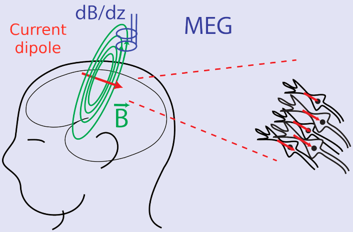
Learning step sizes for unfolded sparse coding

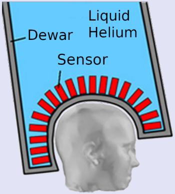
Thomas Moreau INRIA Saclay

Joint work with Pierre Ablin; Mathurin Massias; Alexandre Gramfort

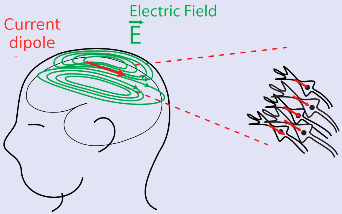
Electrophysiology

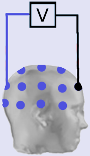
Magnetoencephalography

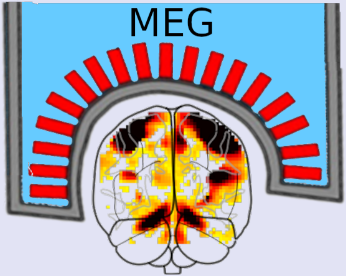




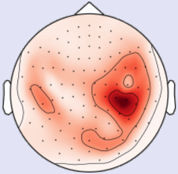
Electroencephalography







Maxwell's Equations $\longrightarrow$



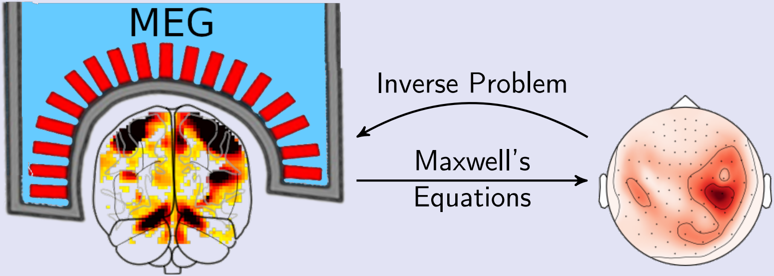
$\boxed{z}$ Electrical activity

$\boxed{D}$

$\boxed{x}$ Observed signal

Forward model: $x = Dz$

Inverse problems



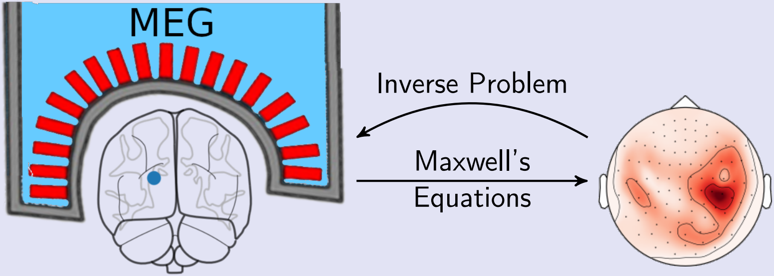
$\boxed{z}$ Electrical activity

$\boxed{D}$

$\boxed{x}$ Observed signal

Forward model: $x = Dz$

Inverse problem: $z = f(x)$ (ill-posed)



$\boxed{z}$ $\boxed{D}$ $\boxed{x}$

Electrical activity **Observed signal**

Forward model: $x = Dz$ **Inverse problem:** $z = f(x)$ (ill-posed)

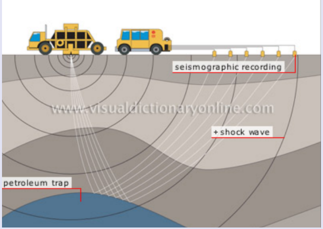
Optimization with a regularization $\mathcal{R}$ encoding prior knowledge

$$\operatorname{argmin}_z \|x - Dz\|_2^2 + \mathcal{R}(z)$$

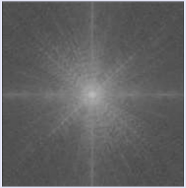
Example: sparsity with $\mathcal{R} = \lambda \| \cdot \|_1$

Other inverse problems

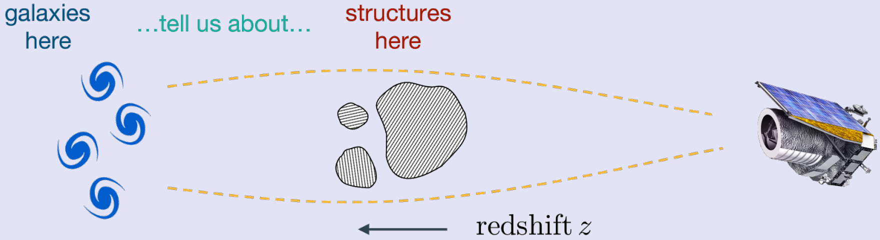
Ultra sound



fMRI - compress sensing



Astrophysic



Some challenges for inverse problems

Evaluation: often there is no ground truth,

- In neuroscience, we cannot access the brain electrical activity.
- How to evaluate how well it is reconstructed?

