Deep-unrolled Successive Convex Approximation for nonconvex sparse learning

Supervisor: prof. Paolo Di Lorenzo

Examinee: Leonardo Di Nino (1919479)

Academic Year 2023/2024

SAPIENZA
UNIVERSITÀ DI ROMA

Sparsity aware learning

1 The problem

$$\min_{\mathbf{x}} V(\mathbf{x}) = \frac{1}{2}||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2 + \lambda G(\mathbf{x}), \tag{1}$$

$$G(x) = \sum_i g_i(x) \text{ s.t. } \begin{cases} g(0) = 0 \\ g(x) \text{ is non decreasing in } [0, +\infty) \\ \frac{g(x)}{x} \text{ is non increasing in } [0, +\infty) \end{cases} \tag{2}$$

- Core of many learning problems in *signal processing* (compressive sensing, source separation, deblurring and super-resolution), *machine learning* (neural network training and pruning, low-rank matrix models) and *digital communication* (compressive random access, massive-MIMO channel estimation):
- Because of its relevance, many solving strategies have been proposed:
  - Iterative solvers (proximal methods, alternated optimization...);
  - **Deep sparse coders** via deep-unrolling

The classic approach: LASSO formulation

LASSO	Difference-of-convex Learning
Proximal gradient descent	Successive Convex Approximation
$G(\mathbf{x}) = \|\|\mathbf{x}\|\|_1$	$G(\mathbf{x}) = \sum_{i=1}^{m} g(x_i), g(x_i) = \eta(\theta)\|x_i\| - g^{-}(x_i)$
$\mathbf{x}^{k+1} = \mathcal{S}_{\frac{\lambda}{L}}\left[\mathbf{x}^k - \frac{1}{L}\left(\mathbf{A}^T\mathbf{A}\mathbf{x}^k - \mathbf{A}^T\mathbf{y}\right)\right]$	$\mathbf{x}^{k+1} = \mathcal{S}_{\frac{\lambda\eta(\theta)}{L}}\left[\mathbf{x}^k - \frac{1}{L}\left(\mathbf{A}^T\mathbf{A}\mathbf{x}^k - \mathbf{A}^T\mathbf{y} + \lambda\Gamma_{\theta,\gamma}(\mathbf{x}^k)\right)\right]$
$\mathbf{x}^{t+1} = \mathcal{S}_{\beta^t}\{\mathbf{W}_1^t\mathbf{x}^t + \mathbf{W}_2^t\mathbf{y}\}$	?

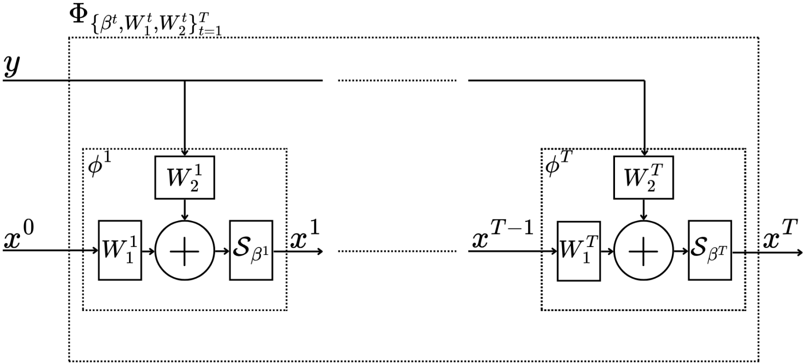


Figure: Structure of an unrolled LISTA network

Nonconvex Sparsity Inducing Penalties: Difference-of-convex Learning

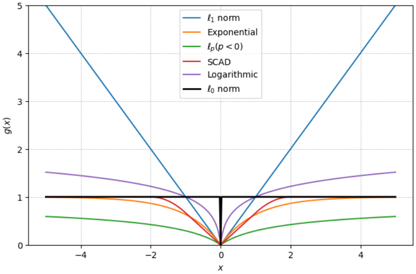


Figure: DC relaxations of ℓ_0 norm

Penalty function	$g(x)$	$\eta(\theta)$
Exp	$1-e^{-\theta\|x\|}$	θ
$\ell_p(p<0)$	$1-(\theta\|x\|+1)^p$	$-p\theta$
SCAD	$\begin{cases} \frac{2\theta}{a+1}\|x\|, & 0 \leq \|x\| \leq \frac{1}{\theta} \\ \frac{-\theta^2\|x\|^2+2a\theta\|x\|-1}{a^2-1}, & \frac{1}{\theta} < \|x\| \leq \frac{a}{\theta} \\ 1, & \|x\| > \frac{a}{\theta} \end{cases}$	$\frac{2\theta}{a+1}$
Log	$\frac{\log(1+\theta\|x\|)}{\log(1+\theta)}$	$\frac{\theta}{\log(1+\theta)}$

