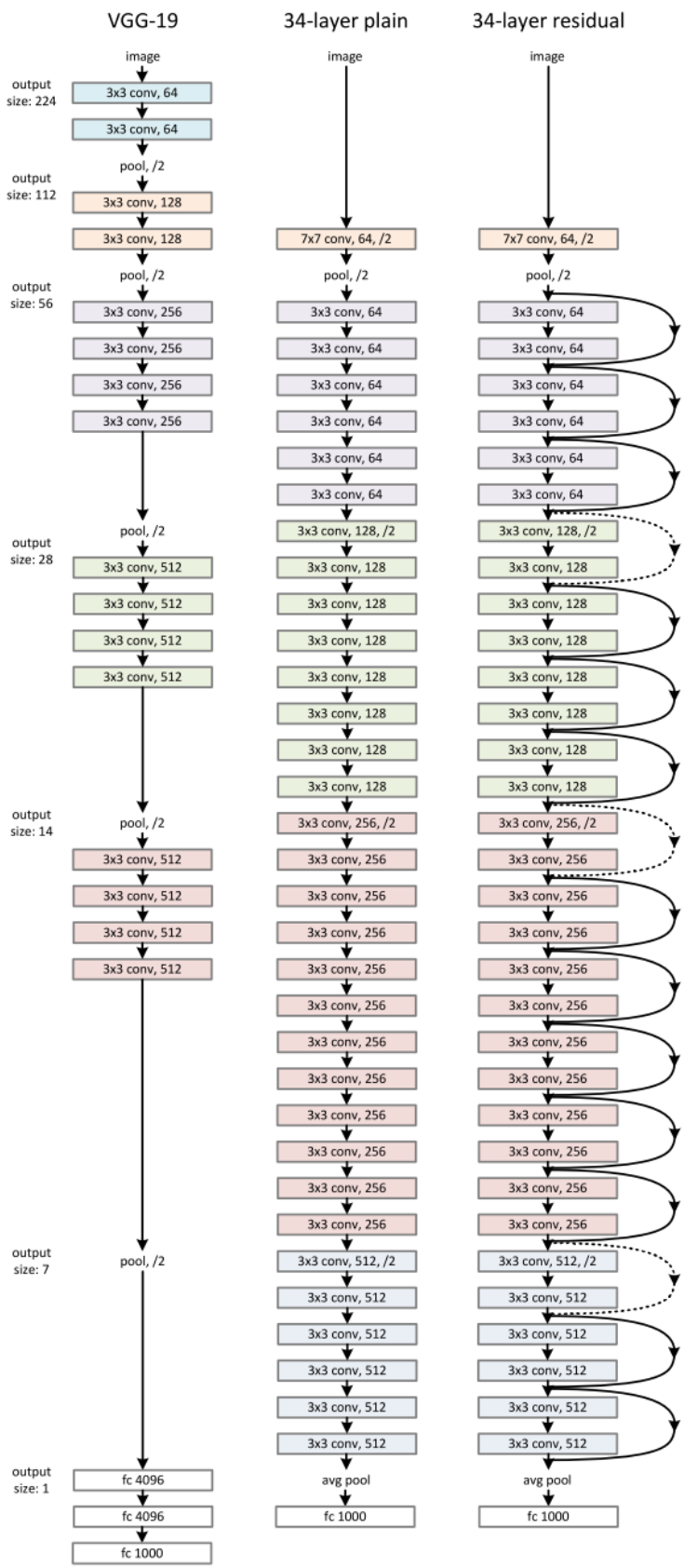


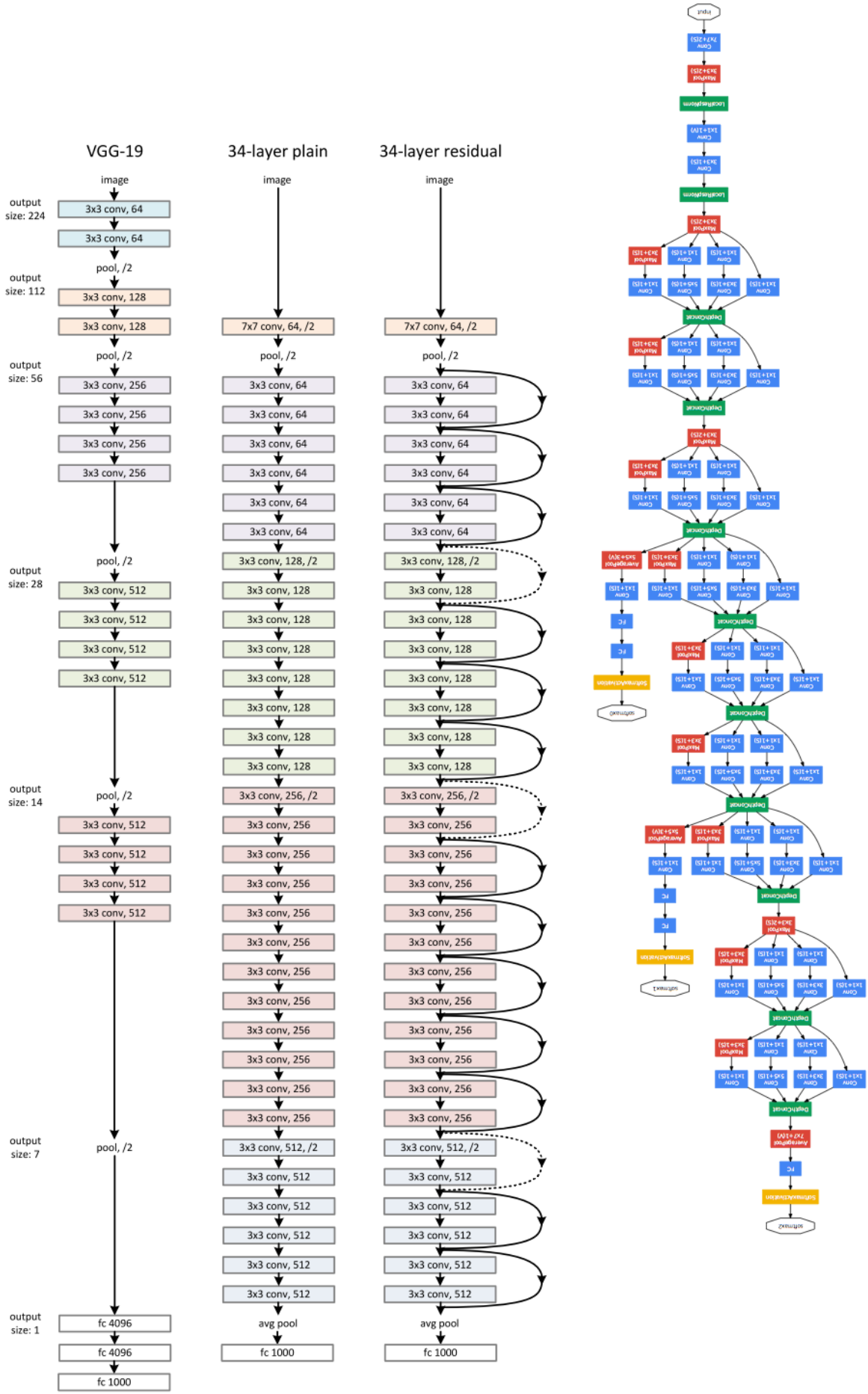
Convolutional Feature Maps

$$\mathbf{h}_i = \sigma(\mathbf{w}_1 * \mathbf{x} + b_i)$$

Max Pooling

$$\mathbf{h}(i) = \max_{j \in \mathcal{N}(s \cdot i)} \mathbf{x}(j)$$

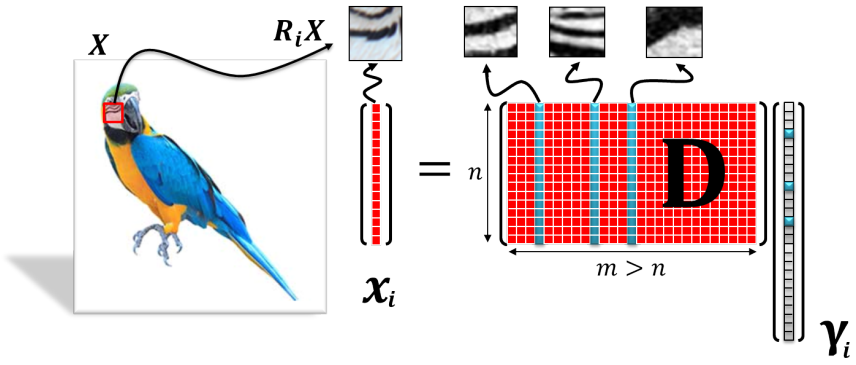

VGG-19
34-layer plain
34-layer residual
image
output size: 224
3x3 conv, 64
3x3 conv, 64
pool, /2
output size: 112
3x3 conv, 128
3x3 conv, 128
pool, /2
output size: 56
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
pool, /2
output size: 28
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
output size: 14
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
output size: 7
fc 4096
fc 4096
fc 1000
output size: 1
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
avg pool
fc 1000


VGG-19
34-layer plain
34-layer residual
image
output size: 224
3x3 conv, 64
3x3 conv, 64
pool, /2
output size: 112
3x3 conv, 128
3x3 conv, 128
pool, /2
output size: 56
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
pool, /2
output size: 28
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
output size: 14
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
output size: 7
fc 4096
fc 4096
fc 1000
output size: 1
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
avg pool
fc 1000

Guiding Questions

- What are the signal models behind these algorithms?
- Why these algorithms and not others?
- Understanding, theoretical guarantees, guidelines

Synthesis Sparse Modeling



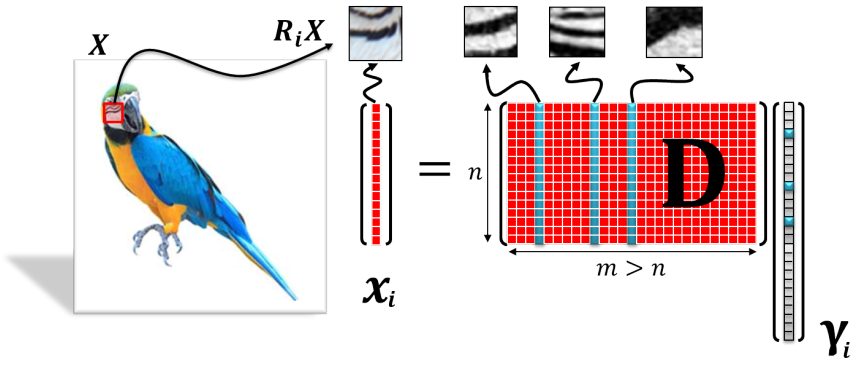
Model:

$$\mathbf{y} = \mathbf{D}\boldsymbol{\gamma} + \mathbf{w}, \quad \|\boldsymbol{\gamma}\|_0 \leq k$$

Pursuit:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\boldsymbol{\gamma}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}\|_0 \leq k$$

Synthesis Sparse Modeling



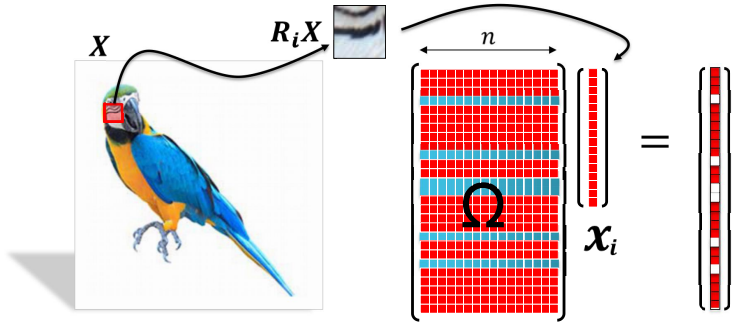
Model:

$$\mathbf{y} = \mathbf{D}\boldsymbol{\gamma} + \mathbf{w}, \quad \|\boldsymbol{\gamma}\|_0 \leq k$$

Pursuit:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\boldsymbol{\gamma}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}\|_0 \leq k$$

Analysis Sparse Modeling



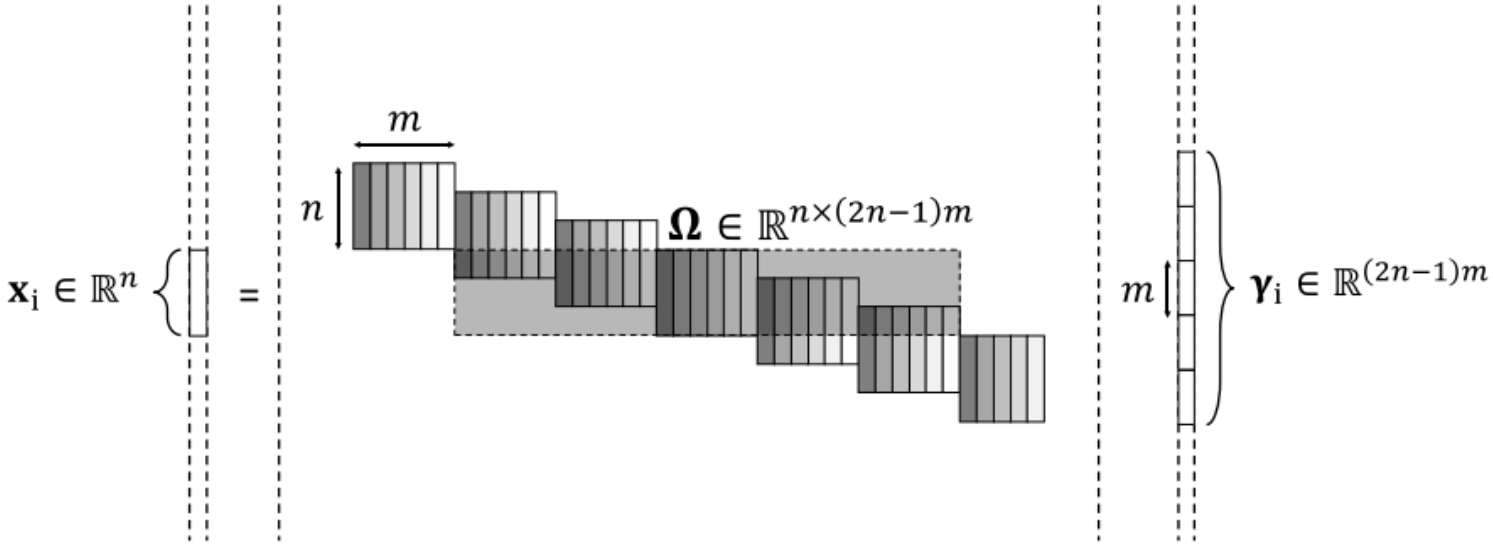
Model:

$$\mathbf{y} = \mathbf{x} + \mathbf{w}, \quad \|\boldsymbol{\Omega}\mathbf{x}\|_0 \leq m - \ell$$

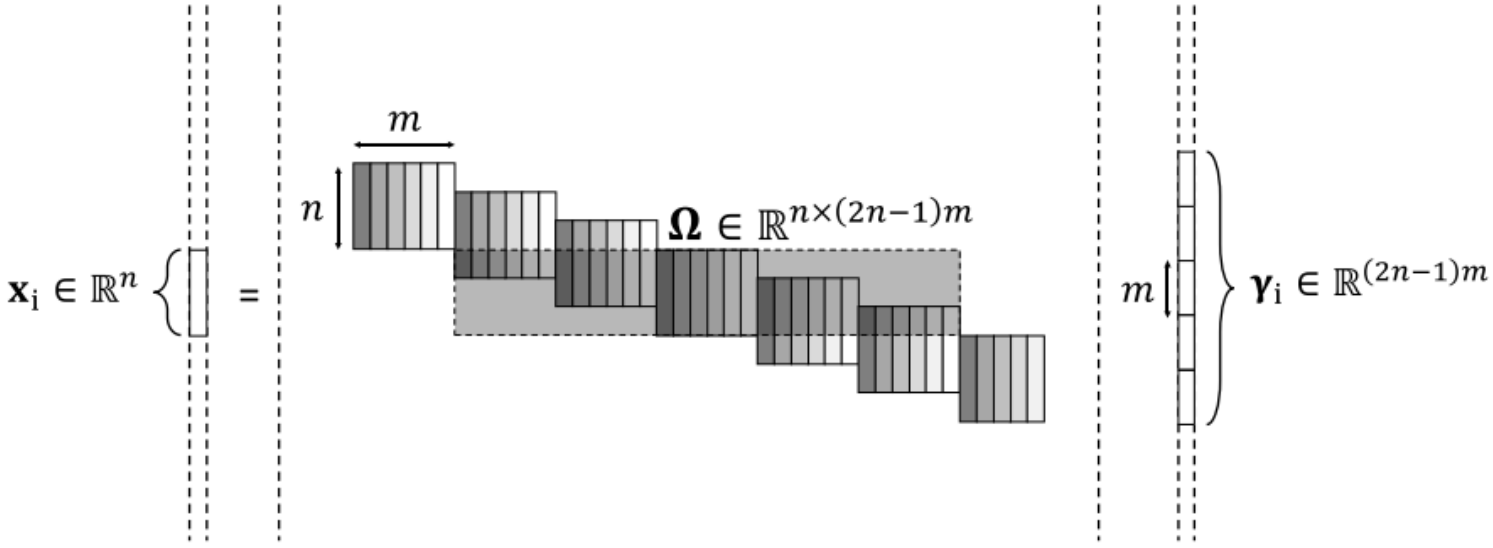
Pursuit:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{x}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\Omega}\mathbf{x}\|_0 \leq m - \ell$$

Recap on Convolutional Sparse Models



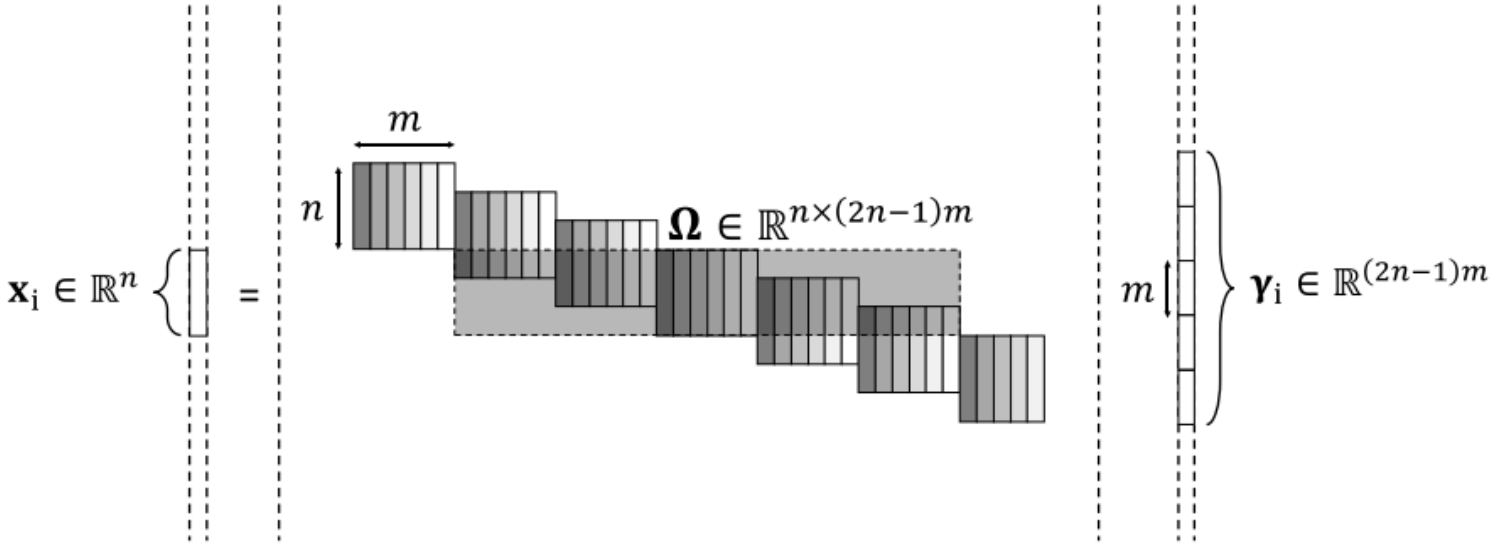
Recap on Convolutional Sparse Models



Shift-Invariant Model:

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

Recap on Convolutional Sparse Models



Shift-Invariant Model:

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

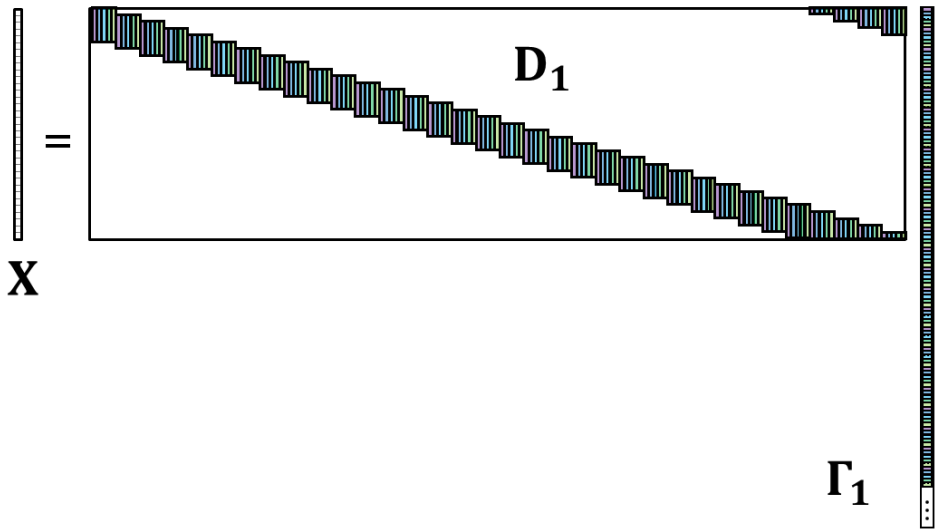
New sparsity measure:

$$\|\boldsymbol{\Gamma}\|_{0,\infty} := \max_i \|\boldsymbol{\gamma}_i\|_0 = \max_i \|\mathbf{S}_i\boldsymbol{\Gamma}\|_0$$