Open problem in unsupervised learning

Modelization: how to better account for the signal structure,

- ℓ_2 reconstruction evaluation does not account for localization
- Optimal transport could help in this case?

Computational: solving these problems can be too long,

- Many problems share the same forward operator $\boldsymbol{D}$
- Can we use the structure of the problem?

Today's talk topic!

Better step sizes for
Iterative Shrinkage-Thresholding Algorithm (ISTA)

For a fixed design matrix $D \in \mathbb{R}^{n\times m}$ and $\lambda > 0$, the Lasso for $x \in \mathbb{R}^n$ is

$$z^* = \underset{z}{\operatorname{argmin}}\, F_x(z) = \underbrace{\frac{1}{2}\|x - Dz\|_2^2}_{f_x(z)} + \lambda\|z\|_1$$

a.k.a. sparse coding, sparse linear regression, ...

We are interested in the over-complete case where $m > n$.

The Lasso

For a fixed design matrix $D \in \mathbb{R}^{n \times m}$ and $\lambda > 0$, the Lasso for $x \in \mathbb{R}^n$ is

$$z^* = \operatorname*{argmin}_z F_x(z) = \underbrace{\frac{1}{2}\|x - Dz\|_2^2}_{f_x(z)} + \lambda \|z\|_1$$

a.k.a. sparse coding, sparse linear regression, ...

We are interested in the over-complete case where $m > n$.

Properties

- The problem is convex in z but not strongly convex in general
- $z = 0$ is solution if and only if $\lambda \geq \lambda_{\max} \doteq \|D^\top x\|_\infty$

ISTA: Iterative Shrinkage-Thresholding Algorithm

[Daubechies et al. 2004]

f_x is a L-smooth function with $L = \|D\|_2^2$ and

$$\nabla f_x(z^{(t)}) = D^\top(Dz^{(t)} - x)$$

The ℓ_1-norm is proximable with a separable proximal operator

$$\mathrm{prox}_{\mu\|\cdot\|_1}(x) = \mathrm{sign}(x)\max(0, |x| - \mu) = ST(x, \mu)$$

ISTA: Iterative Shrinkage-Thresholding Algorithm

[Daubechies et al. 2004]

f_x is a L-smooth function with $L = \|D\|_2^2$ and

$$\nabla f_x(z^{(t)}) = D^\top(Dz^{(t)} - x)$$

The ℓ_1-norm is proximable with a separable proximal operator

$$\mathrm{prox}_{\mu\|\cdot\|_1}(x) = \mathrm{sign}(x)\max(0, |x| - \mu) = ST(x, \mu)$$

We can use the proximal gradient descent algorithm (ISTA)

$$z^{(t+1)} = \mathrm{ST}\left(z^{(t)} - \rho\underbrace{\nabla f_x(z^{(t)})}_{D^\top(Dz^{(t)}-x)}, \rho\lambda\right)$$

Here, ρ play the role of a step size (in $[0, \frac{2}{L}[$).

ISTA: Majoration-Minimization

Taylor expansion of f_x in $z^{(t)}$

$$
\begin{aligned}
F_x(z) &= f_x(z^{(t)}) + \nabla f_x(z^{(t)})^\top (z - z^{(t)}) + \frac{1}{2}\|D(z - z^{(t)})\|_2^2 + \lambda\|z\|_1 \\
&\leq f_x(z^{(t)}) + \nabla f_x(z^{(t)})^\top (z - z^{(t)}) + \frac{L}{2}\|z - z^{(t)}\|_2^2 + \lambda\|z\|_1
\end{aligned}
$$

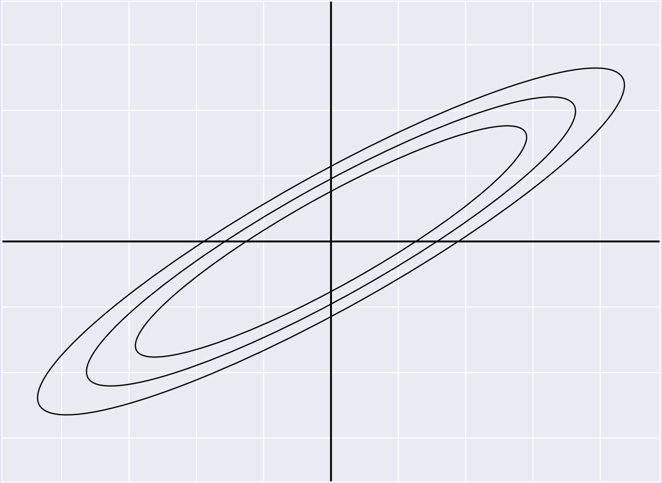
$\Rightarrow$ Replace the Hessian $D^\top D$ by L **Id**.