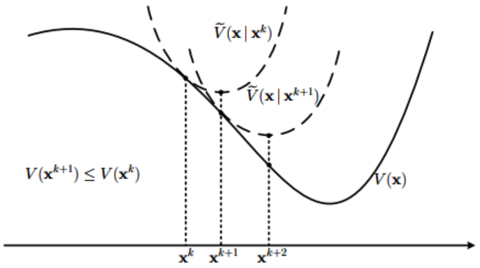
Table: Functional forms for nonconvex relaxations of ℓ_0 norm enjoying a DC structure

SCA aims at solving an optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} V(\mathbf{x}) \tag{3}$$

in an iterative way, defining at each iterate a strongly convex surrogate of the function whose parametrization depends on the design choices we make.

$$\mathbf{x}^{k+1} = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}}\ \widetilde{V}(\mathbf{x}|\mathbf{x}^k, \theta^k) \tag{4}$$



Nonconvex Sparsity Inducing Penalties: Difference-of-convex Learning

2 Methodological background

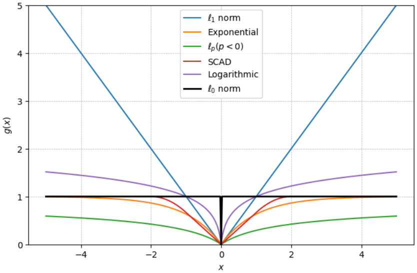


Figure: DC relaxations of ℓ_0 norm

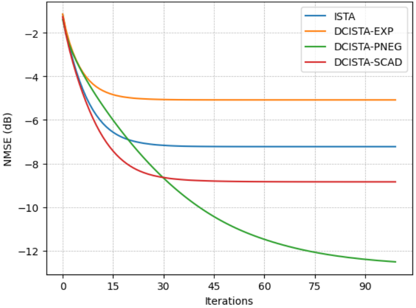


Figure: Reconstruction error for different sparse learning algorithms

LASSO	Difference-of-convex Learning
Proximal gradient descent	Successive Convex Approximation
$G(\mathbf{x}) = \|\|\mathbf{x}\|\|_1$	$G(\mathbf{x}) = \sum_{i=1}^{m} g(x_i), g(x_i) = \eta(\theta)\|x_i\| - g^-(x_i)$
$\mathbf{x}^{k+1} = \mathcal{S}_{\frac{\lambda}{L}}\left[\mathbf{x}^k - \frac{1}{L}\left(\mathbf{A}^T\mathbf{A}\mathbf{x}^k - \mathbf{A}^T\mathbf{y}\right)\right]$	$\mathbf{x}^{k+1} = \mathcal{S}_{\frac{\lambda\eta(\theta)}{L}}\left[\mathbf{x}^k - \frac{1}{L}\left(\mathbf{A}^T\mathbf{A}\mathbf{x}^k - \mathbf{A}^T\mathbf{y} + \lambda\Gamma_{\theta,\gamma}(\mathbf{x}^k)\right)\right]$
$\mathbf{x}^{t+1} = \mathcal{S}_{\beta^t}\{\mathbf{W}_1^t\mathbf{x}^t + \mathbf{W}_2^t\mathbf{y}\}$	**?**

From LISTA to ALISTA

Model	Recursion	Complexity	Key enabler
LISTA[1]	$\mathbf{x}^{t+1} = \mathcal{S}_{\beta^t}\{\mathbf{W}_1^t\mathbf{x}^t + \mathbf{W}_2^t\mathbf{y}\}$	$\mathcal{O}(TM^2 + TNM + T)$	-
LISTA-CPSS[2]	$\mathbf{x}^{t+1} = \mathcal{S}_{\beta^t}\{\mathbf{x}^t + (\mathbf{W}^t)^T(\mathbf{y} - \mathbf{A}\mathbf{x}^t)\}$	$\mathcal{O}(TNM + T)$	Necessary conditions of convergence
TiLISTA[3]	$\mathbf{x}^{t+1} = \mathcal{S}_{\beta^t}\{\mathbf{x}^t + \gamma^t\mathbf{W}^T(\mathbf{y} - \mathbf{A}\mathbf{x}^t)\}$	$\mathcal{O}(NM + T)$	Optimal upper bound on reconstruction error
ALISTA[3]	$\mathbf{x}^{t+1} = \mathcal{S}_{\beta^t}\{\mathbf{x}^t + \gamma^t\overline{\mathbf{W}}^T(\mathbf{y} - \mathbf{A}\mathbf{x}^t)\}$, $\overline{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}}\|\mathbf{W}^T\mathbf{A}\|_F^2,\ \text{s.t.}[\mathbf{W}]_{:,j}^T[\mathbf{A}]_{:,j} = 1, \forall j$	$\mathcal{O}(T)$	

[1] **Karol Gregor and Yann LeCun**, *Learning fast approximations of sparse coding*, 2010.
[2] **Xiaohan Chen, Jialin Liu, Zhangyang Wang, and Wotao Yin**, *Theoretical linear convergence of unfolded ISTA and its practical weights and thresholds*, 2018.
[3] **Jialin Liu, Xiaohan Chen, Zhangyang Wang, and Wotao Yin**, *ALISTA: Analytic weights are as good as learned weights in LISTA*, 2019.

