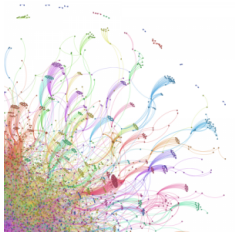
Neural Networks as Sparsity Enforcing Algorithms

Jeremias Sulam



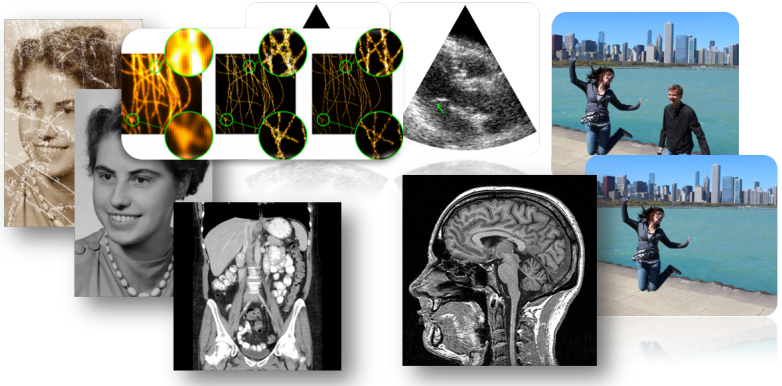
Deep Geometric Learning of Big Data and Applications
IPAM – 2019

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Center for
IMAGING
SCIENCE
JOHNS HOPKINS UNIVERSITY

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

Inverse Problems in Image Processing



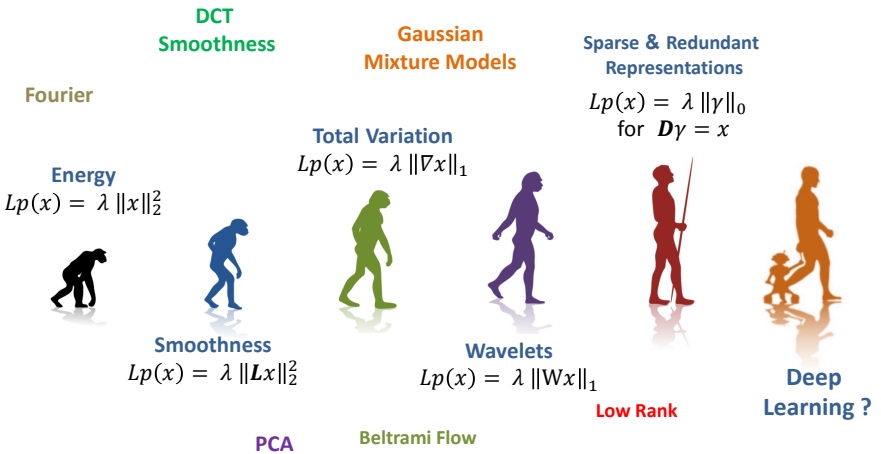
- All data has inherent **structure** than can be exploited
- This structure enables different **processing** tasks to be carried out

} **Signal Models**

find image that : is *related* to the input + looks like a *"good"* image

$$\min_{\mathbf{x}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \quad - \quad \log P(\mathbf{x})$$

Image Models



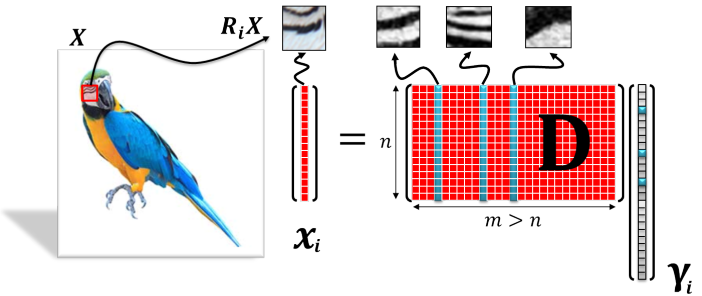
Sparsity $\in$ (artificial) Neural Networks