Separable function that can be minimized in close form

$$
\begin{aligned}
\operatorname*{argmin}_z \frac{L}{2}\left\| z^{(t)} - \frac{1}{L}\nabla f_x(z^{(t)}) - z \right\|_2^2 + \lambda\|z\|_1 &= \mathrm{ST}\left(z^{(t)} - \frac{1}{L}\nabla f_x(z^{(t)}), \frac{\lambda}{L} \right) \\
&= \mathrm{prox}_{\frac{\lambda}{L}}\left(z^{(t)} - \frac{1}{L}\nabla f_x(z^{(t)}) \right)
\end{aligned}
$$

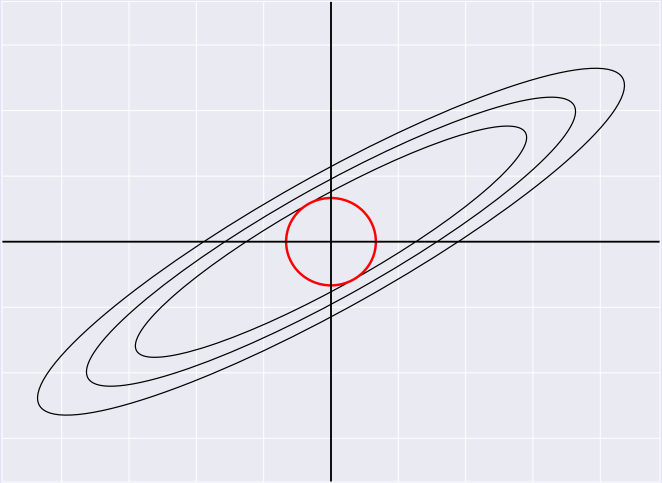
ISTA: Majoration for the data-fit

- Level lines form $z^\top D^\top Dz$



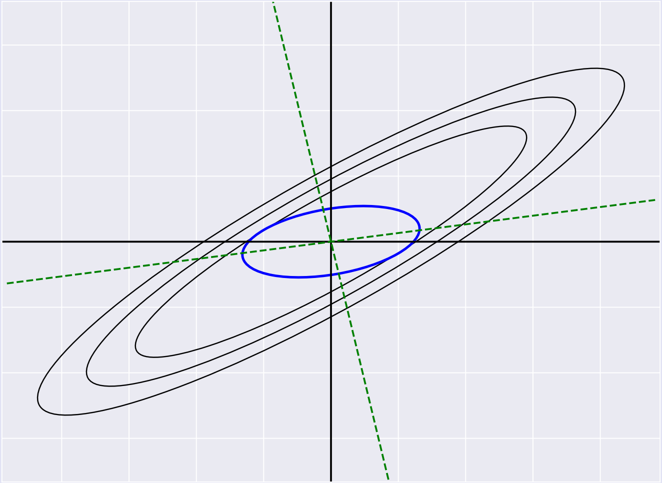
ISTA: Majoration for the data-fit

- Level lines form $z^\top D^\top Dz \leq L\|z\|_2$



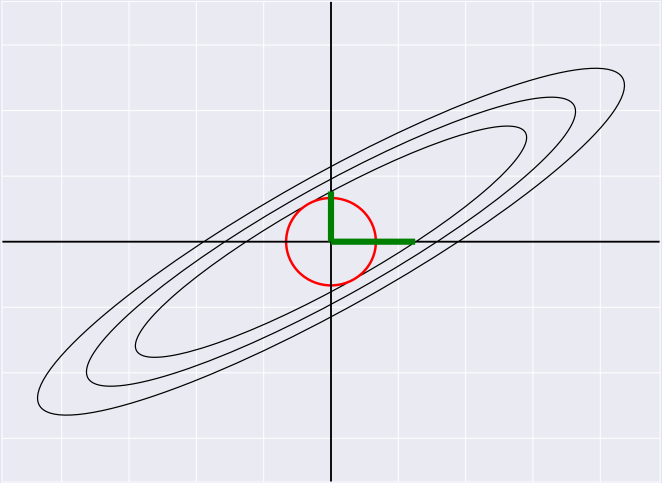
ISTA: Majoration for the data-fit

- Level lines form $z^\top D^\top D z \leq z^\top A^\top \Lambda A z$ [Moreau and Bruna 2017]