Multilayer (& convolutional) sparse models

Given $\{\mathbf{D}_i\}_{i=1}^{L}$, a signal $\mathbf{X} \in \mathbb{R}^N$,

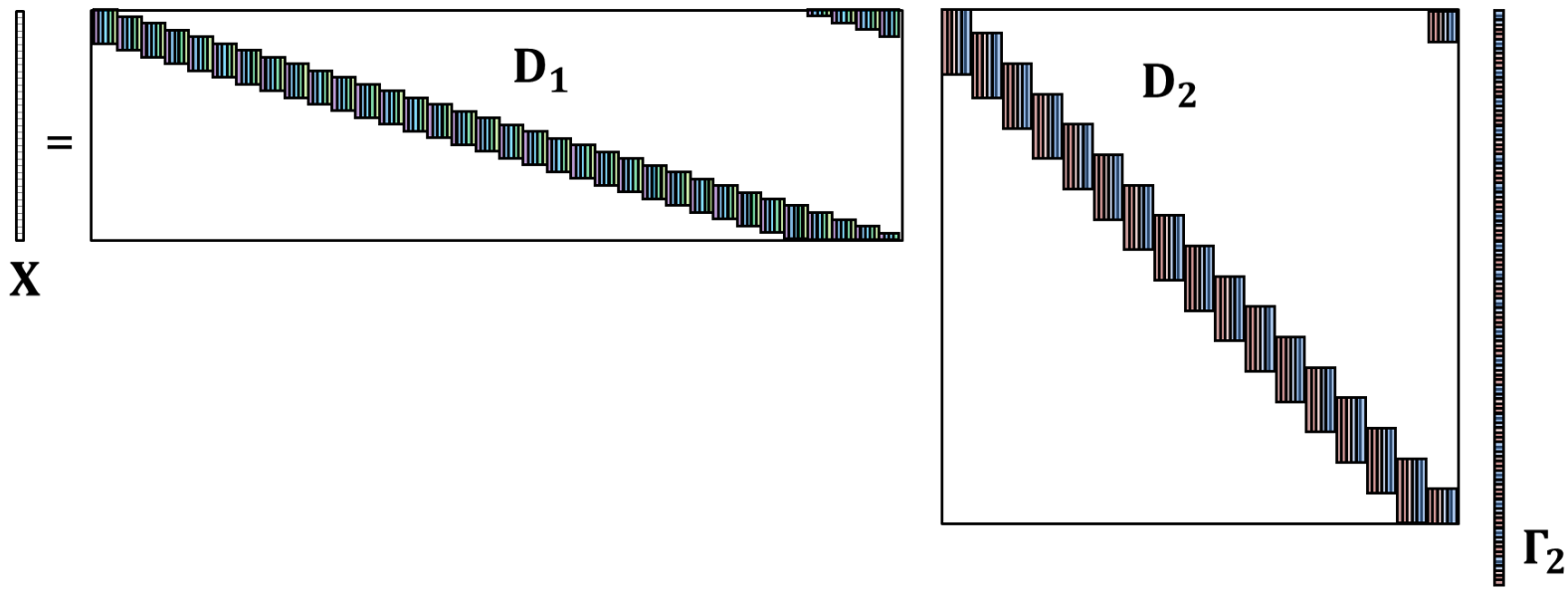
$$\mathbf{X} = \mathbf{D}_1\boldsymbol{\Gamma}_1, \quad \|\boldsymbol{\Gamma}_1\|_{0,\infty}^{s} \leq \lambda_1,$$



Multilayer (& convolutional) sparse models

Given $\{\mathbf{D}_i\}_{i=1}^{L}$, a signal $\mathbf{X} \in \mathbb{R}^N$,

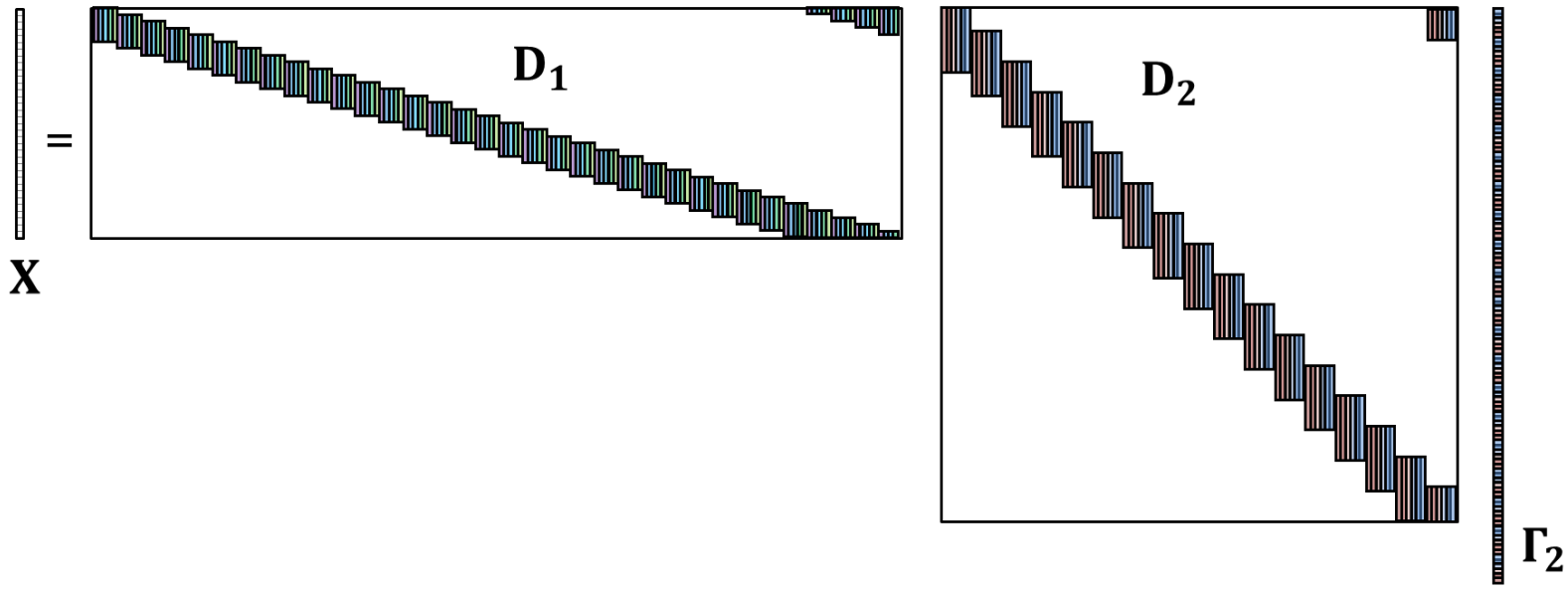
$$\mathbf{X} = \mathbf{D}_1\boldsymbol{\Gamma}_1, \quad \|\boldsymbol{\Gamma}_1\|_{0,\infty}^{s} \leq \lambda_1,$$
$$\boldsymbol{\Gamma}_1 = \mathbf{D}_2\boldsymbol{\Gamma}_2, \quad \|\boldsymbol{\Gamma}_2\|_{0,\infty}^{s} \leq \lambda_2,$$



Multilayer (& convolutional) sparse models

Given $\{\mathbf{D}_i\}_{i=1}^{L}$, a signal $\mathbf{X} \in \mathbb{R}^N$,

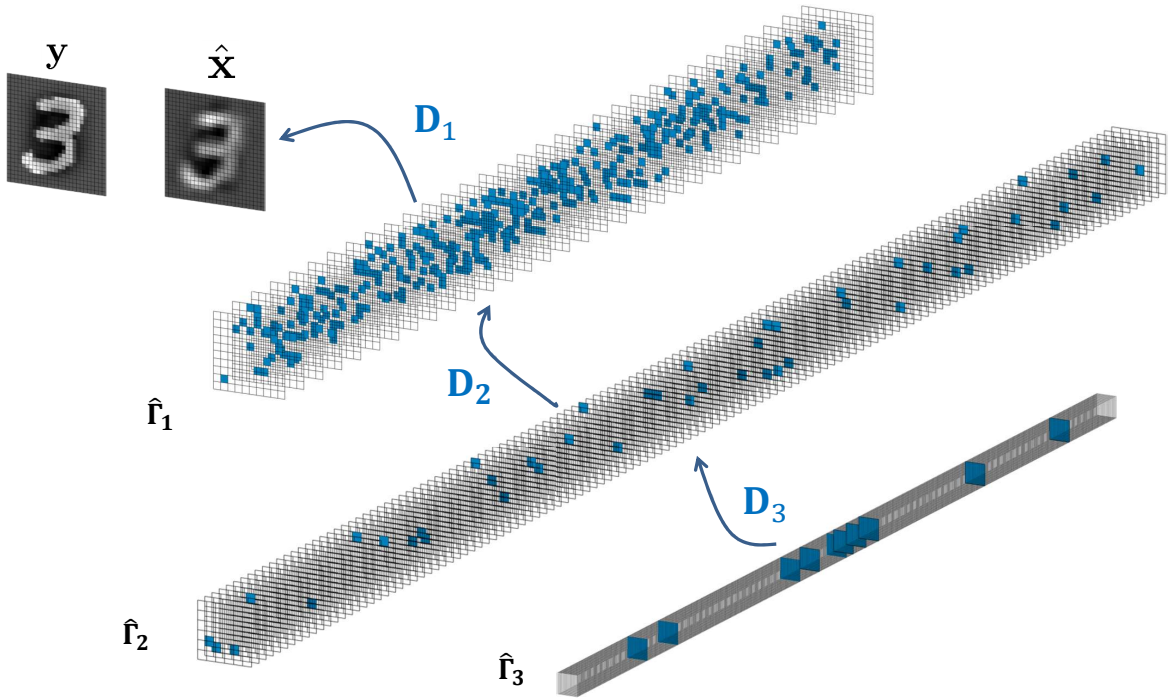
$$\begin{aligned}
\mathbf{X} &= \mathbf{D}_1\boldsymbol{\Gamma}_1, \quad \|\boldsymbol{\Gamma}_1\|_{0,\infty}^{s} \leq \lambda_1, \\
\boldsymbol{\Gamma}_1 &= \mathbf{D}_2\boldsymbol{\Gamma}_2, \quad \|\boldsymbol{\Gamma}_2\|_{0,\infty}^{s} \leq \lambda_2, \\
&\vdots \\
\boldsymbol{\Gamma}_{K-1} &= \mathbf{D}_K\boldsymbol{\Gamma}_K,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



Multilayer (& convolutional) sparse models

Given $\{\mathbf{D}_i\}_{i=1}^L$, a signal $\mathbf{X} \in \mathbb{R}^N$,

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



A Multi-Layer Pursuit

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

A Multi-Layer Pursuit

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

Deep Coding Problem

$$
\begin{aligned}
find \quad \{\mathbf{\Gamma}_i\}_{i=1}^{K} \qquad \text{s.t.} \qquad & \|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}_1\|_2^2 \le \epsilon_0, && \|\mathbf{\Gamma}_1\|_{0,\infty}^{s} \le \lambda_1 \\
& \|\mathbf{\Gamma}_1 - \mathbf{D}_2\mathbf{\Gamma}_2\|_2^2 \le \epsilon_1, && \|\mathbf{\Gamma}_2\|_{0,\infty}^{s} \le \lambda_2 \\
& \qquad\qquad \vdots && \qquad\qquad \vdots \\
& \|\mathbf{\Gamma}_{K-1} - \mathbf{D}_K\mathbf{\Gamma}_K\|_2^2 \le \epsilon_{K-1}, && \|\mathbf{\Gamma}_K\|_{0,\infty}^{s} \le \lambda_K,
\end{aligned}
$$

A Multi-Layer Pursuit

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

Deep Coding Problem

$$
\begin{aligned}
find \quad \{\mathbf{\Gamma}_i\}_{i=1}^{K} \qquad \text{s.t.} \qquad & \|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}_1\|_2^2 \le \epsilon_0, && \|\mathbf{\Gamma}_1\|_{0,\infty}^{s} \le \lambda_1 \\
& \|\mathbf{\Gamma}_1 - \mathbf{D}_2\mathbf{\Gamma}_2\|_2^2 \le \epsilon_1, && \|\mathbf{\Gamma}_2\|_{0,\infty}^{s} \le \lambda_2 \\
& \qquad\qquad \vdots && \qquad\qquad \vdots \\
& \|\mathbf{\Gamma}_{K-1} - \mathbf{D}_K\mathbf{\Gamma}_K\|_2^2 \le \epsilon_{K-1}, && \|\mathbf{\Gamma}_K\|_{0,\infty}^{s} \le \lambda_K,
\end{aligned}
$$

Consider the first layer: $\mathbf{Y} = \mathbf{D}_1\mathbf{\Gamma}_1 + \mathbf{E}$, how to find $\mathbf{\Gamma}_1$?

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

How good is this solution?

A Multi-Layer Pursuit

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

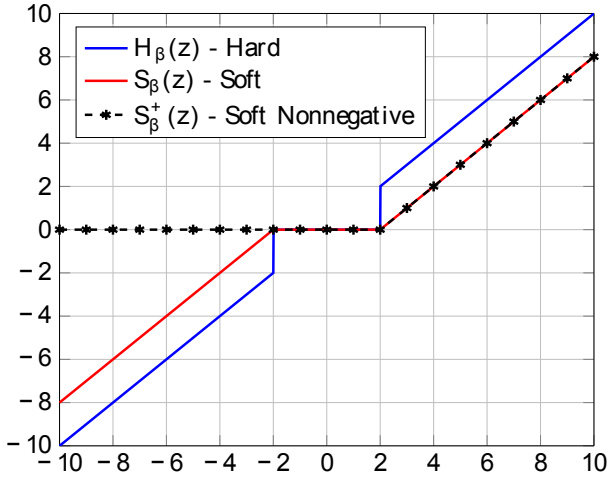
Deep Coding Problem

$$
\begin{aligned}
find \quad \{\mathbf{\Gamma}_i\}_{i=1}^{K} \qquad \text{s.t.} \qquad & \|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}_1\|_2^2 \le \epsilon_0, && \|\mathbf{\Gamma}_1\|_{0,\infty}^{s} \le \lambda_1 \\
& \|\mathbf{\Gamma}_1 - \mathbf{D}_2\mathbf{\Gamma}_2\|_2^2 \le \epsilon_1, && \|\mathbf{\Gamma}_2\|_{0,\infty}^{s} \le \lambda_2 \\
& \vdots && \vdots \\
& \|\mathbf{\Gamma}_{K-1} - \mathbf{D}_K\mathbf{\Gamma}_K\|_2^2 \le \epsilon_{K-1}, && \|\mathbf{\Gamma}_K\|_{0,\infty}^{s} \le \lambda_K,
\end{aligned}
$$

Consider the first layer: $\mathbf{Y} = \mathbf{D}_1\mathbf{\Gamma}_1 + \mathbf{E}$, how to find $\mathbf{\Gamma}_1$?

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

How good is this solution?



A Multi-Layer Pursuit

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

Deep Coding Problem

$$
\begin{aligned}
find \quad \{\mathbf{\Gamma}_i\}_{i=1}^{K} \qquad \text{s.t.} \qquad & \|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}_1\|_2^2 \le \epsilon_0, && \|\mathbf{\Gamma}_1\|_{0,\infty}^{s} \le \lambda_1 \\
& \|\mathbf{\Gamma}_1 - \mathbf{D}_2\mathbf{\Gamma}_2\|_2^2 \le \epsilon_1, && \|\mathbf{\Gamma}_2\|_{0,\infty}^{s} \le \lambda_2 \\
& \qquad\vdots && \qquad\vdots \\
& \|\mathbf{\Gamma}_{K-1} - \mathbf{D}_K\mathbf{\Gamma}_K\|_2^2 \le \epsilon_{K-1}, && \|\mathbf{\Gamma}_K\|_{0,\infty}^{s} \le \lambda_K,
\end{aligned}
$$

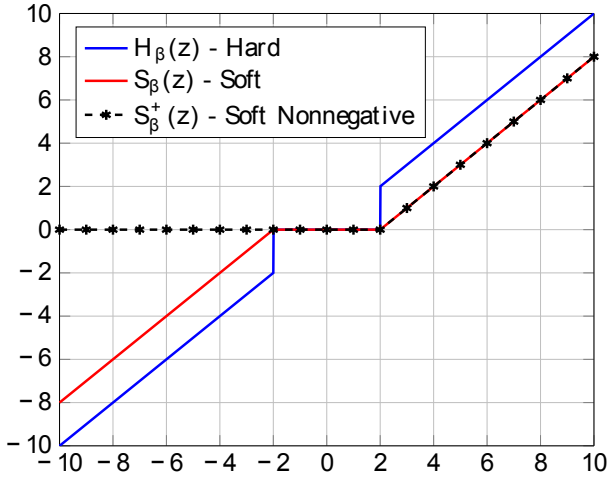
Consider the first layer: $\mathbf{Y} = \mathbf{D}_1\mathbf{\Gamma}_1 + \mathbf{E}$, how to find $\mathbf{\Gamma}_1$?

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

How good is this solution?

Consider then $\hat{\mathbf{\Gamma}}_1 = \mathbf{D}_2\mathbf{\Gamma}_2 + \mathbf{E}_1$, how to find $\mathbf{\Gamma}_2$?

Second Layer pursuit: $\hat{\mathbf{\Gamma}}_2 \leftarrow \mathcal{P}_{\beta_2}(\mathbf{D}_2^T\hat{\mathbf{\Gamma}}_1)$



Layered Thresholding algorithm

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

Layered Thresholding algorithm

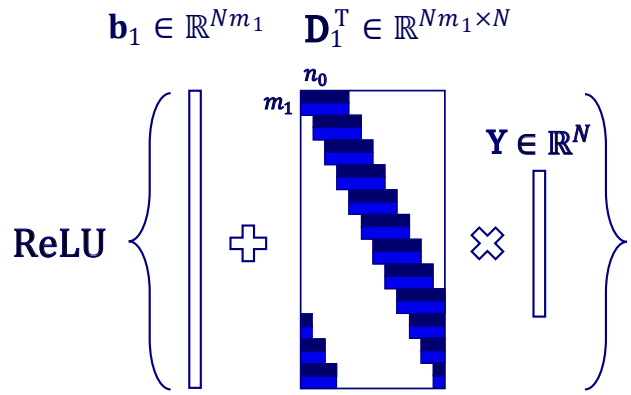
$$\hat{\mathbf{\Gamma}}_2 = \mathcal{P}_{\beta_2}(\mathbf{D}_2^T \mathcal{P}_{\beta_1}(\mathbf{D}_1^T \mathbf{Y}))$$

Layered Thresholding algorithm

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

Written differently,

$$\text{ReLu}(\mathbf{D}_1^T \mathbf{Y} + \mathbf{b}_1)$$

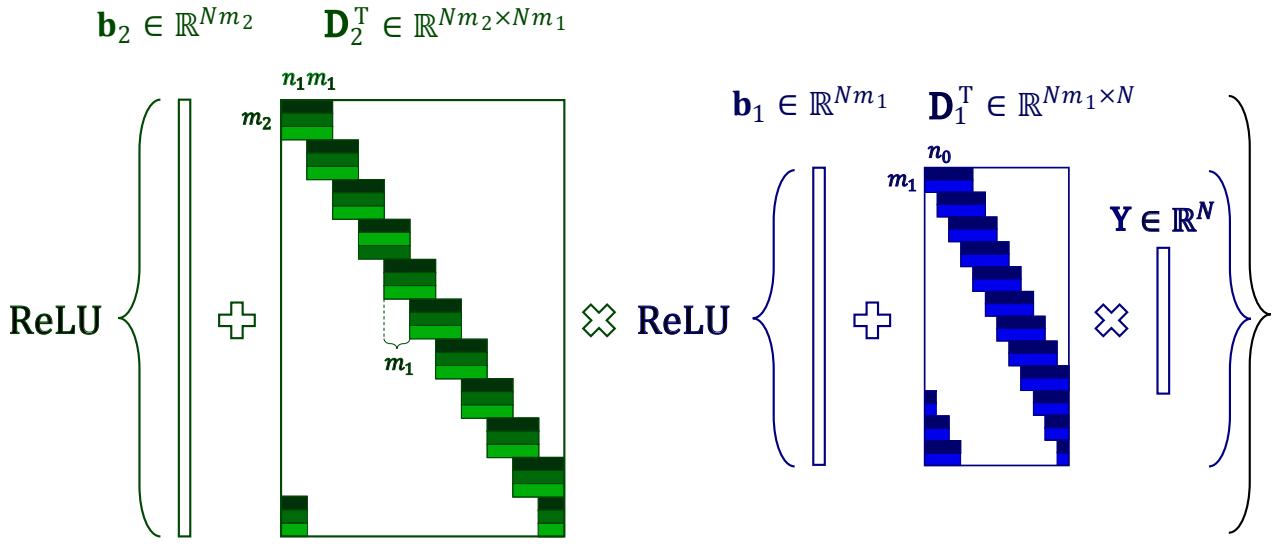


Layered Thresholding algorithm

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

Written differently,

$$\text{ReLu}(\mathbf{D}_2^T \ \text{ReLu}(\mathbf{D}_1^T \mathbf{Y} + \mathbf{b}_1) + \mathbf{b}_2)$$