- Ranzato & LeCun et al, *Efficient Learning of Sparse Representations with an Energy-Based Model* NIPS '07
- Koray & LeCun et al. "Learning convolutional feature hierarchies for visual recognition" NIPS '10.
- Lee & Ng et al, "Sparse deep belief net model for visual area V2" NIPS '08
- ...
- He, A. Szlam et al "Unsupervised feature learning by deep sparse coding" ICDM '14.
- Zeiler & Fergus "Deconvolutional networks" CVPR. '10
- Alireza & Frey. "K-sparse autoencoders" '13, "Winner Take All Auto-encoders" NIPS '15

Neural Networks $\in$ Sparsity

- Gregor & LeCun "Learning fast approximations of sparse coding" ICML '10
- Moreau & Bruna "Understanding trainable sparse coding via matrix factorization" ICLR '17
- Liu et al. "On the convergence of learning-based iterative methods for nonconvex inverse problems" (2018).

Sparse Representations (synthesis)



Pursuit Problems: How to find γ?

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

Pursuit Algorithms

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

- **Greedy Algorithms**
  - (Orthogonal) Matching Pursuit
  - Iterative Hard Thresholding (P.G.D.)

- **Relaxation Approaches**

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

  - Convex optimization tools
  - Soft Thresholding
  - **Iterative Soft Thresholding (ISTA/FISTA)**

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

Sparse Representations

Characterization of the Dictionary

- **Restricted Isometry Property (R.I.P.)** smallest δ_s so that $\forall\ \boldsymbol{\gamma}:\ \|\boldsymbol{\gamma}\|_0 = s$,

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

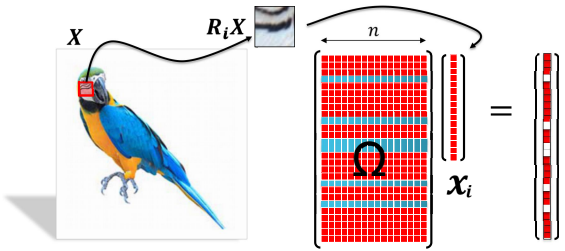
- **Mutual Coherence** $\mu(\mathbf{D}) = \max_{i \neq j} \frac{|\mathbf{d}_i^T \mathbf{d}_j|}{\|\mathbf{d}_i\|_2 \|\mathbf{d}_j\|_2}$

Sparse Representations are useful

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

$$\mathbb{E}\ \|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\|_2^2 \leq \frac{\sigma^2 s}{1-\delta_{\lambda_s}} \ll \sigma^2 n$$

Analysis Sparsity



Model

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

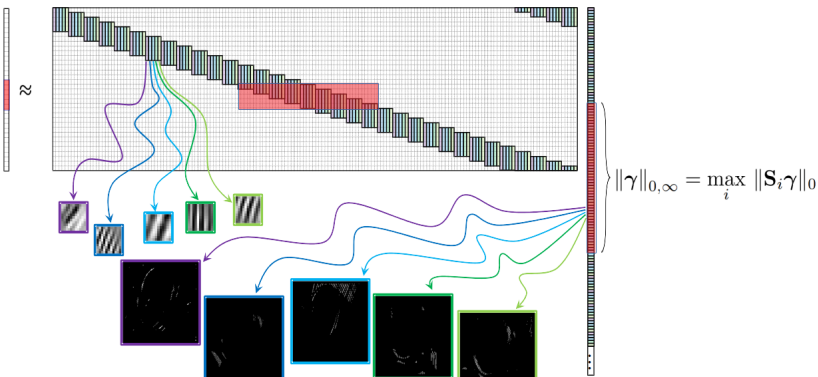
Pursuit

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

Convolutional Sparse Modeling of Signals

Model

$$\mathbf{y} = \mathbf{D}\boldsymbol{\gamma} + \mathbf{w}, \qquad \|\boldsymbol{\gamma}\|_{0,\infty} \leq s$$



Convolutional Sparse Modeling of Signals

Model

$$\mathbf{y} = \mathbf{D}\boldsymbol{\gamma} + \mathbf{w}, \qquad \|\boldsymbol{\gamma}\|_{0,\infty} \leq s$$