ISTA: Majoration for the data-fit

- Level lines form $z^\top D^\top D z \leq \color{green}{L_S \|z\|_2}$ for $Supp(z) \subset S$

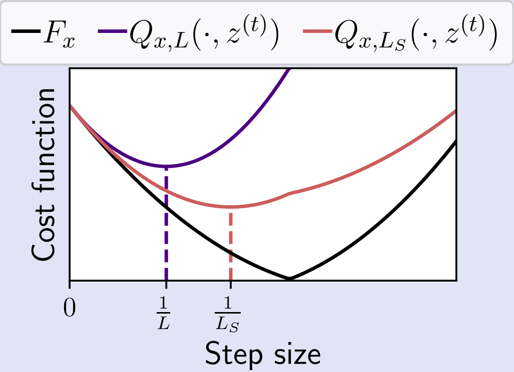


Oracle ISTA: Majoration-Minimization

For all z such that $\mathrm{Supp}(z) \subset S \doteq \mathrm{Supp}(z^{(t)})$,

$$F_x(z) \leq f_x(z^{(t)}) + \nabla f_x(z^{(t)})^\top (z - z^{(t)}) + \frac{L_S}{2}\|z - z^{(t)}\|_2^2 + \lambda\|z\|_1$$

with $L_S = \|D_{\cdot,S}\|_2^2$.



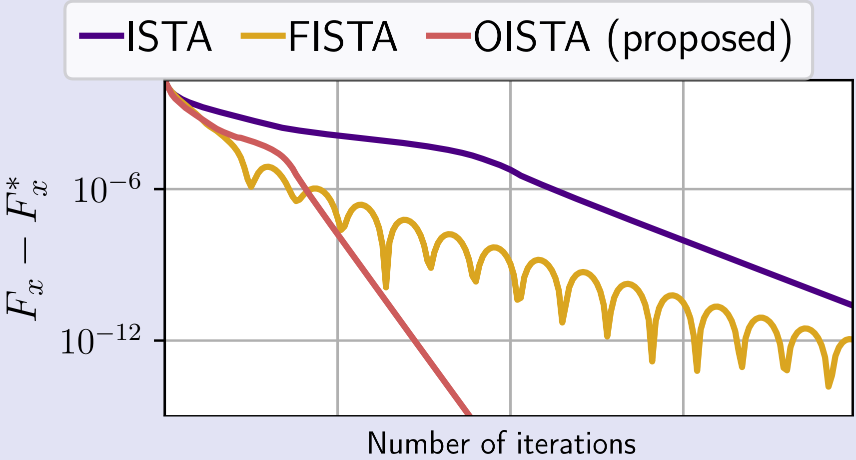
Oracle ISTA (OISTA):

1. Get the Lipschitz constant L_S associated with support $S = \mathrm{Supp}(z^{(t)})$.
2. Compute $y^{(t+1)}$ as a step of ISTA with a step-size of $1/L_S$

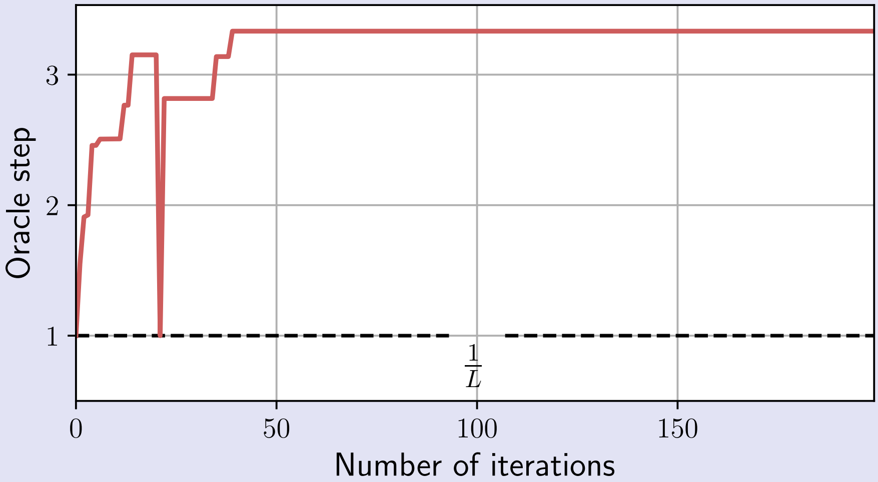
$$y^{(t+1)} = \mathrm{ST}\left(z^{(t)} - \frac{1}{L_S} D^\top (Dz^{(t)} - x), \frac{\lambda}{L_S}\right)$$

3. If $\mathrm{Supp}(y^{t+1}) \subset S$, accept the update $z^{(t+1)} = y^{(t+1)}$.
4. Else, $z^{(t+1)}$ is computed with step size $1/L$.