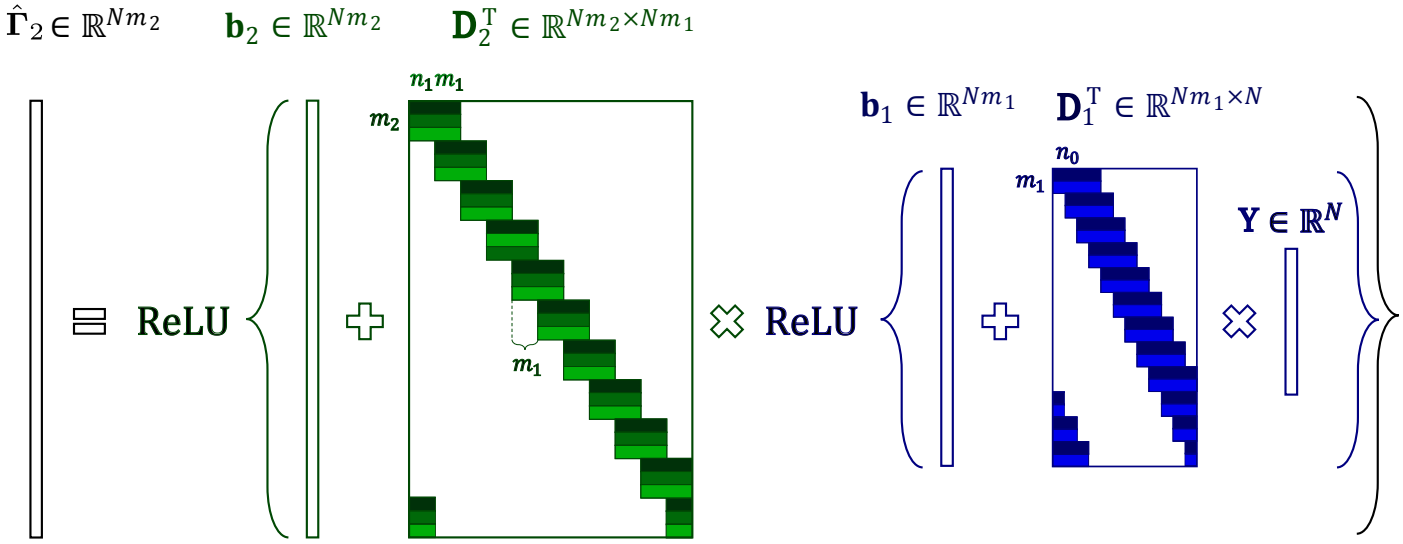


Layered Thresholding algorithm

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

Written differently,

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

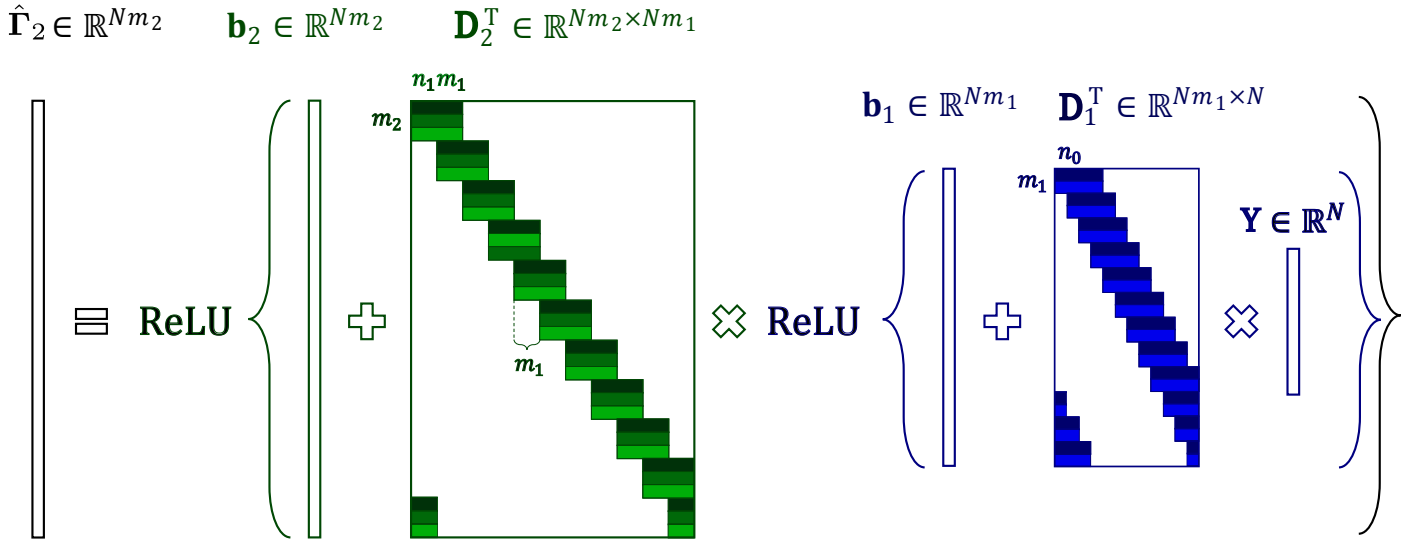


Layered Thresholding algorithm

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

Written differently,

$$\hat{\mathbf{\Gamma}}_2 = \text{ReLu}(\mathbf{D}_2^T \ \text{ReLu}(\mathbf{D}_1^T \mathbf{Y} + \mathbf{b}_1) + \mathbf{b}_2) \qquad \textbf{Forward Pass of CNN}$$



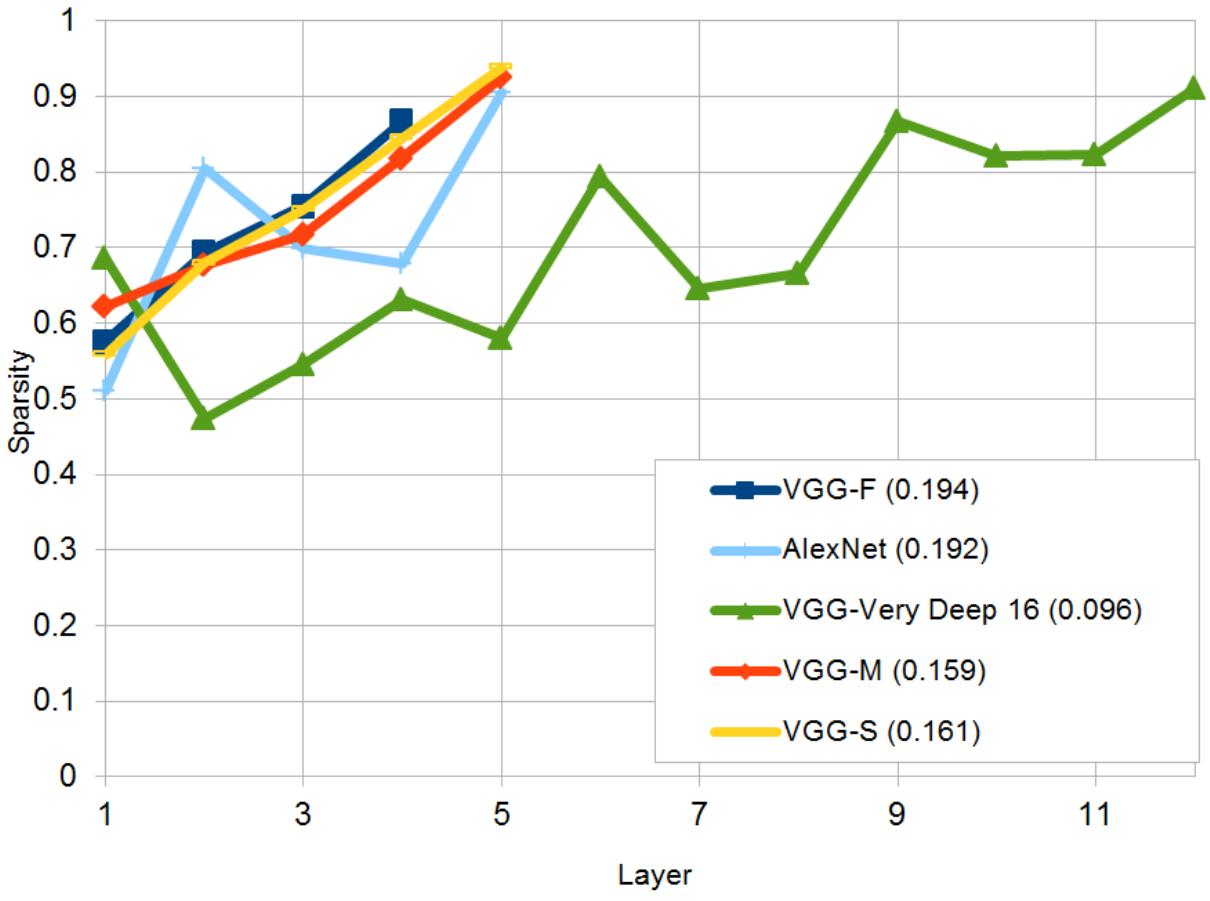
The forward pass can be seen as a pursuit seeking for the sparse representations under the ML-CSC model

Looking into the Networks

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

Looking into the Networks

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



Stability of the Multi-Layer Thresholding (a.k.a Forward Pass)

If a set of solutions $\{\mathbf{\Gamma}_i\}_{i=1}^K$ satisfy

$$\|\mathbf{\Gamma}_i\|_{0,\infty} \leq \frac{1}{2}\left(1+\frac{1}{\mu(\mathbf{D}_i)}\frac{|\mathbf{\Gamma}_i^{min}|}{|\mathbf{\Gamma}_i^{max}|}\right) - \frac{1}{\mu(\mathbf{D}_i)}\frac{\epsilon_L^{i-1}}{|\mathbf{\Gamma}_i^{max}|}$$

then, for proper thresholds/biases β_i^*

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

✓ Recovery of the support

✓ Stable recovery of representations

Stability of the Multi-Layer Thresholding (a.k.a Forward Pass)

If a set of solutions $\{\mathbf{\Gamma}_i\}_{i=1}^K$ satisfy

$$\|\mathbf{\Gamma}_i\|_{0,\infty} \leq \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}_i)}\frac{|\mathbf{\Gamma}_i^{min}|}{|\mathbf{\Gamma}_i^{max}|}\right) - \frac{1}{\mu(\mathbf{D}_i)}\frac{\epsilon_L^{i-1}}{|\mathbf{\Gamma}_i^{max}|}$$

then, for proper thresholds/biases β_i^*

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

✓ Recovery of the support
✓ Stable recovery of representations

× Signal contrast
× No perfect recovery in noiseless case
× Bounds increase with depth

Layered Basis Pursuit

$$\hat{\mathbf{\Gamma}}_1 = \arg\min_{\mathbf{\Gamma}} \quad \frac{1}{2}\|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}\|_2^2 + \lambda_1\|\mathbf{\Gamma}\|_1$$

$$\hat{\mathbf{\Gamma}}_2 = \arg\min_{\mathbf{\Gamma}} \quad \frac{1}{2}\|\hat{\mathbf{\Gamma}}_1 - \mathbf{D}_2\mathbf{\Gamma}\|_2^2 + \lambda_2\|\mathbf{\Gamma}\|_1$$

$$\vdots$$

$$\hat{\mathbf{\Gamma}}_L = \arg\min_{\mathbf{\Gamma}} \quad \frac{1}{2}\|\hat{\mathbf{\Gamma}}_{L-1} - \mathbf{D}_L\mathbf{\Gamma}\|_2^2 + \lambda_L\|\mathbf{\Gamma}\|_1$$

Layered Basis Pursuit

$$\hat{\mathbf{\Gamma}}_1 = \arg\min_{\mathbf{\Gamma}} \quad \frac{1}{2}\|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}\|_2^2 + \lambda_1\|\mathbf{\Gamma}\|_1$$

$$\hat{\mathbf{\Gamma}}_2 = \arg\min_{\mathbf{\Gamma}} \quad \frac{1}{2}\|\hat{\mathbf{\Gamma}}_1 - \mathbf{D}_2\mathbf{\Gamma}\|_2^2 + \lambda_2\|\mathbf{\Gamma}\|_1$$

$$\vdots$$

$$\hat{\mathbf{\Gamma}}_L = \arg\min_{\mathbf{\Gamma}} \quad \frac{1}{2}\|\hat{\mathbf{\Gamma}}_{L-1} - \mathbf{D}_L\mathbf{\Gamma}\|_2^2 + \lambda_L\|\mathbf{\Gamma}\|_1$$

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

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

Layered Basis Pursuit

$$\hat{\mathbf{\Gamma}}_1 = \arg\min_{\mathbf{\Gamma}} \quad \frac{1}{2}\|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}\|_2^2 + \lambda_1\|\mathbf{\Gamma}\|_1$$

$$\hat{\mathbf{\Gamma}}_2 = \arg\min_{\mathbf{\Gamma}} \quad \frac{1}{2}\|\hat{\mathbf{\Gamma}}_1 - \mathbf{D}_2\mathbf{\Gamma}\|_2^2 + \lambda_2\|\mathbf{\Gamma}\|_1$$

$$\vdots$$

$$\hat{\mathbf{\Gamma}}_L = \arg\min_{\mathbf{\Gamma}} \quad \frac{1}{2}\|\hat{\mathbf{\Gamma}}_{L-1} - \mathbf{D}_L\mathbf{\Gamma}\|_2^2 + \lambda_L\|\mathbf{\Gamma}\|_1$$

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

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

× Bound increases with depth

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

Given $\mathbf{Y} = \mathbf{X}(\boldsymbol{\Gamma}_i) + \mathbf{E}$, $\mathbf{X} \in \mathcal{M}_{\boldsymbol{\lambda}}$ and convolutional dictionaries $\{\mathbf{D}_i\}_{i=1}^K$,

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

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

Given $\mathbf{Y} = \mathbf{X}(\mathbf{\Gamma}_i) + \mathbf{E}$, $\quad \mathbf{X} \in \mathcal{M}_{\boldsymbol{\lambda}}$ and convolutional dictionaries $\{\mathbf{D}_i\}_{i=1}^K$,

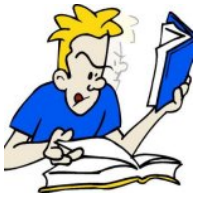
$$\mathcal{P}_{\mathcal{M}_{\boldsymbol{\lambda}}}: \quad \min_{\{\mathbf{\Gamma}_i\}} \quad \|\mathbf{Y} - \mathbf{X}(\mathbf{\Gamma}_i)\|_2 \quad \text{s.t.} \quad \mathbf{X}(\mathbf{\Gamma}_i) \in \mathcal{M}_{\boldsymbol{\lambda}}.$$

Stability of the $P_{\mathcal{M}}$ problem

If $\|\mathbf{E}\|_2 \leq \mathcal{E}_0$, and $\|\mathbf{\Gamma}_i\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}^{(i)})}\right)$, for $1 \leq i \leq K$, then

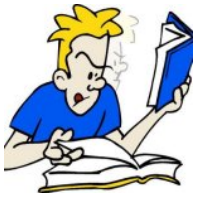
$$\|\mathbf{\Gamma}_i - \hat{\mathbf{\Gamma}}_i\|_2^2 \leq \frac{4\mathcal{E}_0^2}{1 - (2\|\mathbf{\Gamma}_i\|_{0,\infty} - 1)\mu(\mathbf{D}^{(i)})}$$

- ✓ Bound is not cumulative across layers
- ✓ Dependence on $\mu(\mathbf{D}^{(L)})$ - Global effective dictionary



How to Learn?

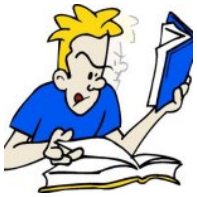
The existence of $\mathbf{X}(\mathbf{\Gamma}_i) \in \mathcal{M}_{\boldsymbol{\lambda}}$ depends on proper dictionaries $\mathbf{D}_i$.



How to Learn?

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

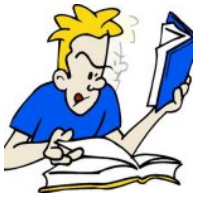
$$\min_{\{\mathbf{\Gamma}_i^t\},\{\mathbf{D}_i\}} \quad \sum_{t=1}^{T} \|\mathbf{Y}^t - \mathbf{D}_1 \dots \mathbf{D}_L \mathbf{\Gamma}_L\|_2^2 \quad \text{s.t.} \quad \|\mathbf{\Gamma}_i\|_{0,\infty} \leq s_i \ \forall \ i$$



How to Learn?

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

$$\min_{\{\mathbf{\Gamma}_i^t\},\{\mathbf{D}_i\}} \quad \sum_{t=1}^{T} \|\mathbf{Y}^t - \mathbf{D}_1 \dots \mathbf{D}_L \mathbf{\Gamma}_L\|_2^2 \quad \text{s.t.} \quad \|\mathbf{\Gamma}_i\|_{0,\infty} \leq s_i \ \forall \ i$$



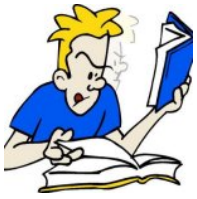
How to Learn?

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

$$\min_{\{\mathbf{\Gamma}_i^t\},\{\mathbf{D}_i\}} \quad \sum_{t=1}^{T} \|\mathbf{Y}^t - \mathbf{D}_1 \dots \mathbf{D}_L \mathbf{\Gamma}_L\|_2^2 \quad \text{s.t.} \quad \|\mathbf{\Gamma}_i\|_{0,\infty} \leq s_i \ \forall \ i$$

Sparsity Proxies

$$\mathbf{\Gamma}_{L-1} = \mathbf{D}_L \mathbf{\Gamma}_L. \qquad \Rightarrow \|\mathbf{\Gamma}_{K-1}\|_{0,\infty}^s \leq c_L \|\mathbf{D}_L\|_0 \|\mathbf{\Gamma}_L\|_{0,\infty}^s$$



How to Learn?

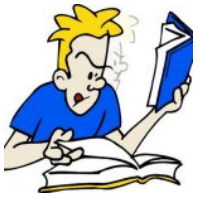
The existence of $\mathbf{X}(\mathbf{\Gamma}_i) \in \mathcal{M}_{\boldsymbol{\lambda}}$ depends on proper dictionaries $\mathbf{D}_i$.

$$\min_{\{\mathbf{\Gamma}_i^t\},\{\mathbf{D}_i\}} \quad \sum_{t=1}^{T} \|\mathbf{Y}^t - \mathbf{D}_1 \dots \mathbf{D}_L \mathbf{\Gamma}_L\|_2^2 \quad \text{s.t.} \quad \|\mathbf{\Gamma}_i\|_{0,\infty} \leq s_i \ \forall \ i$$

Sparsity Proxies

$$\mathbf{\Gamma}_{L-1} = \mathbf{D}_L \mathbf{\Gamma}_L. \qquad \Rightarrow \|\mathbf{\Gamma}_{K-1}\|_{0,\infty}^s \leq c_L \|\mathbf{D}_L\|_0 \|\mathbf{\Gamma}_L\|_{0,\infty}^s$$

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



How to Learn?

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

$$\min_{\{\mathbf{\Gamma}_i^t\},\{\mathbf{D}_i\}} \quad \sum_{t=1}^{T} \|\mathbf{Y}^t - \mathbf{D}_1 \dots \mathbf{D}_L \mathbf{\Gamma}_L\|_2^2 \quad \text{s.t.} \quad \|\mathbf{\Gamma}_i\|_{0,\infty} \leq s_i \ \forall \ i$$

Sparsity Proxies

$$\mathbf{\Gamma}_{L-1} = \mathbf{D}_L \mathbf{\Gamma}_L. \qquad \Rightarrow \|\mathbf{\Gamma}_{K-1}\|_{0,\infty}^s \leq c_L \|\mathbf{D}_L\|_0 \|\mathbf{\Gamma}_L\|_{0,\infty}^s$$

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

Problem formulation

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

Multi-Layer Convolutional Dictionary Learning

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

Multi-Layer Convolutional Dictionary Learning

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

Algorithm

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

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

Draw $\mathbf{Y}^t$ at random

Sparse Coding: $\hat{\mathbf{\Gamma}}_L \leftarrow \arg\min_{\mathbf{\Gamma}} \|\mathbf{Y}^t - \mathbf{D}^{(K)}\mathbf{\Gamma}\|_2$ s.t. $\|\mathbf{\Gamma}\|_{0,\infty}^s \leq \lambda_L$