Convolutional Pursuit

$$\min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{D}\boldsymbol{\gamma}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}\|_{0,\infty} \leq s$$

$$\Downarrow$$

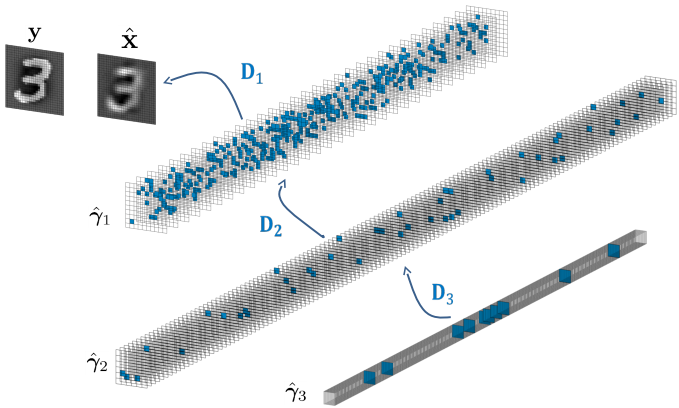
$$\min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{D}\boldsymbol{\gamma}\|_2^2 + \lambda\|\boldsymbol{\gamma}\|_1$$

[Papyan, Sulam, Elad, IEEE TSP 2017]

Multi-Layer Sparse Modeling of Signals

Model

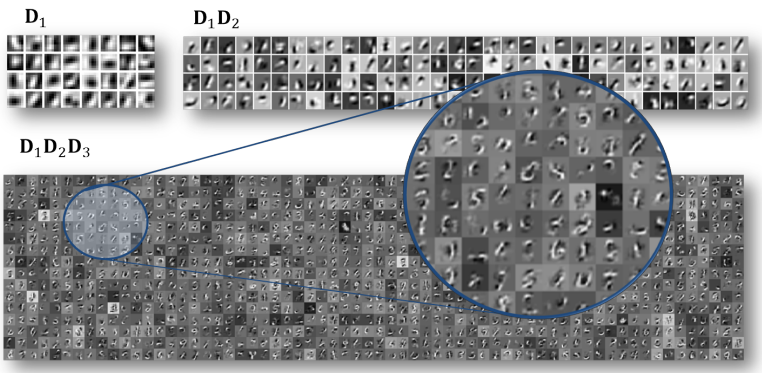
$$\mathbf{y} = \mathbf{D}_1\boldsymbol{\gamma}_1 + \mathbf{w}, \qquad \{\boldsymbol{\gamma}_{i-1} = \mathbf{D}_i\boldsymbol{\gamma}_i, \quad \|\boldsymbol{\gamma}_i\|_{0,\infty} \leq s_i\}_{i=1}^{L}$$



Multi-Layer Sparse Modeling of Signals

Model

$$\mathbf{y} = \mathbf{D}_1\boldsymbol{\gamma}_1 + \mathbf{w}, \qquad \{\boldsymbol{\gamma}_{i-1} = \mathbf{D}_i\boldsymbol{\gamma}_i, \quad \|\boldsymbol{\gamma}_i\|_{0,\infty} \leq s_i\}_{i=1}^{L}$$



Multi-Layer Sparse Modeling of Signals

Model

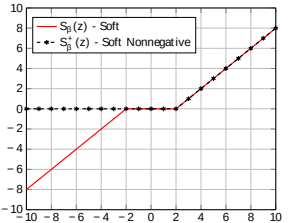
$$\mathbf{y} = \mathbf{D}_1\boldsymbol{\gamma}_1 + \mathbf{w}, \qquad \{\boldsymbol{\gamma}_{i-1} = \mathbf{D}_i\boldsymbol{\gamma}_i, \quad \|\boldsymbol{\gamma}_i\|_{0,\infty} \leq s_i\}_{i=1}^{L}$$