OISTA: Performances



OISTA – Step-size



OISTA – Improved-convergence rates

$S^* = Supp(Z^*)$

$\mu^* = \min \|Dz\|_2^2$ for $\|z\|_2 = 1$ and $Supp(z) \subset S^*$.

If $\mu^* > 0$, OISTA converges with a linear rate

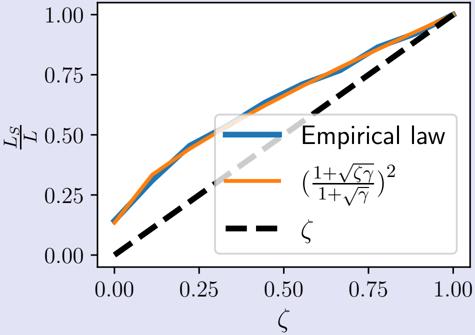
$$F_x(z^{(t)}) - F_x(z^*) \leq (1 - \frac{\mu^*}{L_{S^*}})^{t-T^*}(F_x(z^{(T^*)}) - F_x(z^*)) \ .$$

OISTA – Gaussian setting

Acceleration quantification with Marchenko-Pastur

Entries in $D \in \mathbb{R}^{n\times m}$ are sampled from $\mathcal{N}(0,1)$ and S is sampled uniformly with $|S| = k$. Denote $m/n \to \gamma$, $k/m \to \zeta$, with $k, m, n \to +\infty$. Then

$$\frac{L_S}{L} \to \left(\frac{1+\sqrt{\zeta\gamma}}{1+\sqrt{\gamma}}\right)^2 \quad . \tag{1}$$



Empirical law

$n = 200, \quad m = 600$

OISTA – Limitation

- In practice, OISTA is not practical, as you need to compute L_S at each iteration and this is costly.
- No precomputation possible: there is an exponential number of supports S.

Using deep learning to approximate OISTA

Solving the Lasso many times

Assume that we want to solve the Lasso for many observation $\{x_1, \dots, x_N\}$ with a fixed direct operator D *i.e.*for each x computes

$$\mathcal{I}_D(x) = \operatorname*{argmin}_z \frac{1}{2}\|x - Dz\| + \lambda\|z\|_1$$

Thus, the goal is not to solve **one** problem but **multiple** problems.

$\Rightarrow$ Can we leverage the problem's structure?

- **ISTA**: worst case algorithm, second order information is L.
- **OISTA**: adaptive algorithm, second order information is L_S (NP-hard).
- **LISTA**: adaptive algorithm, use DL to adapt to second order information?

ISTA is a Neural Network

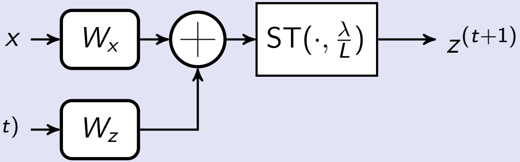
ISTA

$$z^{(t+1)} = \mathrm{ST}\left(z^{(t)} - \frac{1}{L}D^\top(Dz^{(t)} - x), \frac{\lambda}{L}\right)$$

Let $W_z = I_m - \frac{1}{L}D^\top D$ and $W_x = \frac{1}{L}D^\top$. Then

$$z^{(t+1)} = \mathrm{ST}(W_z z^{(t)} + W_x x, \frac{\lambda}{L})$$

One step of ISTA



ISTA is a Neural Network

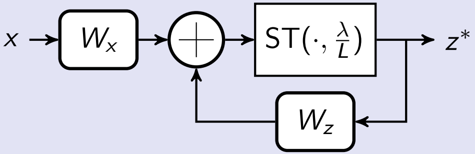
ISTA

$$z^{(t+1)} = \mathrm{ST}\left(z^{(t)} - \frac{1}{L} D^\top (D z^{(t)} - x), \frac{\lambda}{L}\right)$$

Let $W_z = I_m - \frac{1}{L} D^\top D$ and $W_x = \frac{1}{L} D^\top$. Then

$$z^{(t+1)} = \mathrm{ST}(W_z z^{(t)} + W_x x, \frac{\lambda}{L})$$

RNN equivalent to ISTA

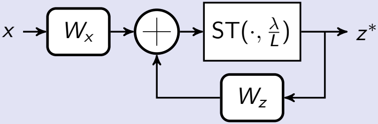


Learned ISTA

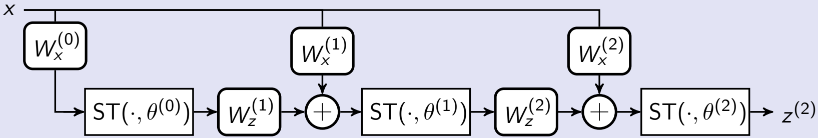
[Gregor and Le Cun 2010]

Recurrence relation of ISTA define a RNN

$$z^{(t+1)} = \mathrm{ST}\left(z^{(t)} - \frac{1}{L}D^\top(Dz^{(t)} - x), \frac{\lambda}{L}\right)$$



This RNN can be unfolded as a feed-forward network.