Multi-Layer Convolutional Dictionary Learning

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

Algorithm

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

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

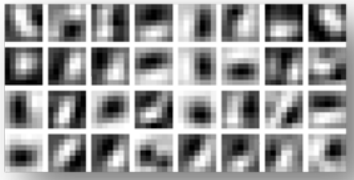
- Draw $\mathbf{Y}^t$ at random
- Sparse Coding: $\hat{\mathbf{\Gamma}}_L \leftarrow \arg\min_{\mathbf{\Gamma}} \|\mathbf{Y}^t - \mathbf{D}^{(K)}\mathbf{\Gamma}\|_2$ s.t. $\|\mathbf{\Gamma}\|_{0,\infty}^s \leq \lambda_L$
- Update Dictionaries:
- **for** $l = 1, \ldots, L$ **do**
  - $\mathbf{D}_l \leftarrow \arg\min_{\mathbf{D}_l} \|\mathbf{Y}^t - \mathbf{D}_1 \ldots \mathbf{D}_l \ldots \mathbf{D}_L\mathbf{\Gamma}_L\|_2 + \zeta_L\|\mathbf{D}_l\|_0$
- **end**

end

Learning Multi-Layer CSC Models

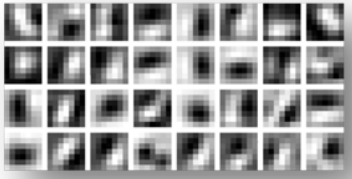
Learning Multi-Layer CSC Models

$\mathbf{D}_1$

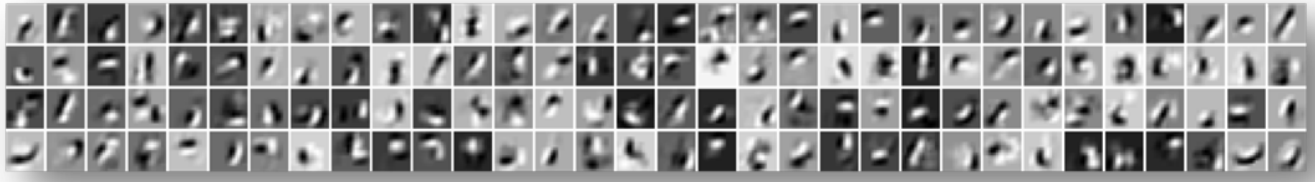


Learning Multi-Layer CSC Models

$\mathbf{D}_1$

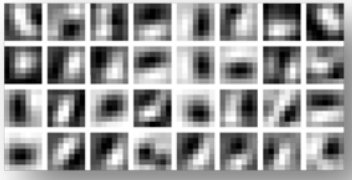


$\mathbf{D}_1\mathbf{D}_2$

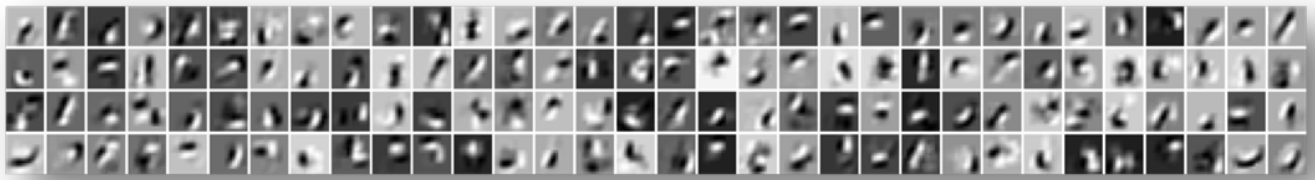


Learning Multi-Layer CSC Models

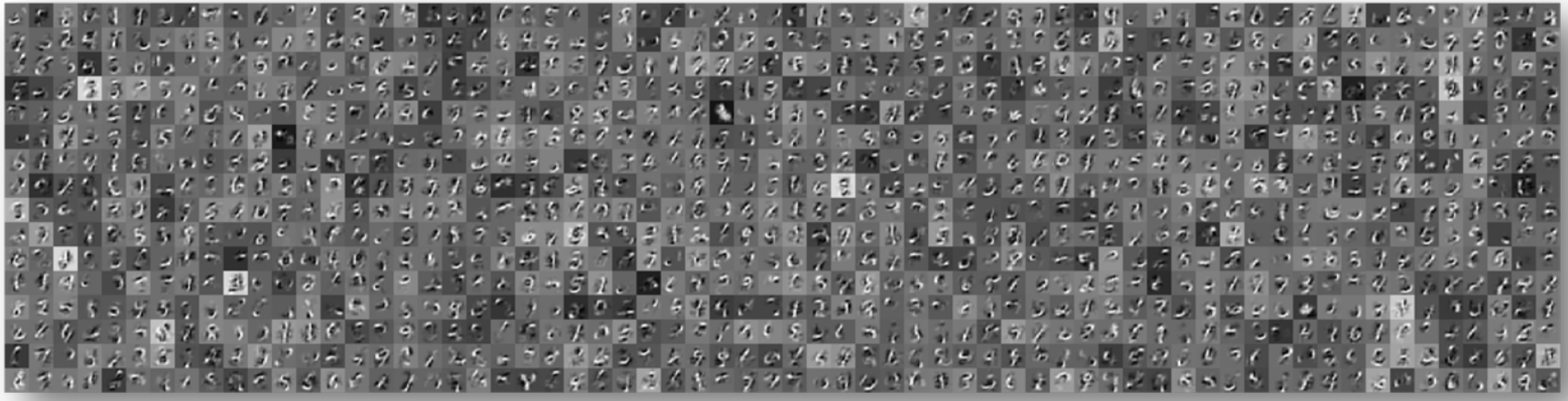
$\mathbf{D}_1$



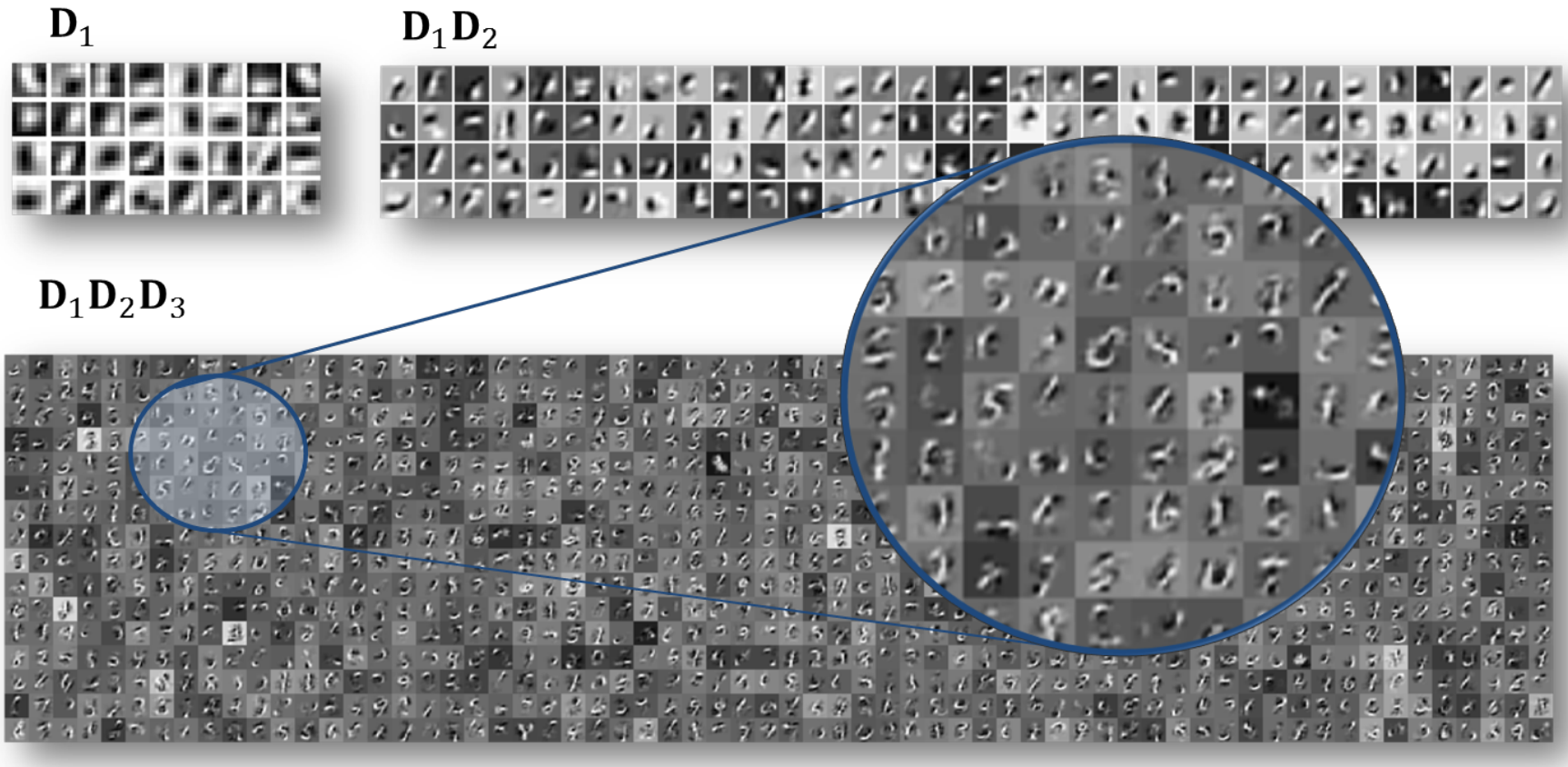
$\mathbf{D}_1\mathbf{D}_2$



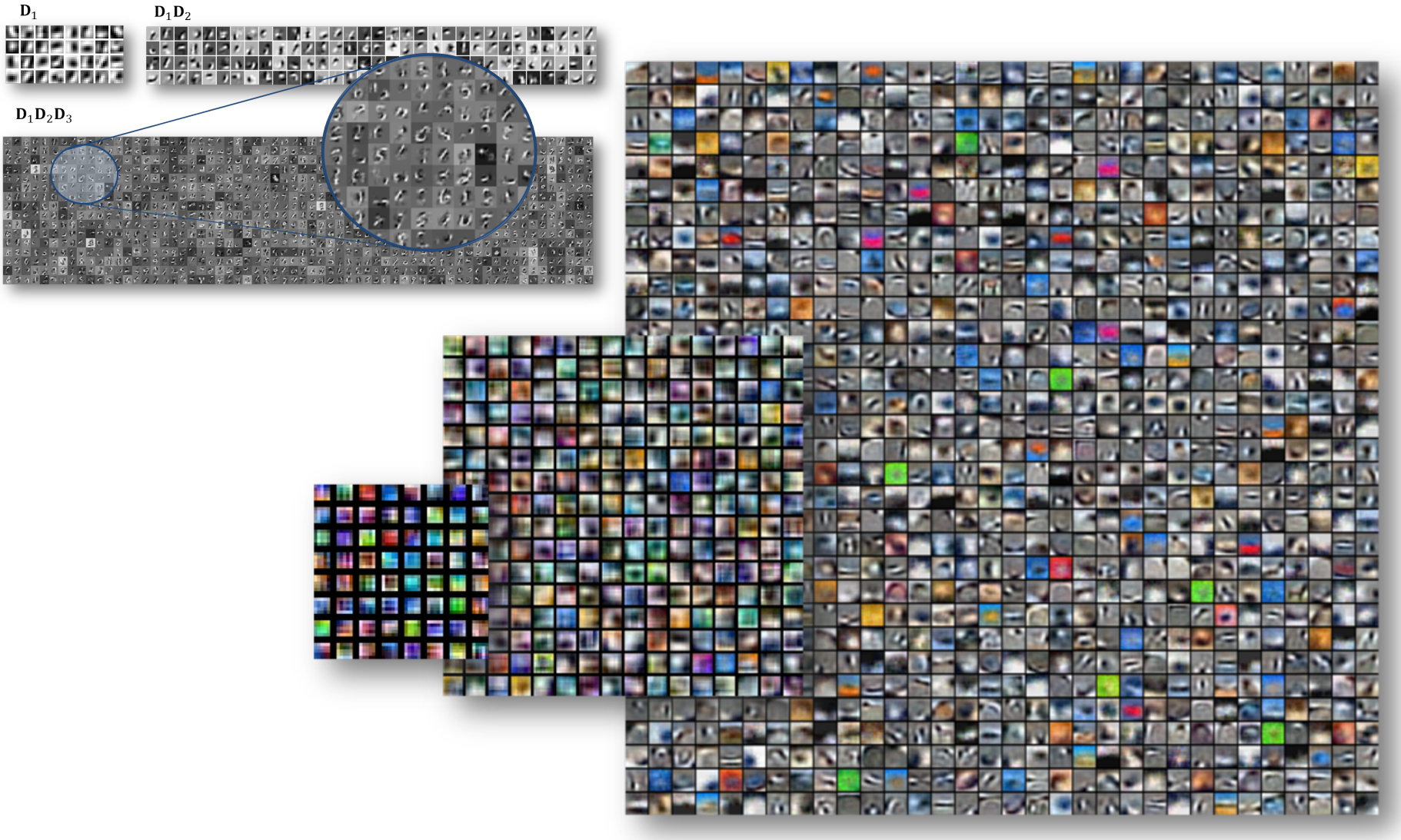
$\mathbf{D}_1\mathbf{D}_2\mathbf{D}_3$



Learning Multi-Layer CSC Models

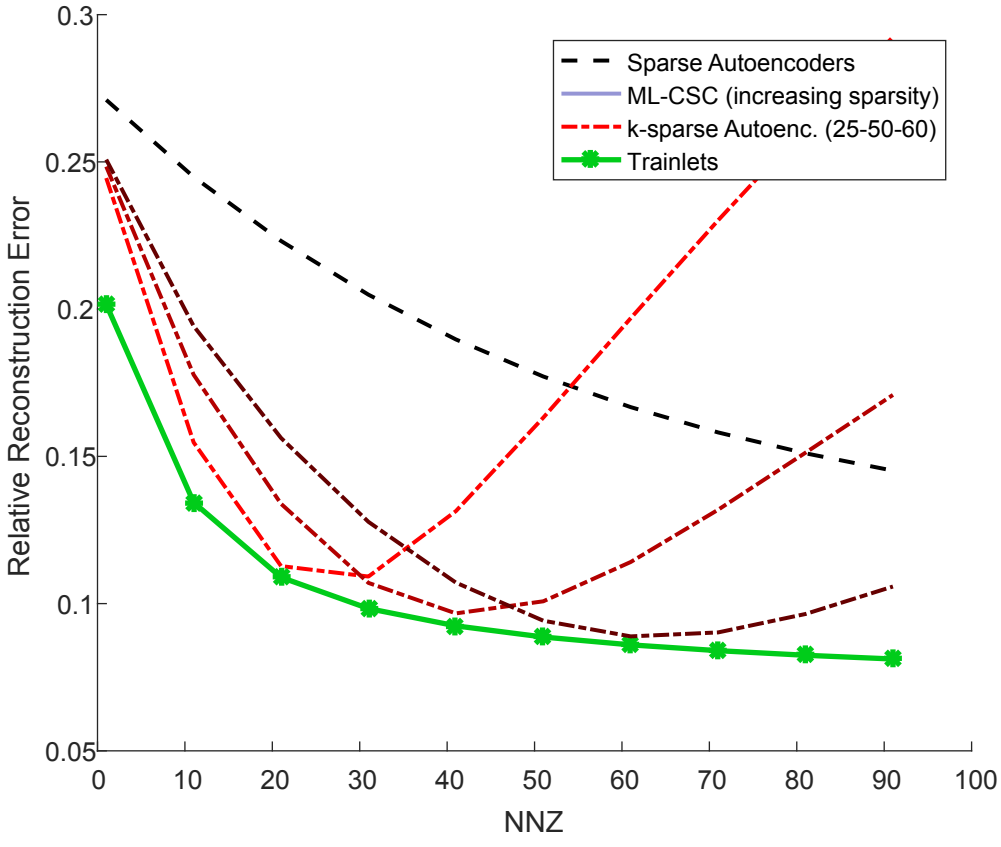


Learning Multi-Layer CSC Models



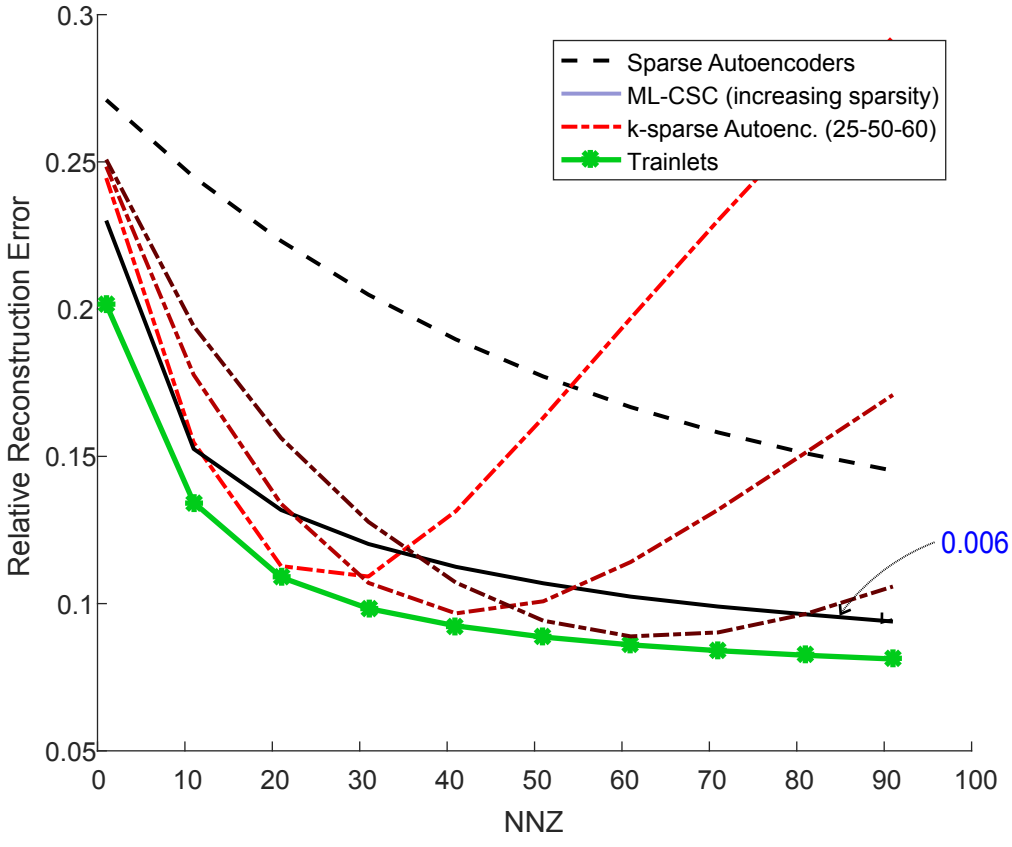
ML-CSC Decompositions

Signal Approximation:



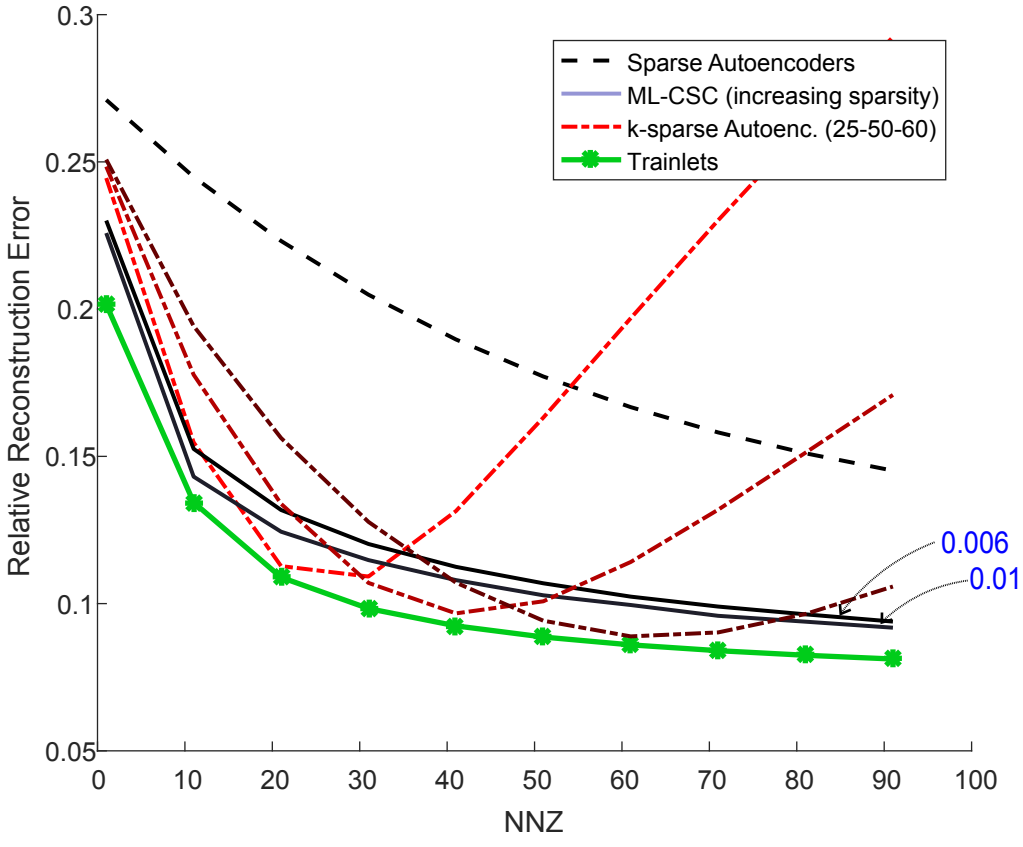
ML-CSC Decompositions

Signal Approximation:



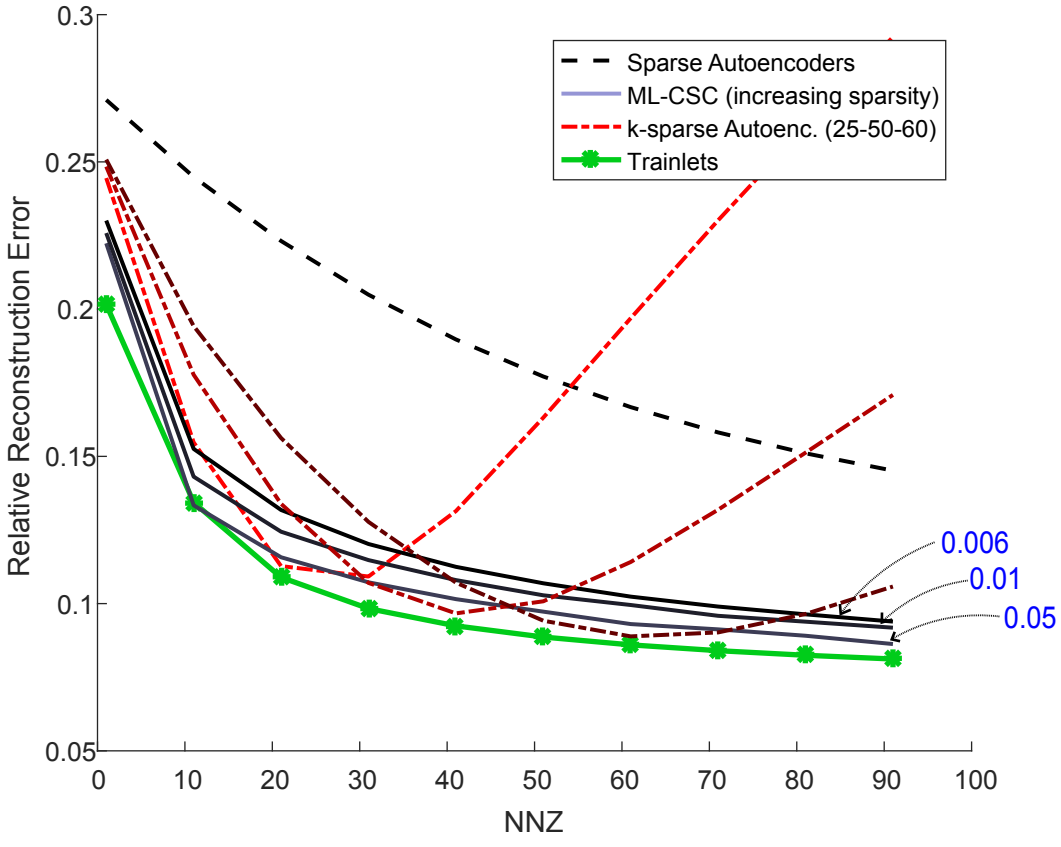
ML-CSC Decompositions

Signal Approximation:



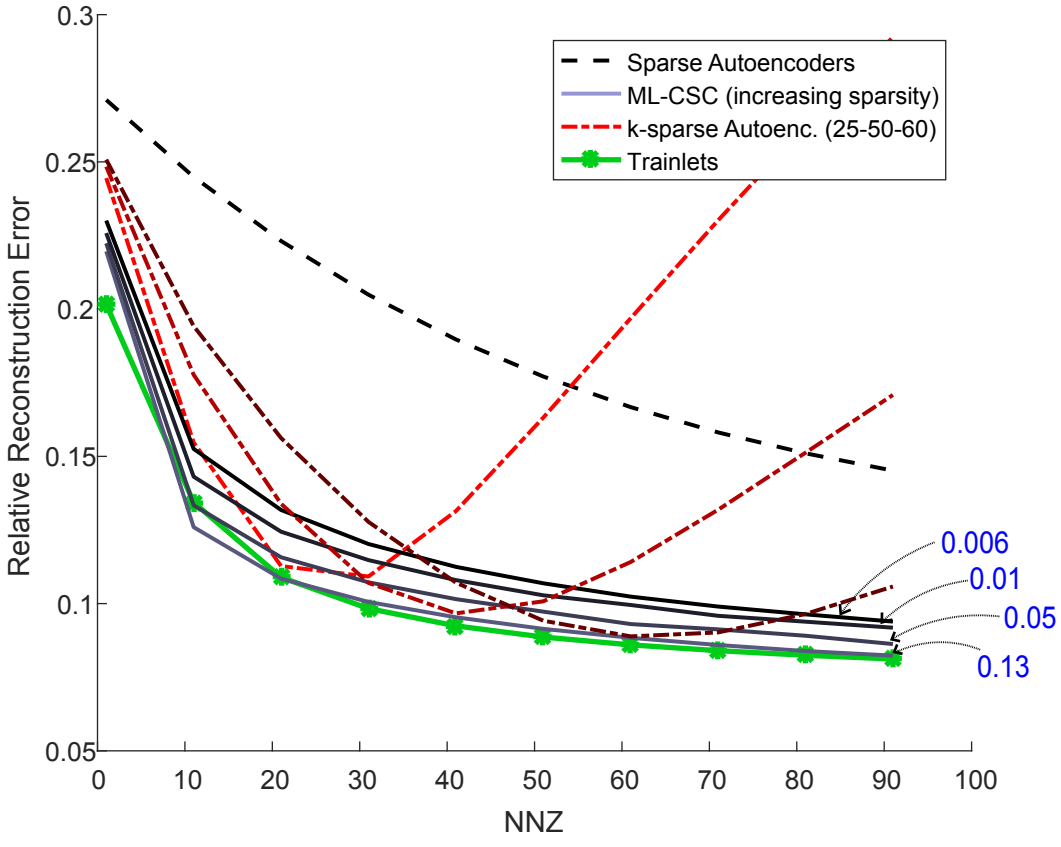
ML-CSC Decompositions

Signal Approximation:



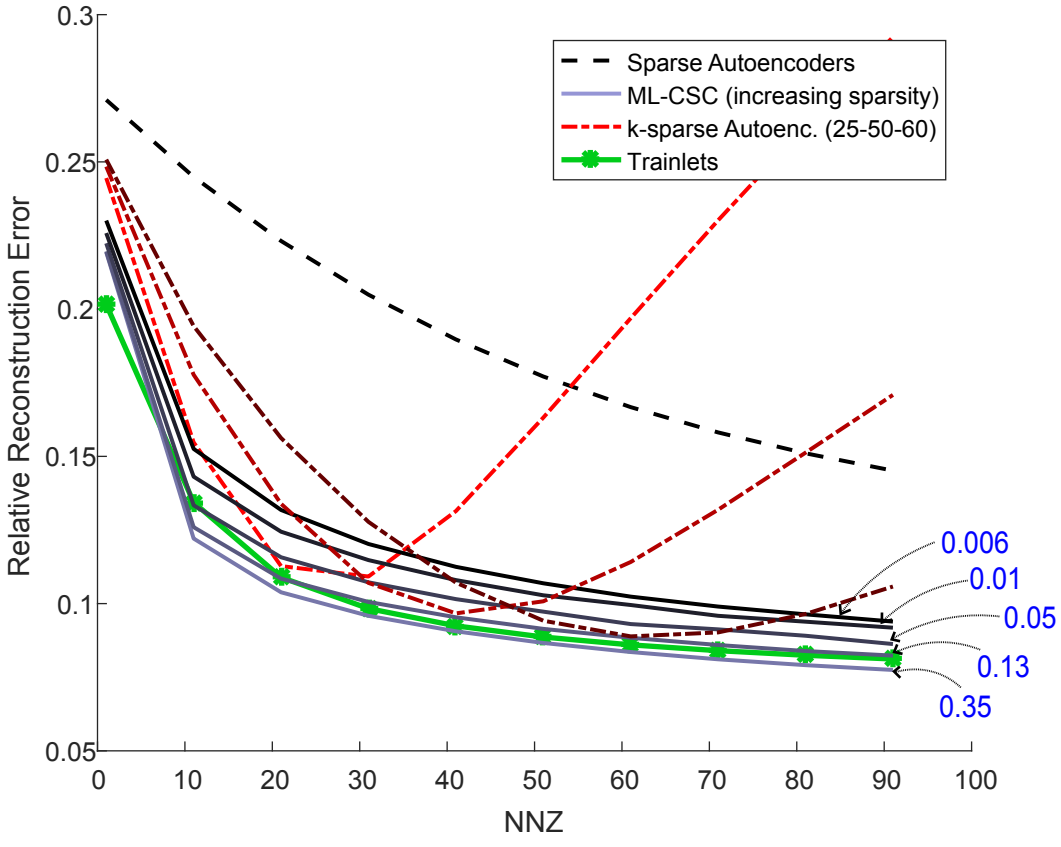
ML-CSC Decompositions

Signal Approximation:



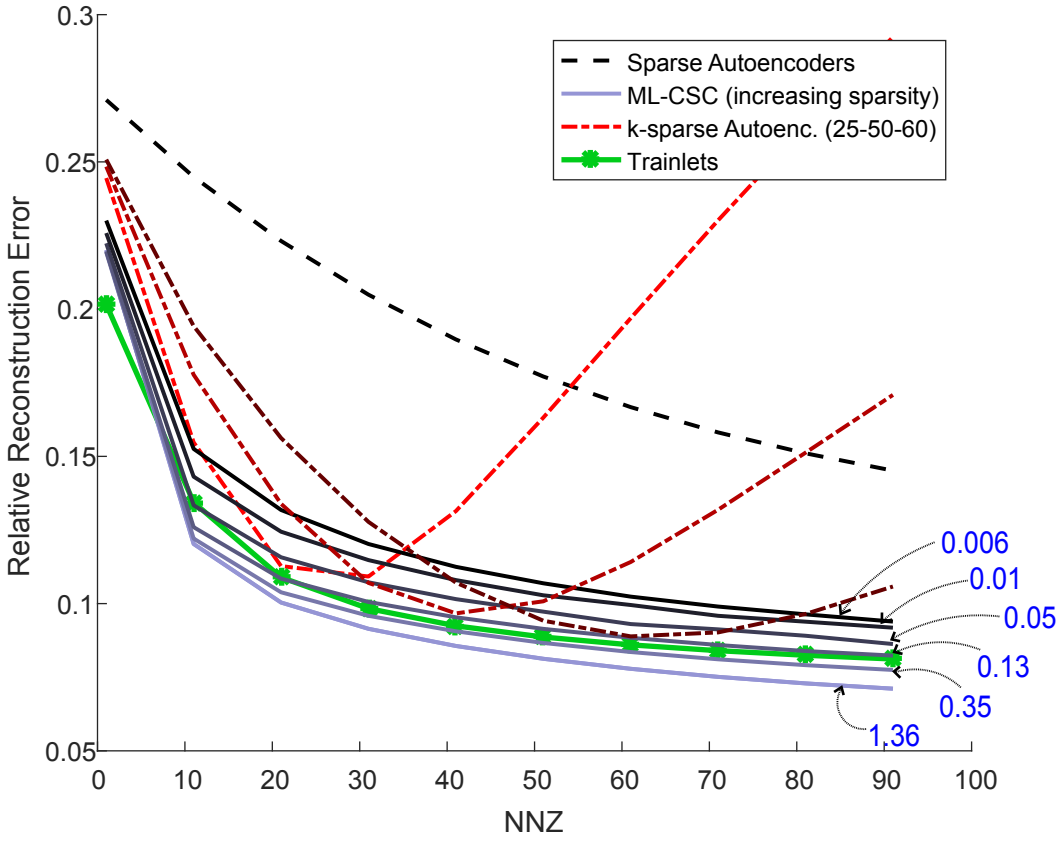
ML-CSC Decompositions

Signal Approximation:



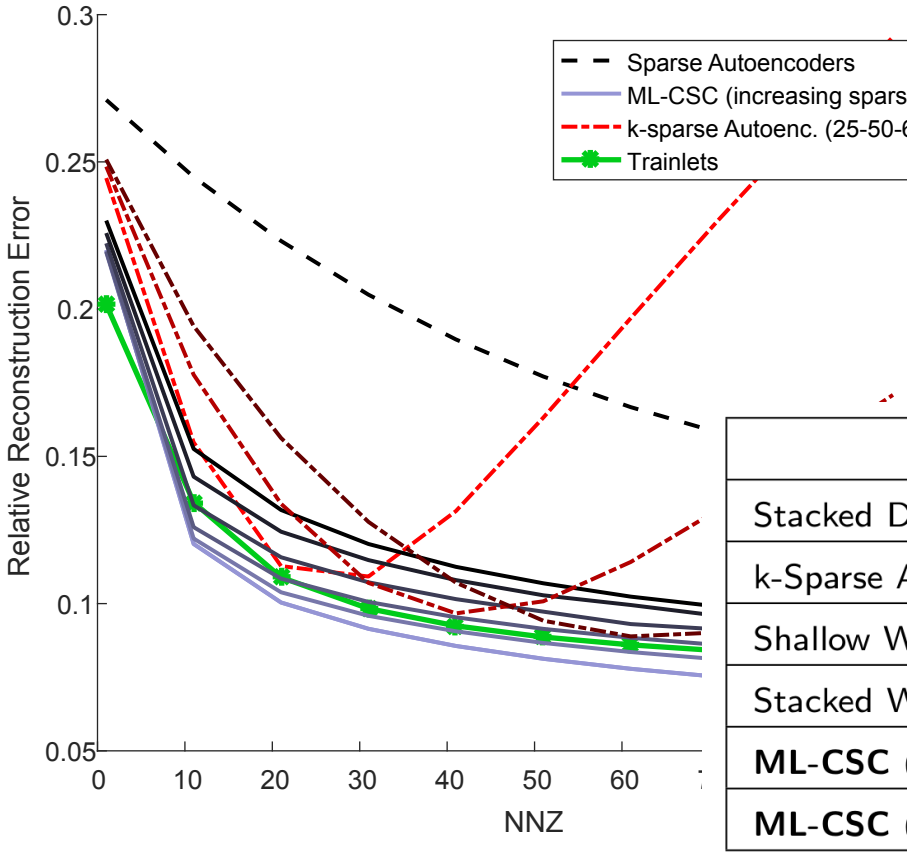
ML-CSC Decompositions

Signal Approximation:



ML-CSC Decompositions

Signal Approximation:



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%
ML-CSC (1K units) - 2nd Layer Rep.	1.30%
ML-CSC (2K units) - 2nd&3rd Layer Rep.	1.15%

Multi-Layer Sparse Modeling of Signals

Model

$$\mathbf{y} = \mathbf{D}_1\boldsymbol{\gamma}_1 + \mathbf{w}, \qquad \{\boldsymbol{\gamma}_{i-1} = \mathbf{D}_i\boldsymbol{\gamma}_i, \quad \|\boldsymbol{\gamma}_i\|_0 \leq s_i\}_{i=1}^{L}$$

Multi-Layer Sparse Modeling of Signals

Model

$$\mathbf{y} = \mathbf{D}_{(1,L)}\boldsymbol{\gamma}_L + \mathbf{w}, \qquad \{\boldsymbol{\gamma}_{i-1} = \mathbf{D}_i\boldsymbol{\gamma}_i, \quad \|\boldsymbol{\gamma}_i\|_0 \leq s_i\}_{i=1}^{L}$$

$$\mathbf{D}_{(1,L)} = \mathbf{D}_1\mathbf{D}_2\ldots\mathbf{D}_L.$$

Multi-Layer Sparse Modeling of Signals

Model

$$\mathbf{y} = \mathbf{D}_{(1,L)}\boldsymbol{\gamma}_L + \mathbf{w}, \qquad \|\boldsymbol{\gamma}_L\|_0 \leq s_L,\ \{\|\mathbf{D}_{(i,L)}\boldsymbol{\gamma}_L\|_0 \leq s_{i-1}\}_{i=1}^{L}$$

$$\mathbf{D}_{(1,L)} = \mathbf{D}_1\mathbf{D}_2\ldots\mathbf{D}_L.$$

Multi-Layer Sparse Modeling of Signals

Model

$$\mathbf{y} = \mathbf{D}_{(1,L)}\boldsymbol{\gamma}_L + \mathbf{w}, \qquad \|\boldsymbol{\gamma}_L\|_0 \le s_L,\ \left\{\|\mathbf{D}_{(i,L)}\boldsymbol{\gamma}_L\|_0 \le s_{i-1}\right\}_{i=1}^{L}$$

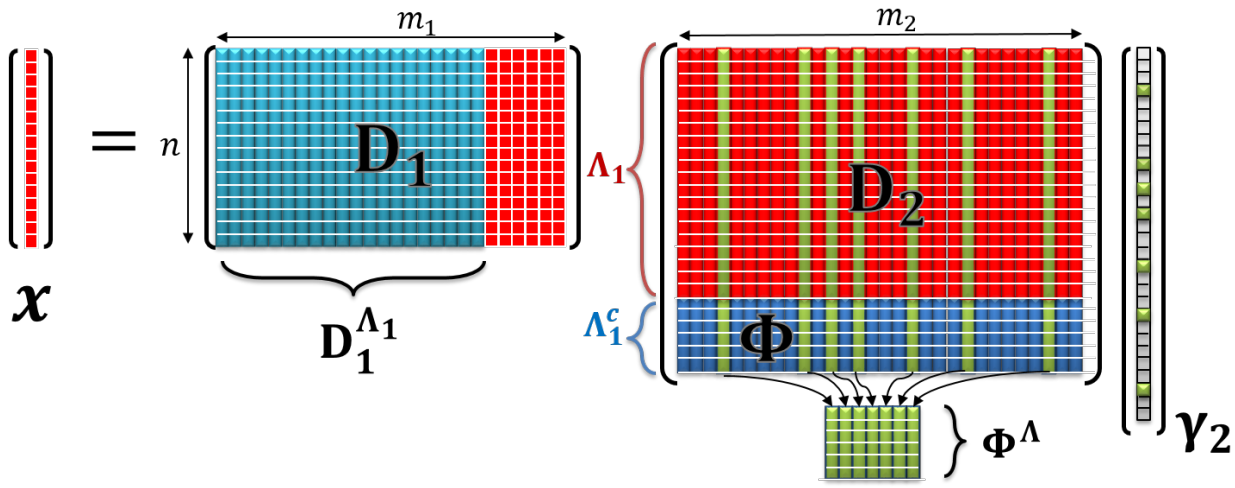
$$\mathbf{D}_{(1,L)} = \mathbf{D}_1\mathbf{D}_2\ldots\mathbf{D}_L.$$

Pursuit

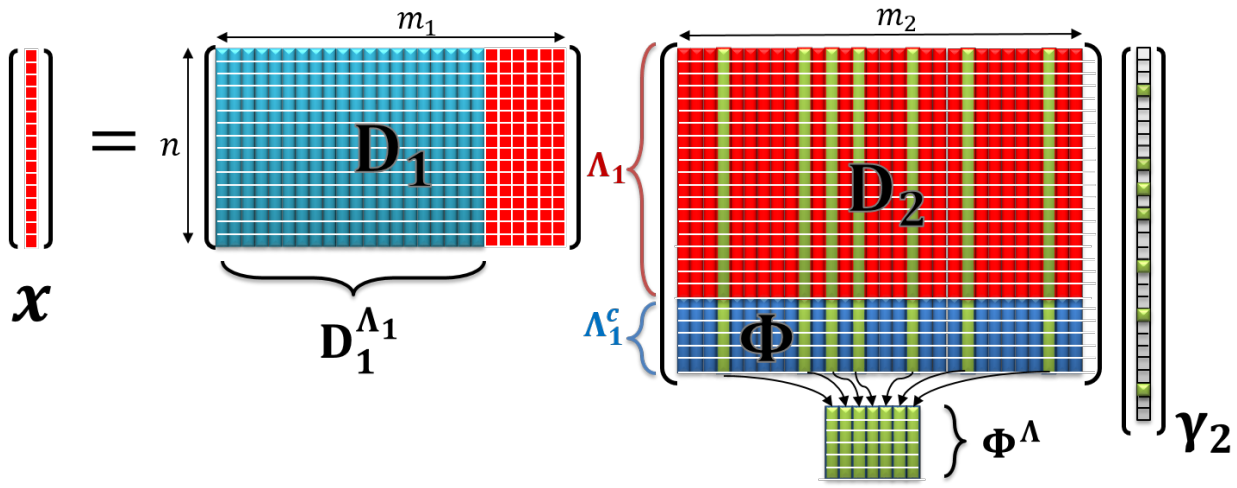
$$\min_{\boldsymbol{\gamma}_L} \quad \|\mathbf{y} - \mathbf{D}_{(1,L)}\boldsymbol{\gamma}_L\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}_L\|_0 \le s_L,\ \left\{\|\mathbf{D}_{(i,L)}\boldsymbol{\gamma}_L\|_0 \le s_{i-1}\right\}_{i=1}^{L}$$

$\Rightarrow$ Coupled Synthesis and Analysis priors!

Synthesis-Analysis Interpretation



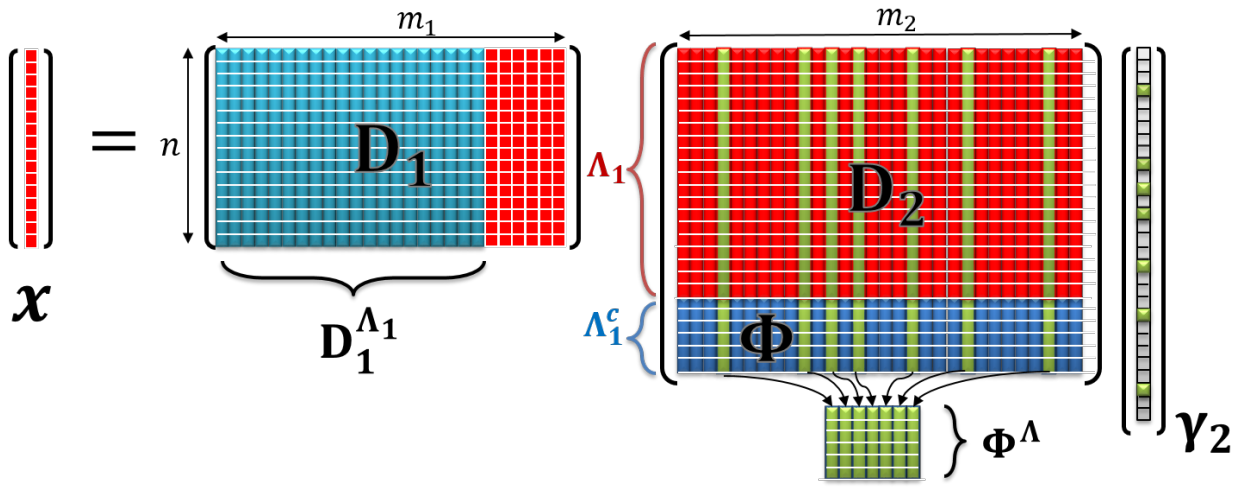
Synthesis-Analysis Interpretation



Uniqueness

- Degrees of freedom $= s_2 - \text{rank}(\Phi^{\Lambda})$

Synthesis-Analysis Interpretation



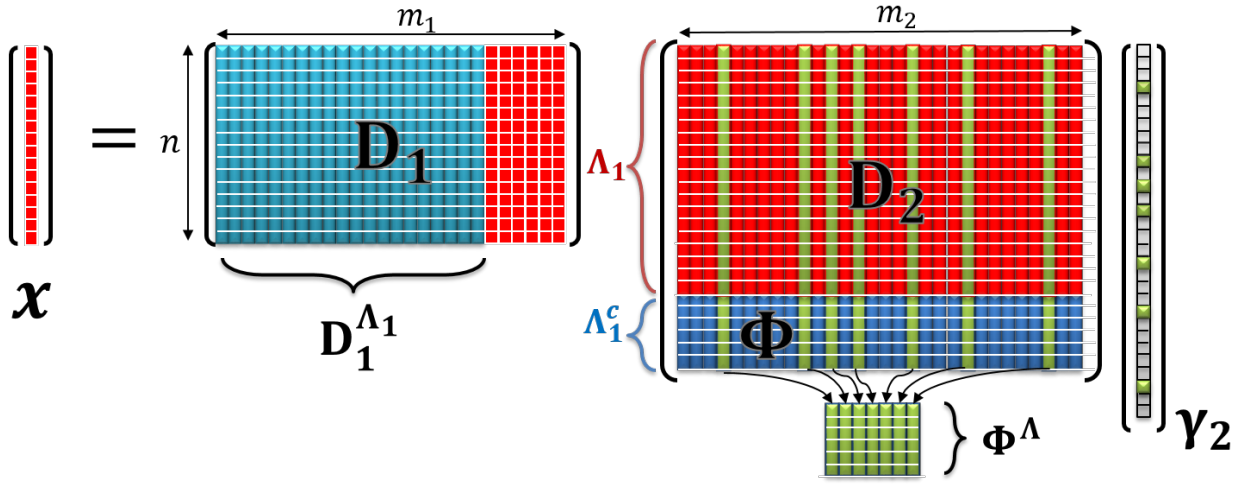
Uniqueness

- Degrees of freedom $= s_2 - \text{rank}(\Phi^{\Lambda})$
- Representations are **unique** if

$$\|\boldsymbol{\gamma}_2\|_0 \leq \frac{\eta(\mathbf{D}_{(1,2)}) - 1}{2}$$

$\eta(\mathbf{D}_i)$: *spark* of $\mathbf{D}_i$ – minimal number of *linearly dependent* columns.