Training the models

- The synthetic experiments are designed to work in a **supervised way**. In particular we design training and test set $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{S}$ in this way:
  - We sample and normalize a gaussian sensing matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, and fix it;
  - We sample gaussian underlying signals $x \in \mathbb{R}^N$ with an assumed fixed underlying sparsity;
  - We generate underdetermined linear observations $\mathbf{y} = \mathbf{A}\mathbf{x} + w$, being w gaussian noise according to a certain SNR.
- The learning task is formulated as $\min\limits_{\Theta} \mathbb{E}\left[||\mathbf{x} - \Phi_{\Theta}(\mathbf{y})||_2^2\right]$;
- Models are trained with a stratified approach on each layer:
  1. We solve for Θ^τ:
  $$\min_{\Theta^\tau} \mathbb{E}[||\mathbf{x} - \Phi_{\{\Theta^t\}_{t=1}^{\tau}}(\mathbf{y})|||_2^2 \tag{5}$$
  2. We fine-tune the network up to the layer τ by solving
  $$\min_{\{\Theta^t\}_{t=1}^{\tau}} \mathbb{E}\left[||\mathbf{x} - \Phi_{\{\Theta^t\}_{t=1}^{\tau}}(\mathbf{y})||_2^2\right. \tag{6}$$

From LISTA to ALISTA

2 Methodological background

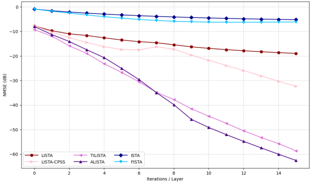


Figure: SNR = ∞

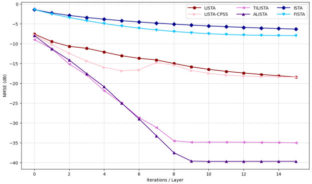


Figure: SNR = 40 db

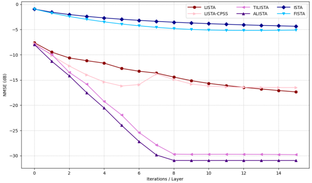


Figure: SNR = 30 db

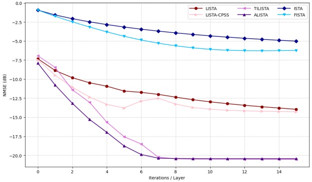


Figure: SNR = 20 db

Learnable Difference of Convex ISTA

3 Proposed models

$$\mathbf{x}^{t+1} = \mathcal{S}_{\frac{\lambda\eta(\theta)}{L}} \left\{ \mathbf{x}^t - \frac{1}{L}\mathbf{A}^T(\mathbf{A}\mathbf{x}^t - \mathbf{y}) + \frac{\lambda}{L}\Gamma_{\theta^t,\gamma^t}(\mathbf{x}^t) \right\}$$

- **Ablation study** for the parametrization
- **Proved differentiability** of the newly built layers

$$\mathbf{x}^{t+1} = \mathcal{S}_{\lambda^t\eta(\theta^t)} \left\{ \mathbf{W}_1^t\mathbf{y} + \mathbf{W}_2^t\mathbf{x}^t + \lambda^t\Gamma_{\theta^t,\gamma^t}(\mathbf{x}^t) \right\}$$

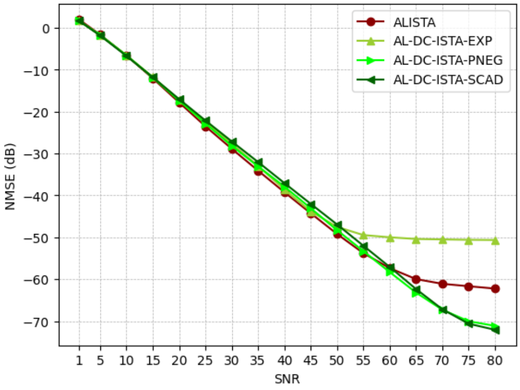
- Proved **necessary conditions of convergence;**
- Proved **optimal upper bound on the reconstruction error**

$$\left.\begin{aligned} \lim_{t\to\infty} \lambda^t\eta(\theta^t) = 0 \\ \lim_{t\to\infty} \mathbf{W}_2^t - (\mathbb{I} - \mathbf{W}_1^t A) = 0 \\ ||x^t - x^*||_2 \le sB\exp\left(-\sum_{k=0}^{t-1} c^k\right),\ t = 1, 2, \dots \end{aligned}\right\} \sup\left\{ \left|\frac{dg^-_{\theta,\gamma}(x)}{dx}\right| \right\} = \eta(\theta,\gamma)$$