Let $\Phi_{\Theta^{(T)}}$ denote a network with T layers parametrized with $\Theta^{(T)}$.

If $W_x^{(i)} = W_x$ and $W_z^{(i)} = W_z$, then $\Phi_{\Theta^T}(x) = z^{(t)}$.

Empirical risk minimization : We need a training set of $\{x_1, \dots x_N$ training sample and our goad is to accelerate ISTA on unseen data $x \sim p$.

The training solves

$$\tilde{\Theta}^{(T)} \in \arg\min_{\Theta^{(T)}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_x(\Phi_{\Theta^{(T)}}(x_i)) \ .$$

for a loss $\mathcal{L}_x$.

$\Rightarrow$ Choice of loss $\mathcal{L}_x$?

LISTA – Training

Supervised: a ground truth $z^*(x)$ is known

$$\mathcal{L}_x(z) = \frac{1}{2}\|z - z^*(x)\|$$

Solving the inverse problem.

LISTA – Training

Supervised: a ground truth $z^*(x)$ is known

$$\mathcal{L}_x(z) = \frac{1}{2}\|z - z^*(x)\|$$

Solving the inverse problem.

Semi-supervised: the solution of the Lasso $z^*(x)$ is known

$$\mathcal{L}_x(z) = \frac{1}{2}\|z - z^*(x)\|$$

Accelerating the resolution of the Lasso.

LISTA – Training

Supervised: a ground truth $z^*(x)$ is known

$$\mathcal{L}_x(z) = \frac{1}{2}\|z - z^*(x)\|$$

Solving the inverse problem.

Semi-supervised: the solution of the Lasso $z^*(x)$ is known

$$\mathcal{L}_x(z) = \frac{1}{2}\|z - z^*(x)\|$$

Accelerating the resolution of the Lasso.

Unsupervised: there is no ground truth

$$\mathcal{L}_x(z) = \frac{1}{2}\|x - Dz\|_2^2 + \lambda\|z\|_1$$

Solving the Lasso.

LISTA – Training

Supervised: a ground truth $z^*(x)$ is known

$$\mathcal{L}_x(z) = \frac{1}{2}\|z - z^*(x)\|$$

Solving the inverse problem.

Semi-supervised: the solution of the Lasso $z^*(x)$ is known

$$\mathcal{L}_x(z) = \frac{1}{2}\|z - z^*(x)\|$$

Accelerating the resolution of the Lasso.

Unsupervised: there is no ground truth

$$\mathcal{L}_x(z) = \frac{1}{2}\|x - Dz\|_2^2 + \lambda\|z\|_1$$

Solving the Lasso.

LISTA – Parametrizations

General LISTA model [Gregor and Le Cun 2010]

$$z^{(t+1)} = \mathrm{ST}\left(\mathbf{W_e^{(t)}} z^{(t)} + \mathbf{W_x^{(t)}} x, \theta^{(\mathbf{t})}\right)$$

The structure of D is lost in the linear transform.

Coupled LISTA [Chen et al. 2018]

$$z^{(t+1)} = \mathrm{ST}\left(z^{(t)} - \alpha^{(\mathbf{t})}\mathbf{W^{(t)}}(Dz^{(t)} - x), \beta^{(\mathbf{t})}\right)$$

Can be seen as learning

- Pre-conditionner
  $W^{(t)} \in \mathbb{R}^{m \times n}$
- Step-size
  $\alpha^{(t)} \in \mathbb{R}_+$
- Threshold
  $\beta^{(t)} \in \mathbb{R}_+$

General LISTA model [Gregor and Le Cun 2010]

$$z^{(t+1)} = \mathrm{ST}\left(\mathbf{W_e^{(t)}} z^{(t)} + \mathbf{W_x^{(t)}} x, \theta^{(t)}\right)$$

The structure of D is lost in the linear transform.