Synthesis-Analysis Interpretation



Uniqueness

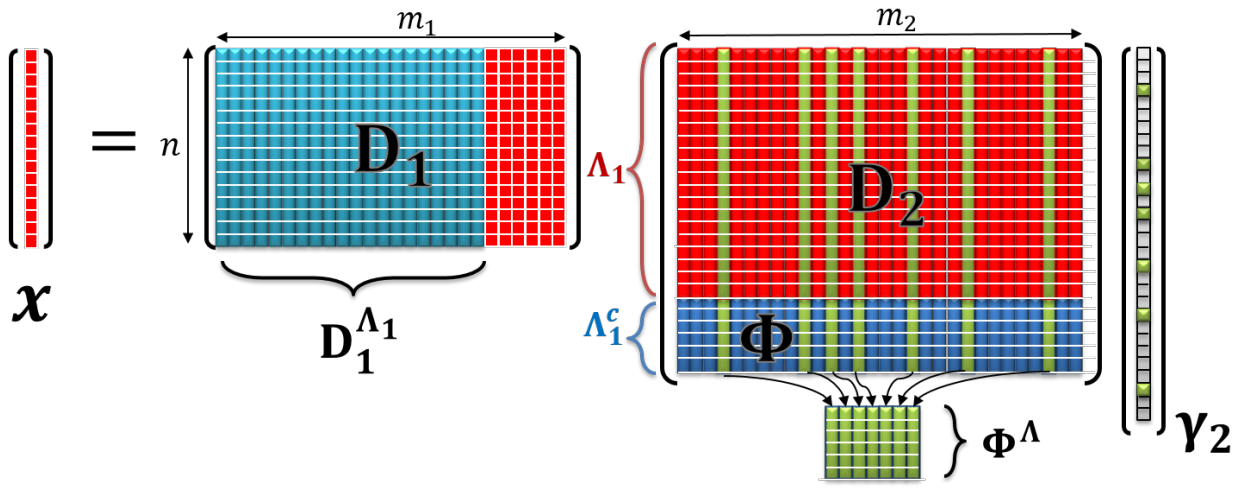
- Degrees of freedom $= s_2 - \text{rank}(\Phi^\Lambda)$
- Representations are **unique** if

$$\|\boldsymbol{\gamma}_2\|_0 \leq \frac{\eta(\mathbf{D}_{(1,2)}) - 1}{2} + \text{rank}(\Phi^\Lambda)$$

$\eta(\mathbf{D}_i)$: *spark* of $\mathbf{D}_i$ – minimal number of *linearly dependent* columns.

- ✓ Improved guarantees
- ✓ Intermediate representations *not as sparse*

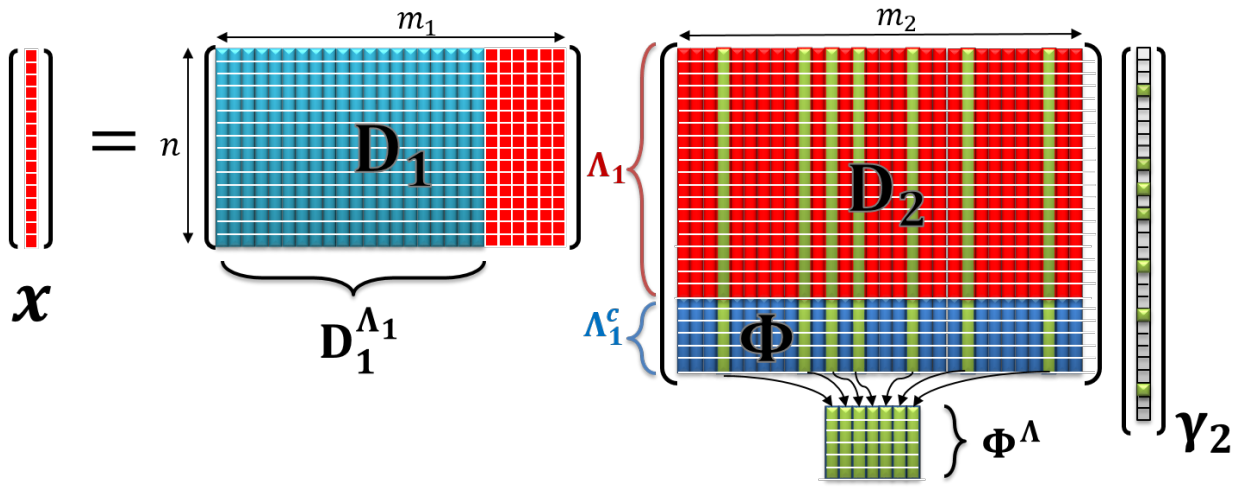
Synthesis-Analysis Interpretation



Stability

$$\mathbf{y} = \mathbf{D}_{(1,L)}\boldsymbol{\gamma}_L + \mathbf{w}, \quad \mathbf{w} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

Synthesis-Analysis Interpretation



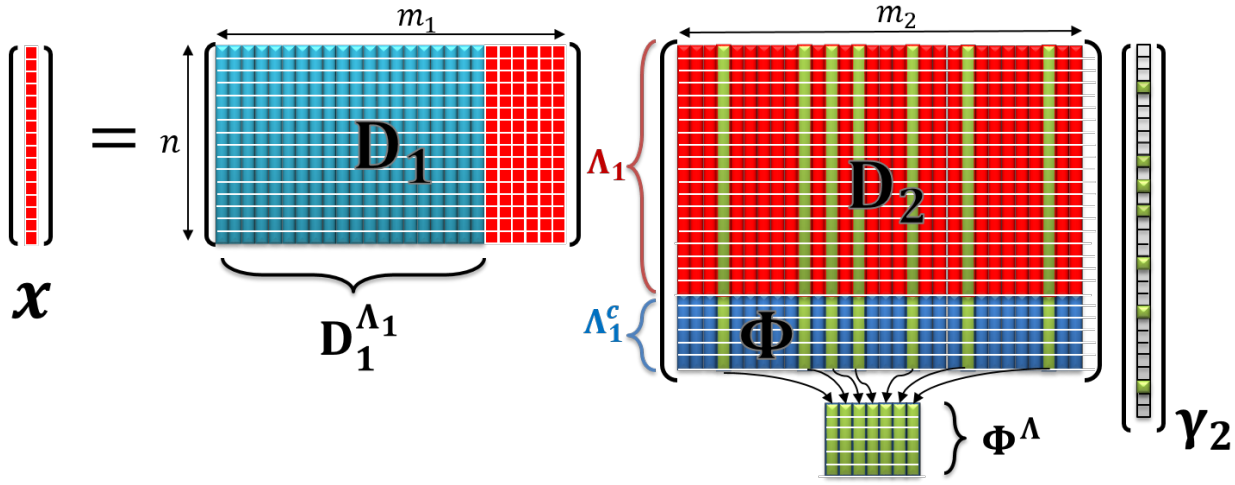
Stability

$\mathbf{y} = \mathbf{D}_{(1,L)}\boldsymbol{\gamma}_L + \mathbf{w}, \quad \mathbf{w} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$

- Layered Synthesis

$$\mathbb{E}\ \|\boldsymbol{\gamma}_i - \hat{\boldsymbol{\gamma}}_i\|_2^2 \lesssim \frac{\sigma^2}{(1-\delta_{s_L})} s_L$$

Synthesis-Analysis Interpretation



Stability

$\mathbf{y} = \mathbf{D}_{(1,L)}\boldsymbol{\gamma}_L + \mathbf{w}, \quad \mathbf{w} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$

- Layered Synthesis

$$\mathbb{E}\ \|\boldsymbol{\gamma}_i - \hat{\boldsymbol{\gamma}}_i\|_2^2 \lesssim \frac{\sigma^2}{(1-\delta_{s_L})} s_L$$

- Synthesis-Analysis

$$\mathbb{E}\ \|\boldsymbol{\gamma}_i - \hat{\boldsymbol{\gamma}}_i\|_2^2 \lesssim \frac{\sigma^2}{(1-\delta_{s_L})} (s_L - rank(\Phi^{\Lambda_L}))$$

Multi-Layer Basis Pursuit

Pursuit

$$\min_{\boldsymbol{\gamma}_L} \quad \|\mathbf{y} - \mathbf{D}_{(1,L)}\boldsymbol{\gamma}_L\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}_L\|_0 \leq s_L, \ \left\{\|\mathbf{D}_{(i,L)}\boldsymbol{\gamma}_L\|_0 \leq s_{i-1}\right\}_{i=1}^{L}$$

with $\mathbf{D}_{(i,L)} = \mathbf{D}_i \dots \mathbf{D}_L$

Multi-Layer Basis Pursuit

Pursuit

$$\min_{\boldsymbol{\gamma}_L} \quad \|\mathbf{y} - \mathbf{D}_{(1,L)}\boldsymbol{\gamma}_L\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}_L\|_0 \leq s_L, \ \left\{\|\mathbf{D}_{(i,L)}\boldsymbol{\gamma}_L\|_0 \leq s_{i-1}\right\}_{i=1}^{L}$$

with $\mathbf{D}_{(i,L)} = \mathbf{D}_i \dots \mathbf{D}_L$

Just relax!

$$(P): \quad \min_{\boldsymbol{\gamma}_L} \|\mathbf{y} - \mathbf{D}_{(1,L)}\boldsymbol{\gamma}_L\|_2^2 + \sum_{i=2}^{L} \lambda_{i-1}\|\mathbf{D}_{(i,L)}\boldsymbol{\gamma}_L\|_1 + \lambda_L\|\boldsymbol{\gamma}_L\|_1$$

Multi-Layer Basis Pursuit

Pursuit

$$\min_{\boldsymbol{\gamma}_L} \quad \|\mathbf{y} - \mathbf{D}_{(1,L)}\boldsymbol{\gamma}_L\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}_L\|_0 \leq s_L, \ \left\{\|\mathbf{D}_{(i,L)}\boldsymbol{\gamma}_L\|_0 \leq s_{i-1}\right\}_{i=1}^{L}$$

with $\mathbf{D}_{(i,L)} = \mathbf{D}_i \ldots \mathbf{D}_L$

Just relax!

$$(P): \quad \min_{\boldsymbol{\gamma}_2} \frac{1}{2}\|\mathbf{y} - \mathbf{D}_1\mathbf{D}_2\boldsymbol{\gamma}_2\|_2^2 + \lambda_1\|\mathbf{D}_2\boldsymbol{\gamma}_2\|_1 + \lambda_2\|\boldsymbol{\gamma}_2\|_1$$

Multi-Layer ISTA

$$\min_{\boldsymbol{\gamma}_2} F(\boldsymbol{\gamma}_2) = f(\mathbf{D}_2\boldsymbol{\gamma}_2) + g_1(\mathbf{D}_2\boldsymbol{\gamma}_2) + g_2(\boldsymbol{\gamma}_2)$$

Multi-Layer ISTA

$$\min_{\boldsymbol{\gamma}_2} F(\boldsymbol{\gamma}_2) = \underbrace{f(\mathbf{D}_2\boldsymbol{\gamma}_2) + g_1(\mathbf{D}_2\boldsymbol{\gamma}_2)}_{h(\boldsymbol{\gamma}_2)} + g_2(\boldsymbol{\gamma}_2)$$

Multi-Layer ISTA

$$\min_{\boldsymbol{\gamma}_2} F(\boldsymbol{\gamma}_2) = \underbrace{f(\mathbf{D}_2\boldsymbol{\gamma}_2) + g_1(\mathbf{D}_2\boldsymbol{\gamma}_2)}_{h(\boldsymbol{\gamma}_2)} + g_2(\boldsymbol{\gamma}_2)$$

Proximal Gradient

$$\boldsymbol{\gamma}_2^{k+1} = \text{prox}_{tg_2}\left(\boldsymbol{\gamma}_2^k - t\nabla h(\boldsymbol{\gamma}_2^k)\right) \quad ??$$

Multi-Layer ISTA

$$\min_{\boldsymbol{\gamma}_2} F(\boldsymbol{\gamma}_2) = f(\mathbf{D}_2\boldsymbol{\gamma}_2) + g_1(\mathbf{D}_2\boldsymbol{\gamma}_2) + g_2(\boldsymbol{\gamma}_2)$$

Proximal Gradient-Mapping

$$\boldsymbol{\gamma}_2^{k+1} = \text{prox}_{tg_2}\left(\boldsymbol{\gamma}_2^k - tG_{1/\mu}^{f,g_1(\mathbf{D}_2\cdot)}(\boldsymbol{\gamma}_2^k)\right)$$

Multi-Layer ISTA

$$\min_{\boldsymbol{\gamma}_2} F(\boldsymbol{\gamma}_2) = f(\mathbf{D}_2\boldsymbol{\gamma}_2) + g_1(\mathbf{D}_2\boldsymbol{\gamma}_2) + g_2(\boldsymbol{\gamma}_2)$$

Proximal Gradient-Mapping

$$\boldsymbol{\gamma}_2^{k+1} = \text{prox}_{tg_2}\left(\boldsymbol{\gamma}_2^k - tG_{1/\mu}^{f,g_1(\mathbf{D}_2\cdot)}(\boldsymbol{\gamma}_2^k)\right)$$

where

$$G_{1/\mu}^{f,g_1}(\boldsymbol{\gamma}_2^k) = \frac{1}{\mu}\left[\boldsymbol{\gamma}_2^k - \text{prox}_{\mu g_1(\mathbf{D}_2\cdot)}\left(\boldsymbol{\gamma}_2^k - \mu\nabla f(\mathbf{D}_2\boldsymbol{\gamma}_2)\right)\right]$$

Multi-Layer ISTA

$$\min_{\boldsymbol{\gamma}_2} F(\boldsymbol{\gamma}_2) = f(\mathbf{D}_2\boldsymbol{\gamma}_2) + g_1(\mathbf{D}_2\boldsymbol{\gamma}_2) + g_2(\boldsymbol{\gamma}_2)$$

Proximal Gradient-Mapping

$$\boldsymbol{\gamma}_2^{k+1} = \text{prox}_{tg_2}\left(\boldsymbol{\gamma}_2^k - t\mathbf{D}_2^T G_{1/\mu}^{f,g_1}(\boldsymbol{\gamma}_1^k)\right)$$

where $\boldsymbol{\gamma}_1^k = \mathbf{D}_2\boldsymbol{\gamma}_2^k$ and

$$G_{1/\mu}^{f,g_1}(\boldsymbol{\gamma}_1^k) = \frac{1}{\mu}\left[\boldsymbol{\gamma}_1^k - \text{prox}_{\mu g_1}\left(\boldsymbol{\gamma}_1^k - \mu\nabla f(\boldsymbol{\gamma}_1^k)\right)\right]$$

Multi-Layer ISTA

$$\min_{\boldsymbol{\gamma}_2} F(\boldsymbol{\gamma}_2) = f(\mathbf{D}_2\boldsymbol{\gamma}_2) + g_1(\mathbf{D}_2\boldsymbol{\gamma}_2) + g_2(\boldsymbol{\gamma}_2)$$

Proximal Gradient-Mapping

$$\boldsymbol{\gamma}_2^{k+1} = \text{prox}_{tg_2}\left(\boldsymbol{\gamma}_2^k - t\mathbf{D}_2^T G_{1/\mu}^{f,g_1}(\boldsymbol{\gamma}_1^k)\right)$$

where $\boldsymbol{\gamma}_1^k = \mathbf{D}_2\boldsymbol{\gamma}_2^k$ and

$$G_{1/\mu}^{f,g_1}(\boldsymbol{\gamma}_1^k) = \frac{1}{\mu}\left[\boldsymbol{\gamma}_1^k - \text{prox}_{\mu g_1}\left(\boldsymbol{\gamma}_1^k - \mu\nabla f(\boldsymbol{\gamma}_1^k)\right)\right]$$

In particular:

$$\boldsymbol{\gamma}_2^{k+1} = \mathcal{T}_{t\lambda_2}\left(\boldsymbol{\gamma}_2^k - \frac{t}{\mu}\mathbf{D}_2^T\left(\boldsymbol{\gamma}_1^k - \mathcal{T}_{\mu\lambda_1}(\boldsymbol{\gamma}_1^k - \mu\mathbf{D}_1^T(\mathbf{D}_1\boldsymbol{\gamma}_1^k - \mathbf{y}))\right)\right)$$

ML-ISTA

Set $\boldsymbol{\gamma}_0^k = \mathbf{y}\ \forall\ k$ and $\boldsymbol{\gamma}_L^1 = 0$

for $k = 1 : K$ **do**

  % for each iteration

    $\hat{\boldsymbol{\gamma}}_i \leftarrow \mathbf{D}_{(i,L)}\boldsymbol{\gamma}_L^k \quad \forall i \in [0, k-1]$

  for $i = 1 : L$ **do**

    % for each layer

      $\boldsymbol{\gamma}_i^{k+1} \leftarrow \mathcal{T}_{\mu_i\lambda_i}\left(\hat{\boldsymbol{\gamma}}_i - \mu_i\mathbf{D}_i^T(\mathbf{D}_i\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i-1}^{k+1})\right)$

  end

end

ML-ISTA

Set $\boldsymbol{\gamma}_0^k = \mathbf{y}\ \forall\ k$ and $\boldsymbol{\gamma}_L^1 = 0$
for $k = 1 : K$ **do**
    % for each iteration
        $\hat{\boldsymbol{\gamma}}_i \leftarrow \mathbf{D}_{(i,L)}\boldsymbol{\gamma}_L^k \quad \forall i \in [0, k-1]$
    for $i = 1 : L$ **do**
        % for each layer
            $\boldsymbol{\gamma}_i^{k+1} \leftarrow \mathcal{T}_{\mu_i\lambda_i}\left(\hat{\boldsymbol{\gamma}}_i - \mu_i\mathbf{D}_i^T(\mathbf{D}_i\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i-1}^{k+1})\right)$
    end
end

✓ *Nested Proximal-Gradient updates*

ML-FISTA

Set $\boldsymbol{\gamma}_0^k = \mathbf{y}\ \forall\ k$ and $\boldsymbol{\gamma}_L^1 = 0$
for $k = 1 : K$ **do**
  % for each iteration
    $\hat{\boldsymbol{\gamma}}_i \leftarrow \mathbf{D}_{(i,L)}\mathbf{z} \quad \forall i \in [0, k-1]$
  for $i = 1 : L$ **do**
    % for each layer
      $\boldsymbol{\gamma}_i^{k+1} \leftarrow \mathcal{T}_{\mu_i \lambda_i}\left(\hat{\boldsymbol{\gamma}}_i - \mu_i \mathbf{D}_i^T(\mathbf{D}_i\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i-1}^{k+1})\right)$
  end
    $\mathbf{z} \leftarrow \boldsymbol{\gamma}_L^{k+1} + \rho^k(\boldsymbol{\gamma}_L^{k+1} - \boldsymbol{\gamma}_L^k)$
end

✓ *Nested Proximal-Gradient updates with momentum*

Convergence for ML-ISTA

Convergence for ML-ISTA

Theorem

Suppose $\{\boldsymbol{\gamma}_2^k\}$ *generated by ML-ISTA with* $\mu \in \left(0, \frac{1}{\|\mathbf{D}_1\|_2^2}\right)$ *and* $t \in \left(0, \frac{4\mu}{3\|\mathbf{D}_2\|_2}\right)$.

If $\|\boldsymbol{\gamma}_2^{k+1} - \boldsymbol{\gamma}_2^k\|_2 \leq t\varepsilon$, *then*

$$F(\boldsymbol{\gamma}_2^{k+1}) - F_{opt} \leq \eta\varepsilon + (\beta + \kappa t)\mu,$$

where η, β *and* κ *are constants depending on* $\mathbf{D}_1$, $\mathbf{D}_2$, g_1, g_2.