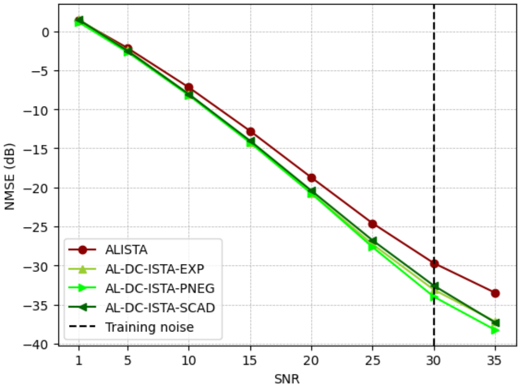
$$\mathbf{x}^{t+1} = \mathcal{S}_{\lambda^t\eta(\theta^t)} \left\{ \mathbf{x}^t + \zeta^t\overline{\mathbf{W}}^T(\mathbf{y} - \mathbf{A}\mathbf{x}^t) + \lambda^t\Gamma_{\theta^t,\gamma^t}(\mathbf{x}^t) \right\}$$

$$\overline{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}} ||\mathbf{W}^T\mathbf{A}||_F^2,\ \text{s. t. } [\mathbf{W}]_{:,j}^T[\mathbf{A}]_{:,j} = 1, \forall\, j$$

Synthetic data results: reconstruction error

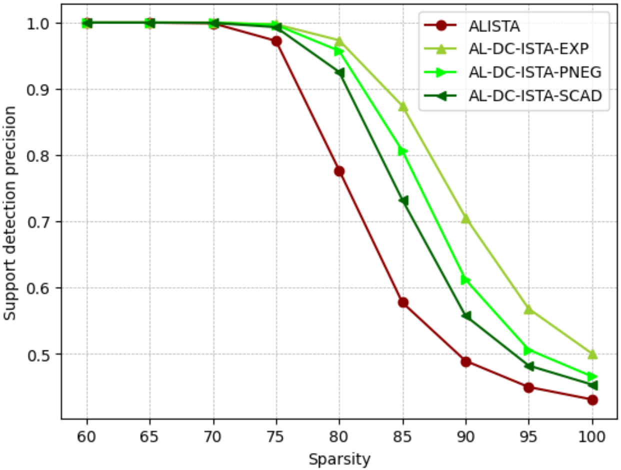


Noiseless training, noisy test



Noisy training, noisy test

Synthetic data results: precision in support detection



Real data results: reconstruction error in image denoising

- **BSD500** (grayscaled, normalized, (16 x 16) patched and vectorized);
- Image denoising (patch-wise);
- $\mathbf{A} = (\mathcal{H}_{256}|\mathbb{I}_{256})$;
- Models Φ_Θ are trained defining the following loss $\min_\Theta \mathbb{E}\left[||\mathbf{y} - \mathbf{A}\Phi_\Theta(\tilde{\mathbf{y}})||_2^2\right]$, given a clean patch $\mathbf{y}$ and its noisy version $\tilde{\mathbf{y}}$:

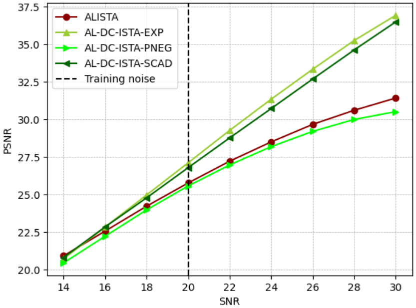


Figure: PSNR when varying SNR in test set

- We defined a **new class of sparse coders hinging on difference-of-convex relaxations of ℓ_0 norm**;
- We connected these models to the state-of-the-art **proving theoretical results**;
- We provided a model, AL-DC-ISTA, which is **capable of outperforming ALISTA** for what concerns certain metrics;
- We provided a **primer** on model-based deep learning architectures leveraging Successive Convex Approximation framework.

...and departures

- Enlarge the framework to **convolutional sparse coding** to natively support structured data (images, sequences, graphs);
- Make the architecture robust to **statistical model perturbations** or even whole **entangled dynamics**;
- Define **neural-building-blocks architecture** hinging on **modularity** to enable **block decompositions** and support **high-dimensional learning**, possibly in **distributed fashion**.

Thank you for listening!
Any questions?