Coupled LISTA [Chen et al. 2018]

$$z^{(t+1)} = \mathrm{ST}\left(z^{(t)} - \alpha^{(\mathbf{t})} \mathbf{W}^{(\mathbf{t})}(Dz^{(t)} - x), \beta^{(\mathbf{t})}\right)$$

Can be seen as learning

- Pre-conditionner $W^{(t)} \in \mathbb{R}^{m \times n}$
- Step-size $\alpha^{(t)} \in \mathbb{R}_+$
- Threshold $\beta^{(t)} \in \mathbb{R}_+$

$\Rightarrow$ Justified theoretically for (un)supervised convergence

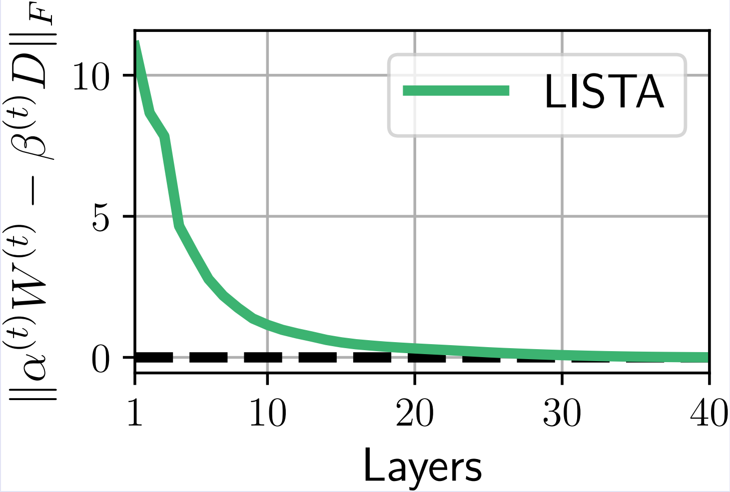
Theorem – Asymptotic convergence of the weights

Consider a sequence of nested networks $\Phi_{\Theta^{(T)}}$ *s.t.* $\Phi_{\Theta^{(t)}}(x) = \phi_{\theta^{(t)}}(\Phi_{\Theta^{(t+1)}}(x), x)$. Assume that

1. the sequence of parameters converges *i.e.* $\theta^{(t)} \xrightarrow[t\to\infty]{} \theta^* = (W^*, \alpha^*, \beta^*)$,
2. the output of the network converges toward a solution $z^*(x)$ of the Lasso uniformly over the equiregularization set $\mathcal{B}_\infty$, *i.e.* $\sup_{x\in\mathcal{B}_\infty} \|\Phi_{\Theta^{(T)}}(x) - z^*(x)\| \xrightarrow[T\to\infty]{} 0$.

Then $\frac{\alpha^*}{\beta^*} W^* = D$.

Sad result: "The deep layers of LISTA only learn a better step size".



40-layers LISTA network trained on a 10×20 problem with $\lambda = 0.1$

The weights $W^{(t)}$ align with D and α, β get coupled.

Restricted parametrization : Only learn a step-size $\alpha^{(t)}$

$$z^{(t+1)} = \mathrm{ST}\left(z^{(t)} - \alpha^{(\mathbf{t})} D^\top (Dz^{(t)} - x), \lambda\alpha^{(\mathbf{t})}\right)$$

Fewer parameters: T instead of $(2 + mn)T$.

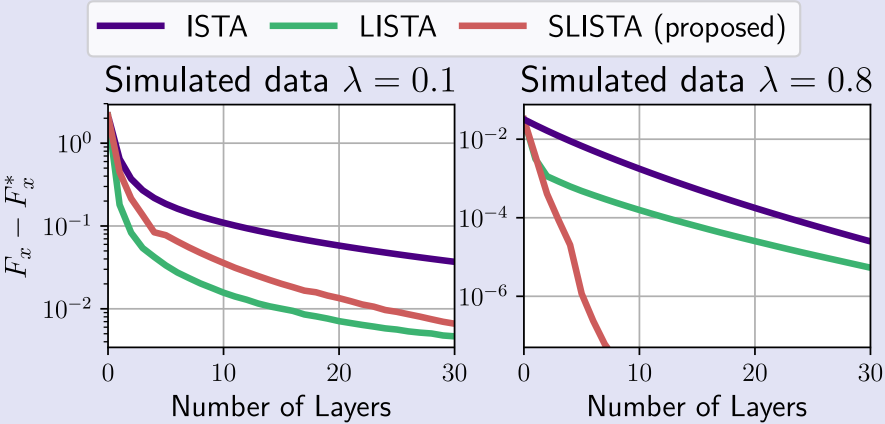
$\Rightarrow$ Easier to learn

$\Rightarrow$ Reduced performances?

Goal: Learn adapted step sizes for ISTA.