Pursuit: *Deep Coding Problem*

$$\begin{aligned} find \quad \{\boldsymbol{\gamma}_i\}_{i=1}^{L} \qquad \text{s.t.} \qquad & \|\mathbf{y} - \mathbf{D}_1\boldsymbol{\gamma}_1\|_2^2 \leq \varepsilon_0, && \|\boldsymbol{\gamma}_1\|_{0,\infty} \leq s_1 \\ & \|\boldsymbol{\gamma}_1 - \mathbf{D}_2\boldsymbol{\gamma}_2\|_2^2 \leq \varepsilon_1, && \|\boldsymbol{\gamma}_2\|_{0,\infty} \leq s_2 \\ & \qquad\qquad \vdots && \qquad\qquad \vdots \\ & \|\boldsymbol{\gamma}_{L-1} - \mathbf{D}_L\boldsymbol{\gamma}_L\|_2^2 \leq \varepsilon_{L-1}, && \|\boldsymbol{\gamma}_L\|_{0,\infty} \leq s_L, \end{aligned}$$

Multi-Layer Thresholding (a.k.a Forward Pass)

$$\hat{\gamma}_1 = \text{ReLU}(\mathbf{D}_1^T\mathbf{y} + \mathbf{b}_1)$$
$$\hat{\gamma}_2 = \text{ReLU}(\mathbf{D}_2^T\hat{\gamma}_1 + \mathbf{b}_2)$$
$$\vdots$$
$$\hat{\gamma}_L = \text{ReLU}(\mathbf{D}_L^T\hat{\gamma}_{L-1} + \mathbf{b}_L)$$



Stability

If a set of solutions $\{\gamma_i\}_{i=1}^L$ satisfy $\|\gamma_i\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}_i)}\frac{|\gamma_i^{min}|}{|\gamma_i^{max}|}\right) - \frac{1}{\mu(\mathbf{D}_i)}\frac{\epsilon_L^{i-1}}{|\gamma_i^{max}|}$, then

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

[Papyan, Romano, Elad, JMLR 2017]

Multi-Layer Basis Pursuit

$$\hat{\gamma}_1 = \arg\min_{\gamma} \|\mathbf{y} - \mathbf{D}_1\gamma\|_2^2 + \lambda_1\|\gamma\|_1$$

$$\hat{\gamma}_2 = \arg\min_{\gamma} \|\hat{\gamma}_1 - \mathbf{D}_2\gamma\|_2^2 + \lambda_2\|\gamma\|_1$$

$$\ldots$$

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

- Solutions **in the model**?
- What is really the **benefit of the model constraints**?
- What can we get from all this **in practice**?

Multi-Layer Sparse Modeling of Signals

Model

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

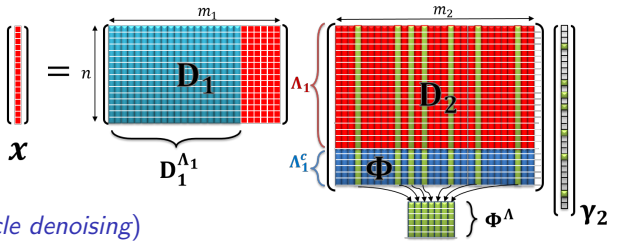
$\mathbf{D}_{(1,L)} = \mathbf{D}_1\mathbf{D}_2\dots\mathbf{D}_L.$

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



Stability (*oracle denoising*)

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

- Synthesis

$$\mathbb{E}\ \|\boldsymbol{\gamma}_2 - \hat{\boldsymbol{\gamma}}_2\|_2^2 \leq \frac{\sigma^2}{1-\delta_{s_2}} s_2$$

- Synthesis-Analysis

$$\mathbb{E}\ \|\boldsymbol{\gamma}_2 - \hat{\boldsymbol{\gamma}}_2\|_2^2 \lesssim \frac{\sigma^2}{1-\delta_{s_2}} (s_2 - rank(\Phi^{\Lambda}))$$

[Aberdam, Sulam, Elad, SIMODS 2019]