Convergence for ML-ISTA

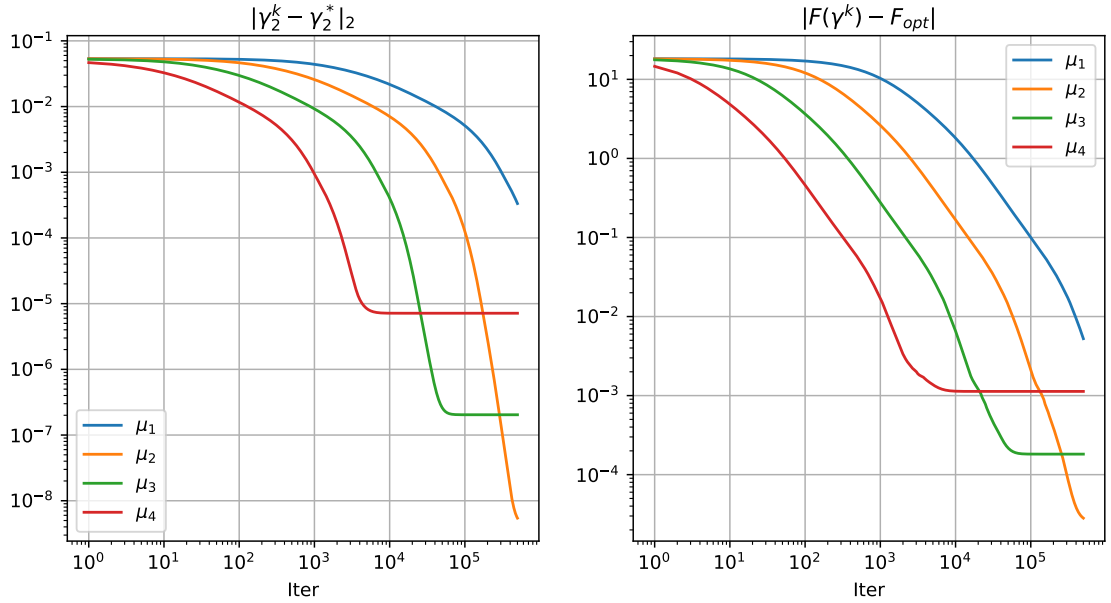
Theorem

Suppose $\{\boldsymbol{\gamma}_2^k\}$ *generated by ML-ISTA with* $\mu \in \left(0, \frac{1}{\|\mathbf{D}_1\|_2^2}\right)$ *and* $t \in \left(0, \frac{4\mu}{3\|\mathbf{D}_2\|_2}\right)$.

If $\|\boldsymbol{\gamma}_2^{k+1} - \boldsymbol{\gamma}_2^k\|_2 \leq t\varepsilon$, *then*

$$F(\boldsymbol{\gamma}_2^{k+1}) - F_{opt} \leq \eta\varepsilon + (\beta + \kappa t)\mu,$$

where η, β *and* κ *are constants depending on* $\mathbf{D}_1$, $\mathbf{D}_2$, g_1, g_2.



Convergence for ML-FISTA

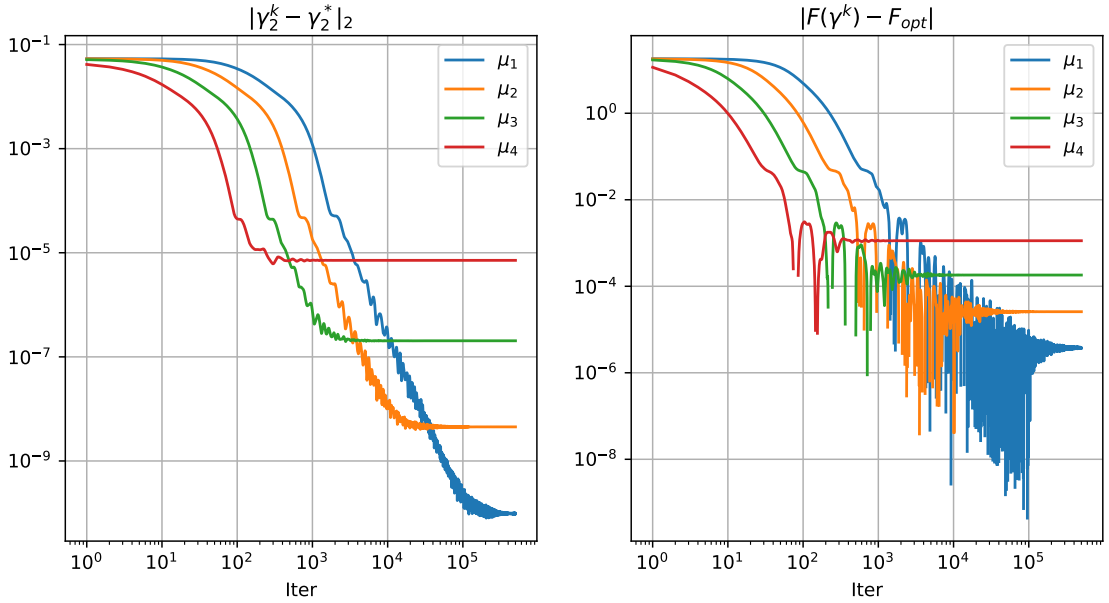
Theorem

Suppose $\{\boldsymbol{\gamma}_2^k\}$ *generated by ML-ISTA with* $\mu \in \left(0, \frac{1}{\|\mathbf{D}_1\|_2^2}\right)$ *and* $t \in \left(0, \frac{4\mu}{3\|\mathbf{D}_2\|_2}\right)$.

If $\|\boldsymbol{\gamma}_2^{k+1} - \boldsymbol{\gamma}_2^k\|_2 \leq t\varepsilon$, *then*

$$F(\boldsymbol{\gamma}_2^{k+1}) - F_{opt} \leq \eta\varepsilon + (\beta + \kappa t)\mu,$$

where η, β *and* κ *are constants depending on* $\mathbf{D}_1$, $\mathbf{D}_2$, g_1, g_2.



Generalizing Recurrent CNNs

Revisiting the ML-ISTA

$$\boldsymbol{\gamma}_2^{k+1} = \mathcal{T}_{t\lambda_2}\left(\boldsymbol{\gamma}_2^k - \mu_2 \mathbf{D}_2^T\left(\boldsymbol{\gamma}_1^k - \mathcal{T}_{\mu_1\lambda_1}\left(\boldsymbol{\gamma}_1^k - \mu \mathbf{D}_1^T\left(\mathbf{D}_1\boldsymbol{\gamma}_1^k - \mathbf{y}\right)\right)\right)\right)$$

Generalizing Recurrent CNNs

Revisiting the ML-ISTA

$$\boldsymbol{\gamma}_2^{(1)} = \mathcal{T}_{t\lambda_2}\Big(\quad - \mu_2 \mathbf{D}_2^T \big(\quad - \mathcal{T}_{\mu_1\lambda_1} \big(\quad - \mu \mathbf{D}_1^T \big(\qquad - \mathbf{y} \big) \big) \big) \Big)$$

Generalizing Recurrent CNNs

Revisiting the ML-ISTA

$$\boldsymbol{\gamma}_2^{(1)} = \mathcal{T}_{t\lambda_2}\left(\mu_2 \mathbf{D}_2^T \mathcal{T}_{\mu\lambda_1}(\mu_1 \mathbf{D}_1^T \mathbf{y})\right)$$

Generalizing Recurrent CNNs

Revisiting the ML-ISTA

$$\boldsymbol{\gamma}_2^{(1)} = \mathsf{ReLU}\Big(\mu_2 \mathbf{D}_2^T \mathsf{ReLU}(\mu_1 \mathbf{D}_1^T \mathbf{y} + \mathbf{b}_1) + \mathbf{b}_2\Big) \qquad \text{(F.P.)}$$

Generalizing Recurrent CNNs

Revisiting the ML-ISTA

$$\boldsymbol{\gamma}_2^{(1)} = \mathsf{ReLU}\Big(\mu_2 \mathbf{D}_2^T \mathsf{ReLU}(\mu_1 \mathbf{D}_1^T \mathbf{y} + \mathbf{b}_1) + \mathbf{b}_2\Big) \qquad \text{(F.P.)}$$

$$\boldsymbol{\gamma}_2^{(2)} = \mathsf{ReLU}\Big(\boldsymbol{\gamma}_2^{(1)} - \mu_2 \mathbf{D}_2^T \big(\boldsymbol{\gamma}_1^{(1)} - \mathsf{ReLU}(\mu_1 \mathbf{D}_1^T (\mathbf{D}_1 \boldsymbol{\gamma}_1^{(1)} - \mathbf{y}) + \mathbf{b}_1)\big) + \mathbf{b}_2\Big)$$

$$\vdots$$

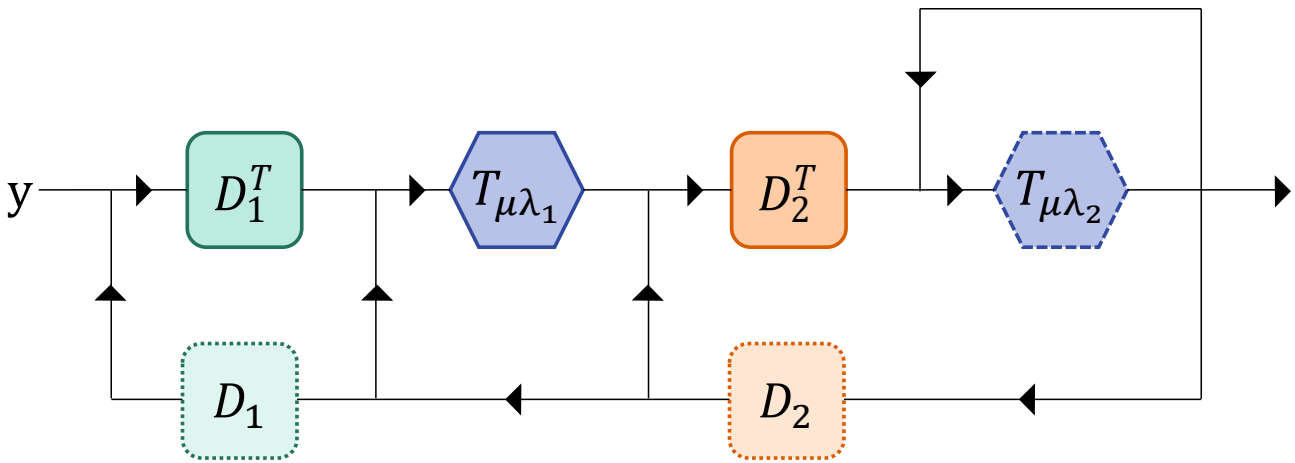
Generalizing Recurrent CNNs

Revisiting the ML-ISTA

$$\boldsymbol{\gamma}_2^{(1)} = \text{ReLU}\Big(\mu_2 \mathbf{D}_2^T \text{ReLU}(\mu_1 \mathbf{D}_1^T \mathbf{y} + \mathbf{b}_1) + \mathbf{b}_2\Big) \qquad \text{(F.P.)}$$

$$\boldsymbol{\gamma}_2^{(2)} = \text{ReLU}\Big(\boldsymbol{\gamma}_2^{(1)} - \mu_2 \mathbf{D}_2^T \big(\boldsymbol{\gamma}_1^{(1)} - \text{ReLU}(\mu_1 \mathbf{D}_1^T (\mathbf{D}_1 \boldsymbol{\gamma}_1^{(1)} - \mathbf{y}) + \mathbf{b}_1)\big) + \mathbf{b}_2\Big)$$

$$\vdots$$



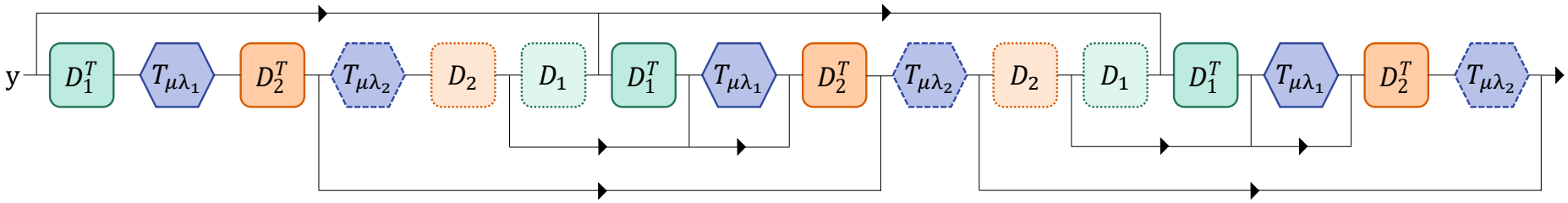
Generalizing Recurrent CNNs

Revisiting the ML-ISTA

$$\boldsymbol{\gamma}_2^{(1)} = \text{ReLU}\Big(\mu_2 \mathbf{D}_2^T \text{ReLU}(\mu_1 \mathbf{D}_1^T \mathbf{y} + \mathbf{b}_1) + \mathbf{b}_2\Big) \qquad \text{(F.P.)}$$

$$\boldsymbol{\gamma}_2^{(2)} = \text{ReLU}\Big(\boldsymbol{\gamma}_2^{(1)} - \mu_2 \mathbf{D}_2^T \big(\boldsymbol{\gamma}_1^{(1)} - \text{ReLU}(\mu_1 \mathbf{D}_1^T (\mathbf{D}_1 \boldsymbol{\gamma}_1^{(1)} - \mathbf{y}) + \mathbf{b}_1)\big) + \mathbf{b}_2\Big)$$

$$\vdots$$



Supervised Learning Formulation

Training examples: $\{(\mathbf{y}_i, h_i)\}_{i=1}^{N}$, $\mathbf{y}$: inputs, h_i : labels

Supervised Learning Formulation

Training examples: $\{(\mathbf{y}_i, h_i)\}_{i=1}^{N}$, $\mathbf{y}$: inputs, h_i : labels

$$\min_{\theta, \{\mathbf{D}_i, \lambda_i\}} \frac{1}{N} \sum_{1=1}^{N} \mathcal{L}\left(h_i, \zeta_\theta(\boldsymbol{\gamma}^*)\right) \quad \text{s.t.}$$

Supervised Learning Formulation

Training examples: $\{(\mathbf{y}_i, h_i)\}_{i=1}^N$, $\mathbf{y}$: inputs, h_i : labels

$$\min_{\theta, \{\mathbf{D}_i, \lambda_i\}} \frac{1}{N} \sum_{1=1}^{N} \mathcal{L}\left(h_i, \zeta_\theta(\boldsymbol{\gamma}^*)\right) \quad \text{s.t.}$$

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{D}_{(1,L)}\boldsymbol{\gamma}\|_2^2 + \sum_{i=1}^{L-1} \lambda_i \|\mathbf{D}_{(i+1,L)}\boldsymbol{\gamma}\|_1 + \lambda_L \|\boldsymbol{\gamma}\|_1.$$

Supervised Learning Formulation

Training examples: $\{(\mathbf{y}_i, h_i)\}_{i=1}^N$, $\mathbf{y}$: inputs, h_i : labels

$$\min_{\theta, \{\mathbf{D}_i, \lambda_i\}} \frac{1}{N} \sum_{1=1}^{N} \mathcal{L}\left(h_i, \zeta_\theta(\boldsymbol{\gamma}^*)\right) \quad \text{s.t.}$$

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{D}_{(1,L)}\boldsymbol{\gamma}\|_2^2 + \sum_{i=1}^{L-1} \lambda_i \|\mathbf{D}_{(i+1,L)}\boldsymbol{\gamma}\|_1 + \lambda_L \|\boldsymbol{\gamma}\|_1.$$

Relaxed to

$$\min_{\theta, \{\mathbf{D}_i, \lambda_i\}} \frac{1}{N} \sum_{1=1}^{N} \mathcal{L}\left(h_i, \zeta_\theta(\hat{\boldsymbol{\gamma}}^k)\right)$$

$\boldsymbol{\gamma}^k$: k^{th} iteration of ML-ISTA.

Supervised Learning Formulation

Training examples: $\{(\mathbf{y}_i, h_i)\}_{i=1}^N$, $\mathbf{y}$: inputs, h_i : labels

$$\min_{\theta, \{\mathbf{D}_i, \lambda_i\}} \frac{1}{N} \sum_{1=1}^{N} \mathcal{L}\left(h_i, \zeta_\theta(\boldsymbol{\gamma}^*)\right) \quad \text{s.t.}$$

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{D}_{(1,L)}\boldsymbol{\gamma}\|_2^2 + \sum_{i=1}^{L-1} \lambda_i \|\mathbf{D}_{(i+1,L)}\boldsymbol{\gamma}\|_1 + \lambda_L \|\boldsymbol{\gamma}\|_1.$$

Relaxed to

$$\min_{\theta, \{\mathbf{D}_i, \lambda_i\}} \frac{1}{N} \sum_{1=1}^{N} \mathcal{L}\left(h_i, \zeta_\theta(\hat{\boldsymbol{\gamma}}^k)\right)$$

$\boldsymbol{\gamma}^k$: k^{th} iteration of ML-ISTA.

- if $k = 1 \rightarrow$ Feed Forward CNN
- if $k > 1 \rightarrow$ Recurrent (ML-ISTA) CNN
  But same number of parameters

Supervised Learning Formulation

Training examples: $\{(\mathbf{y}_i, h_i)\}_{i=1}^N$, $\mathbf{y}$: inputs, h_i : labels

$$\min_{\theta, \{\mathbf{D}_i, \lambda_i\}} \frac{1}{N} \sum_{1=1}^{N} \mathcal{L}\left(h_i, \zeta_\theta(\boldsymbol{\gamma}^*)\right) \quad \text{s.t.}$$

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{D}_{(1,L)}\boldsymbol{\gamma}\|_2^2 + \sum_{i=1}^{L-1} \lambda_i \|\mathbf{D}_{(i+1,L)}\boldsymbol{\gamma}\|_1 + \lambda_L \|\boldsymbol{\gamma}\|_1.$$

Relaxed to

$$\min_{\theta, \{\mathbf{D}_i, \lambda_i\}} \frac{1}{N} \sum_{1=1}^{N} \mathcal{L}\left(h_i, \zeta_\theta(\hat{\boldsymbol{\gamma}}^k)\right)$$

$\boldsymbol{\gamma}^k$: k^{th} iteration of ML-ISTA.

- if $k = 1 \rightarrow$ Feed Forward CNN
- if $k > 1 \rightarrow$ Recurrent (ML-ISTA) CNN
  But same number of parameters

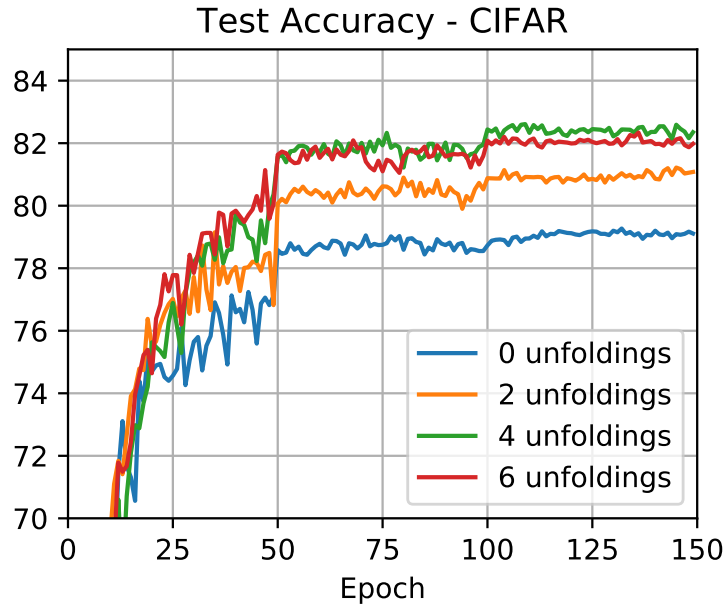
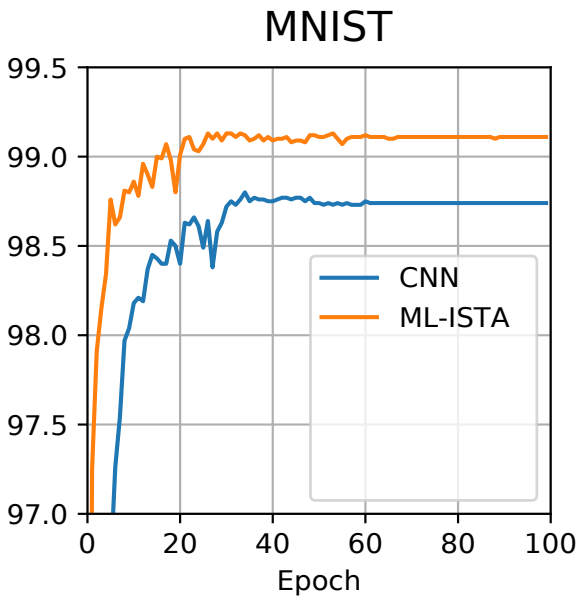
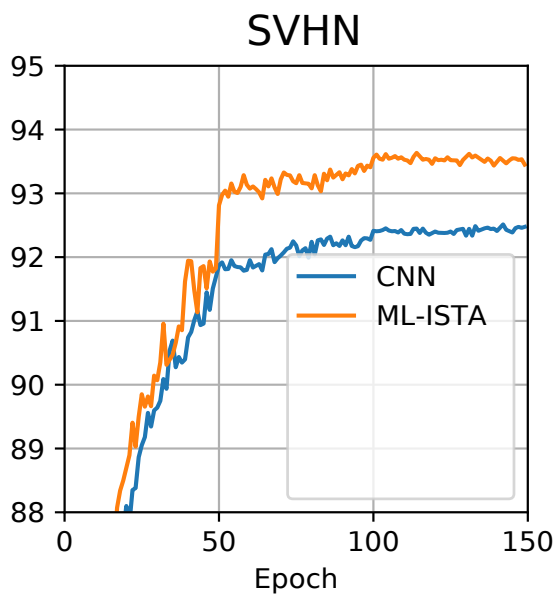