Multi-Layer Pursuit

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

Just relax: Multi-Layer Basis Pursuit

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

Related problems

- **Analysis Lasso** [Candes et al, 2010], **Robust Sparse Analysis Regularization** [Vaiter et al, 2012]
- **Inverse Problems with Compound Regularizers** [Figueiredo et al, 2010], [Haeffele et al, 2014]

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**

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

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

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

✓ *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

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

$$\mathbf{z} \leftarrow \boldsymbol{\gamma}_L^{k+1} + \rho^k(\boldsymbol{\gamma}_L^{k+1} - \boldsymbol{\gamma}_L^k)$$

✓ *Nested Proximal-Gradient updates with momentum*

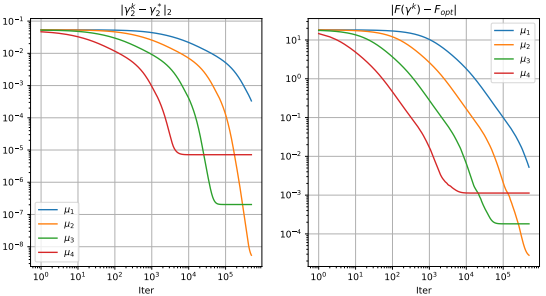
Convergence for ML-ISTA

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.



[Sulam, Aberdam, Beck, Elad, IEEE TPAMI 2019]

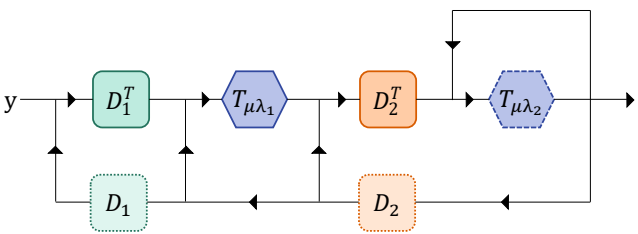
Generalizing Recurrent CNNs

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

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

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

$$\Downarrow$$

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

$\boldsymbol{\gamma}^k$: 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

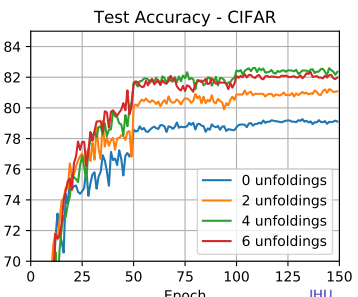
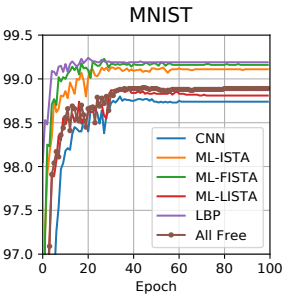
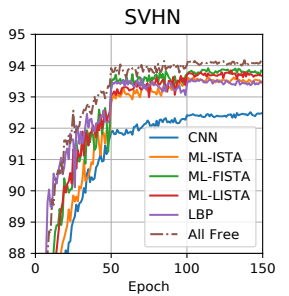
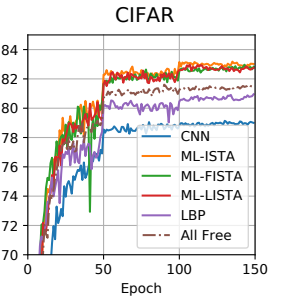


Image Classification







Take-Aways

- Multi-Layer Sparse Model puts forward a generative sparse model and framework for analysis
- Deep learning architectures can be understood analyzed as inference algorithms under this model
- ML-ISTA generalizes feed-forward networks, with improved performance

Joint work with

Aviad Aberdam
Technion

Amir Beck
Tel Aviv University

Yaniv Romano
Stanford

Vardan Papyan
Stanford

Miki Elad
Technion

Johns Hopkins Whiting School of Engineering

Center for Imaging Science, Johns Hopkins University

Johns Hopkins Mathematical Institute for Data Science

"Essentially, all models and wrong, but some are useful"

George E.P. Box