Image Classification





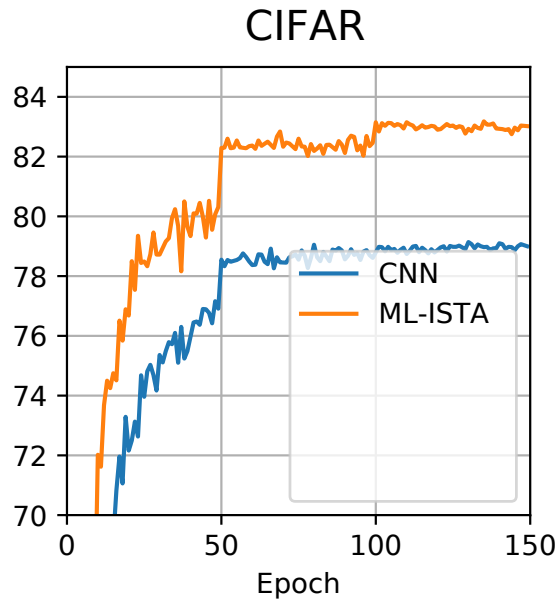
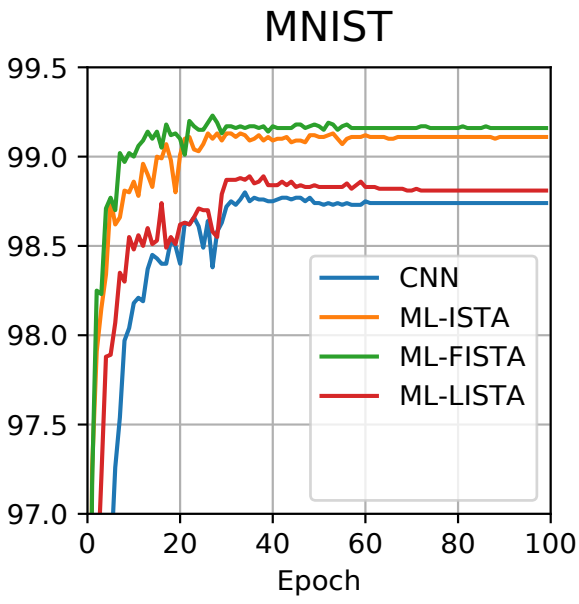
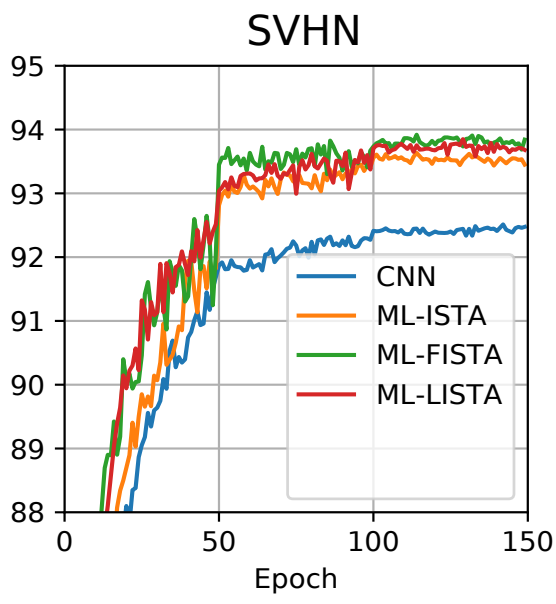


Image Classification





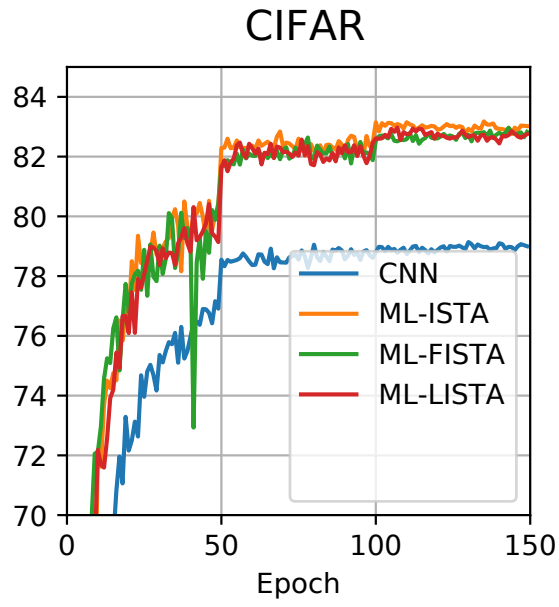
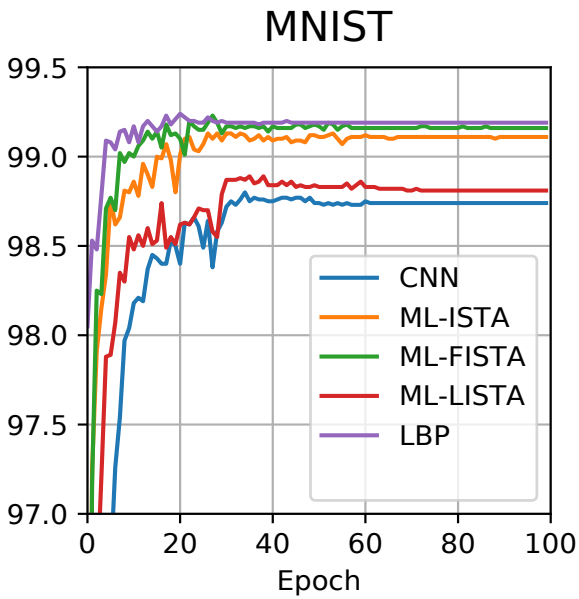
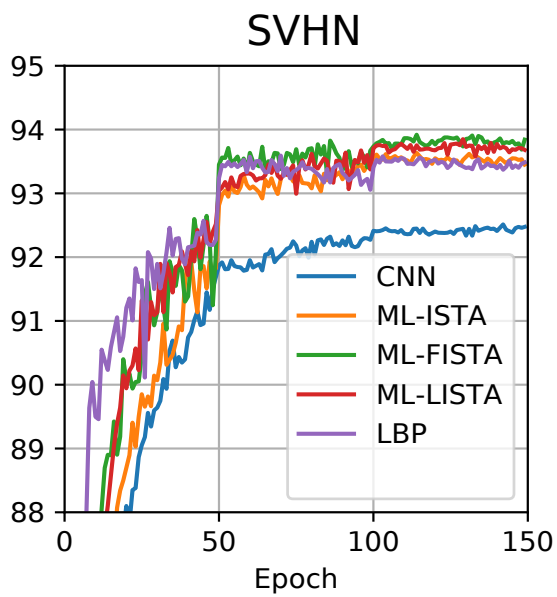


Image Classification





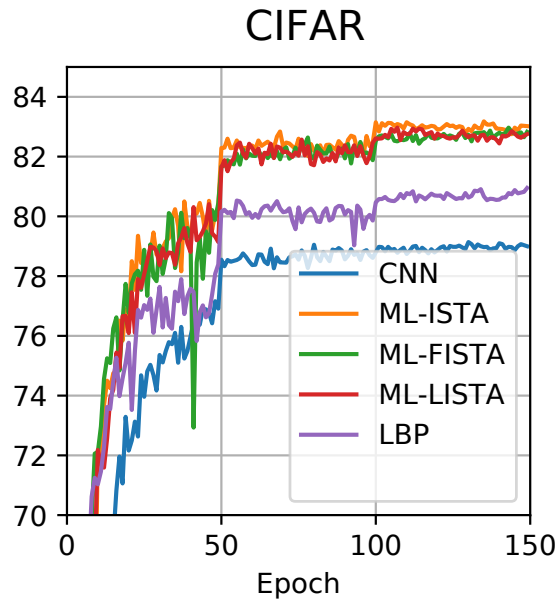
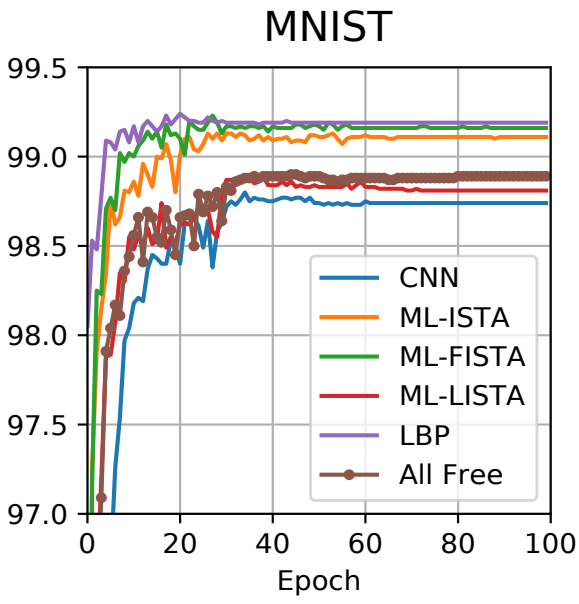
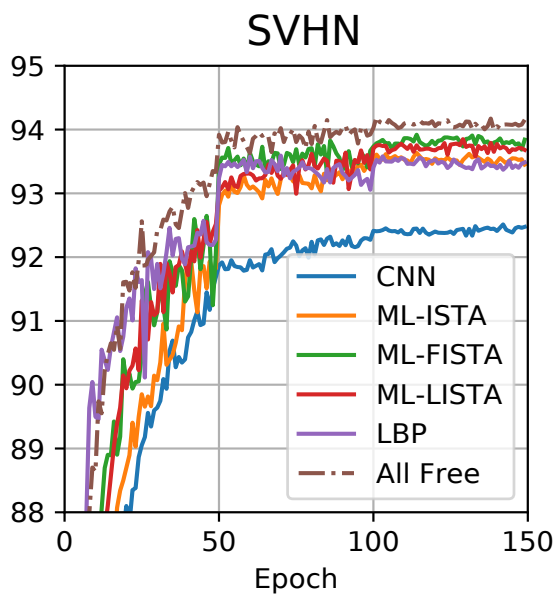
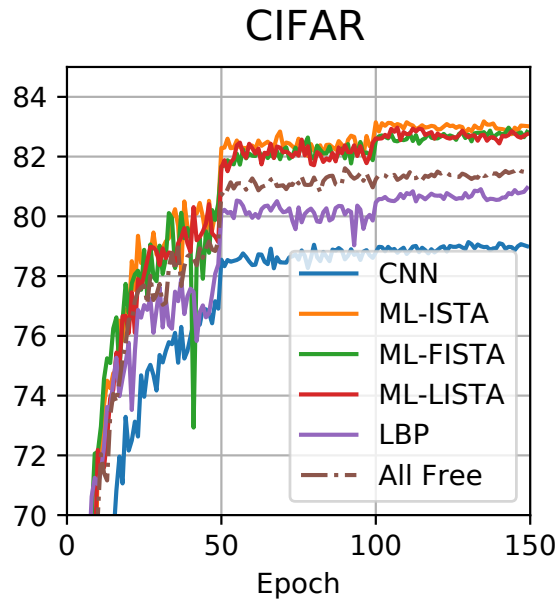


Image Classification

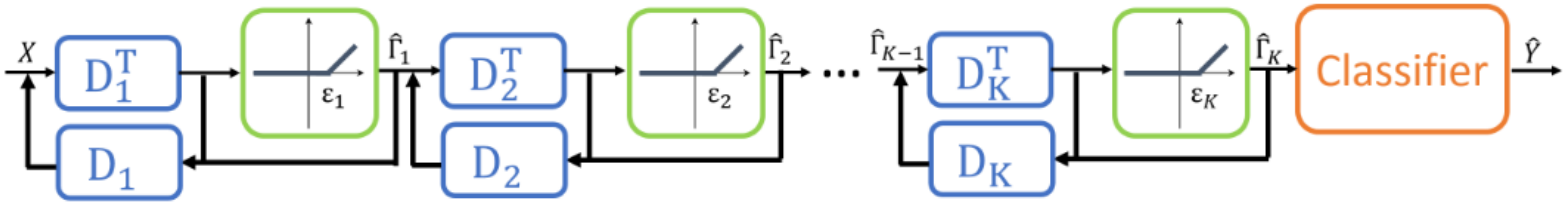






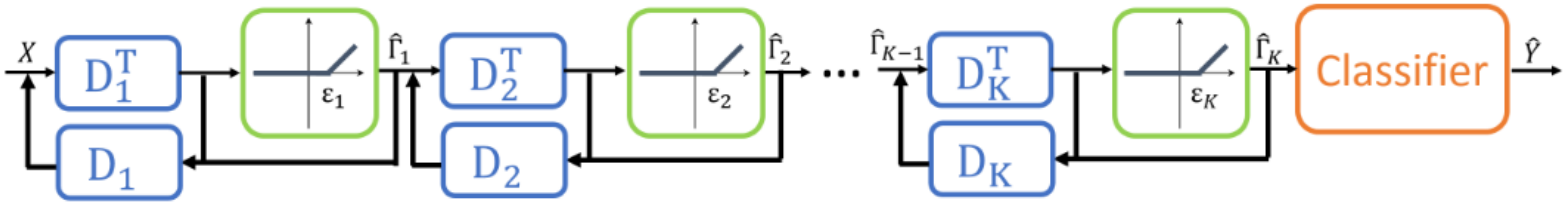
Adversarial Robustness Analysis

Adversarial Robustness Analysis



- Compute Representation: $\hat{\boldsymbol{\gamma}}_i = \arg\min_{\boldsymbol{\gamma}} \frac{1}{2}\|\hat{\boldsymbol{\gamma}}_{i-1} - \mathbf{D}_i\boldsymbol{\gamma}\|_2^2 + \lambda_i\|\boldsymbol{\gamma}\|_1$
- Compute Label: $\hat{y} = sign(\mathbf{w}^T\hat{\boldsymbol{\gamma}}_L)$

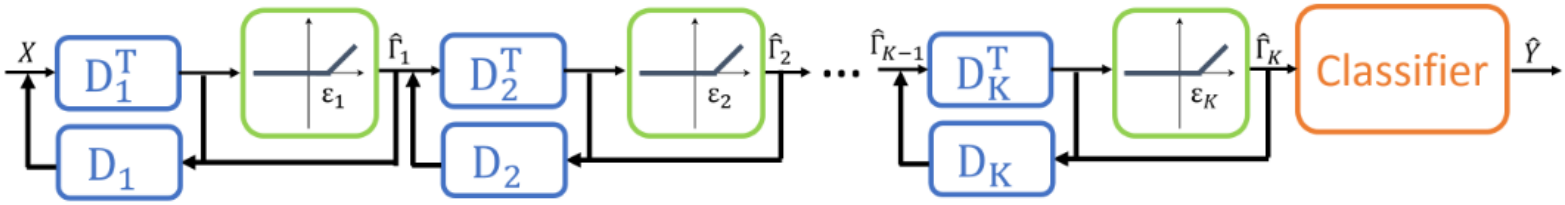
Adversarial Robustness Analysis



- Compute Representation: $\hat{\boldsymbol{\gamma}}_i = \arg\min_{\boldsymbol{\gamma}} \frac{1}{2}\|\hat{\boldsymbol{\gamma}}_{i-1} - \mathbf{D}_i\boldsymbol{\gamma}\|_2^2 + \lambda_i\|\boldsymbol{\gamma}\|_1$
- Compute Label: $\hat{y} = sign(\mathbf{w}^T\hat{\boldsymbol{\gamma}}_L)$

Theorem (1 layer, FC): Let a signal with label y that is (adversarily) contaminated as $\mathbf{x} = \mathbf{D}\boldsymbol{\gamma} + \mathbf{v}$, such that $\|\mathbf{v}\|_2 \leq \epsilon$.

Adversarial Robustness Analysis



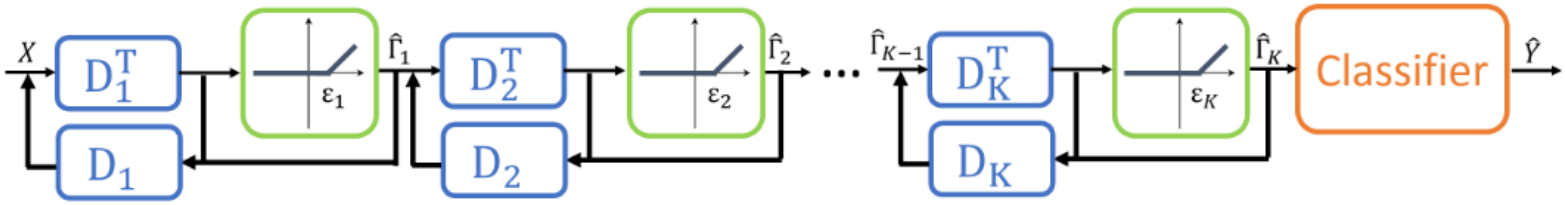
- Compute Representation: $\hat{\boldsymbol{\gamma}}_i = \arg\min_{\boldsymbol{\gamma}} \frac{1}{2}\|\hat{\boldsymbol{\gamma}}_{i-1} - \mathbf{D}_i\boldsymbol{\gamma}\|_2^2 + \lambda_i\|\boldsymbol{\gamma}\|_1$
- Compute Label: $\hat{y} = sign(\mathbf{w}^T\hat{\boldsymbol{\gamma}}_L)$

Theorem (1 layer, FC): Let a signal with label y that is (adversarily) contaminated as $\mathbf{x} = \mathbf{D}\boldsymbol{\gamma} + \mathbf{v}$, such that $\|\mathbf{v}\|_2 \leq \epsilon$. If $\|\boldsymbol{\gamma}_i\|_0 \leq \frac{1}{3}(1 + \frac{1}{\mu(\mathbf{D})})$, positive class margin $O_{\mathcal{B}} > 0$, and

$$\epsilon \leq \frac{O_{\mathcal{B}}}{7.5\|\boldsymbol{\gamma}\|_0\|\mathbf{w}\|_2}$$

Then $sign(\mathbf{w}^T\hat{\boldsymbol{\gamma}}) = sign(\mathbf{w}^T\boldsymbol{\gamma})$.

Adversarial Robustness Analysis

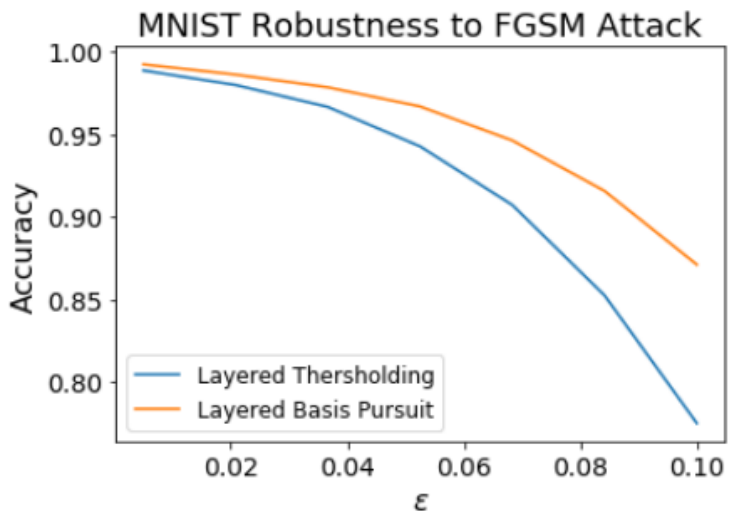


- Compute Representation: $\hat{\boldsymbol{\gamma}}_i = \arg\min_{\boldsymbol{\gamma}} \frac{1}{2}\|\hat{\boldsymbol{\gamma}}_{i-1} - \mathbf{D}_i\boldsymbol{\gamma}\|_2^2 + \lambda_i\|\boldsymbol{\gamma}\|_1$
- Compute Label: $\hat{y} = sign(\mathbf{w}^T\hat{\boldsymbol{\gamma}}_L)$

Theorem (1 layer, FC): Let a signal with label y that is (adversarily) contaminated as $\mathbf{x} = \mathbf{D}\boldsymbol{\gamma} + \mathbf{v}$, such that $\|\mathbf{v}\|_2 \leq \epsilon$. If $\|\boldsymbol{\gamma}_i\|_0 \leq \frac{1}{3}(1 + \frac{1}{\mu(\mathbf{D})})$, positive class margin $O_{\mathcal{B}} > 0$, and

$$\epsilon \leq \frac{O_{\mathcal{B}}}{7.5\|\boldsymbol{\gamma}\|_0\|\mathbf{w}\|_2}$$

Then $sign(\mathbf{w}^T\hat{\boldsymbol{\gamma}}) = sign(\mathbf{w}^T\boldsymbol{\gamma})$.



Take-Aways

Take-Aways

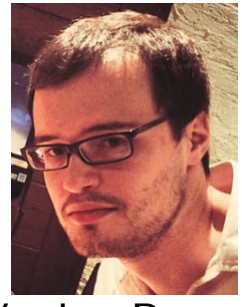
- Multi-Layer Sparse Model puts forward a generative sparse models for real signals
- Deep learning architectures can be understood analyzed as inference algorithms under this model

Joint work with

Aviad Aberdam
Technion

Amir Beck
Tel Aviv University

Yaniv Romano
Stanford



Vardan Papyan
Stanford

Miki Elad
Technion

References

- V. Papyan, Y. Romano, J. Sulam and M. Elad, **Theoretical Foundations of Deep Learning via Sparse Representations**, in IEEE Signal Processing Magazine, vol. 35, no. 4, pp. 72-89, July 2018.

- V. Papyan, J. Sulam, M. Elad. **Working Locally Thinking Globally: Theoretical Guarantees for Convolutional Sparse Coding**. IEEE Transactions on Signal Processing, 65(21), 5687-5701, 2017.
- V. Papyan, Y. Romano, and M. Elad. **Convolutional Neural Networks Analyzed via Convolutional Sparse Coding**. Journal of Machine Learning Research (JMLR), vol. 18, no. 83, pp. 1-52, 2017.
- J. Sulam, V. Papyan, Y. Romano, M. Elad. **Multi-Layer Convolutional Sparse Modeling: Pursuit and Dictionary Learning**. in IEEE TSP, vol. 66, no. 15, pp. 4090-4104, Aug.1, 1 2018.
- A. Aberdam, J. Sulam, M. Elad. **Multi Layer Sparse Coding: the Holistic Way**. SIAM Journal on Mathematics of Data Science, 2018.
- J. Sulam, A. Aberdam, A. Beck, M. Elad. **On Multi-Layer Basis Pursuit, Efficient Algorithms and Convolutional Neural Networks**. IEEE TPAMI (2019).
- Y. Romano, A. Aberdam, J. Sulam, M. Elad. **Adversarial Noise Attacks of Deep Learning Architectures-Stability Analysis via Sparse Modeled Signals**. arXiv preprint arXiv:1805.11596. (2019)