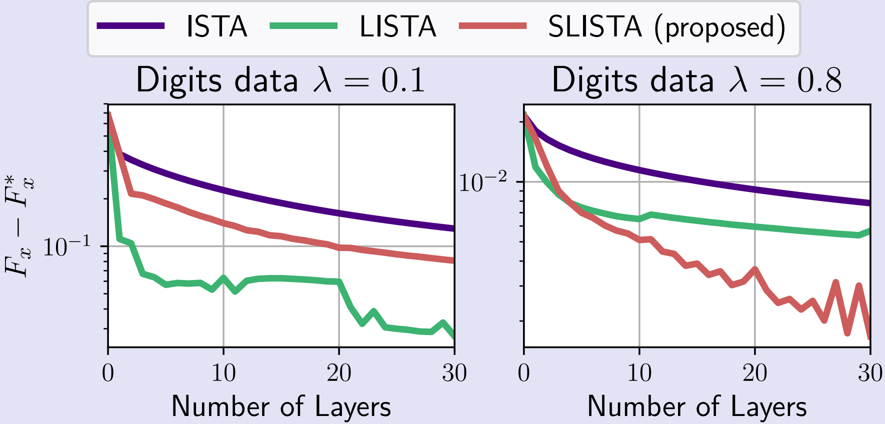
Simulated data: $m = 256$ and $n = 64$

$$D_k \sim \mathcal{U}(\mathcal{S}^{n-1}) \text{ and } x = \frac{\widetilde{x}}{\|D^\top \widetilde{x}\|_\infty} \text{ with } \widetilde{x}_i \sim \mathcal{N}(0,1)$$

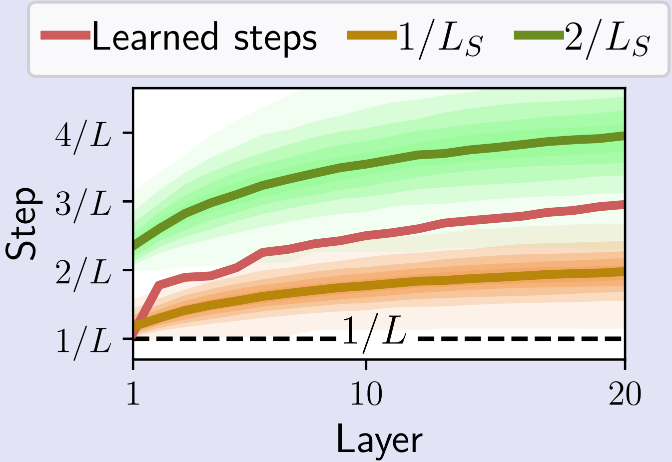


Digits: 8×8 images [Pedregosa et al. 2011]

$$D_k \text{ and } \widetilde{x} \text{ sampled uniformly from the digits and } x = \frac{\widetilde{x}}{\|D^\top \widetilde{x}\|_\infty}.$$



Link with OISTA



The learned step-sizes are linked to the distribution of $1/L_S$

Conclusion

- Using $1/L$ as a step size is not always the fastest.
- Structure of the sparsity can help choose a better step size.
- This structure can be accessed with DL.

Take home message:

First order structure is needed in optimization.
No hope to learn an algorithm better than ISTA.

(except for step-sizes!)

Conclusion

Related work: [Moreau and Bruna 2017]

- It is possible to find a better starting point for ISTA.
- There exists some adversarial cases.
- It is harder and harder as you get closer to the solution.

Code to reproduce the figures is available online:

adopty : github.com/tommoral/adopty

Slides are on my web page:

tommoral.github.io @tomamoral

ISTA – Convergence

Convergence rates

If f_x is μ-strongly convex, *i.e.* $\sigma_{\min}(D^T D) \geq \mu > 0$

$$F_x(z^{(t)}) - F_x(z^*) \leq \left(1 - \frac{\mu}{L}\right)^t (F_x(0) - F_x(z^*))$$

In the general case, $F_x(z^{(t)}) - F_x(z^*) \leq \frac{L\|z^*\|_2}{t}$

OISTA – Convergence

Proposition 3.1: Convergence

When D is such that the solution is unique for all x and $\lambda > 0$, the sequence $(z^{(t)})$ generated by the algorithm converges to $z^* = \operatorname{argmin} F_x$.
Further, there exists an iteration T^* such that for $t \geq T^*$, $\operatorname{Supp}(z^{(t)}) = \operatorname{Supp}(z^*) \triangleq S^*$.

Proposition 3.2: Convergence rate

For $t > T^*$,

$$F_x(z^{(t)}) - F_x(z^*) \leq L_{S^*} \frac{\|z^* - z^{(T^*)}\|^2}{2(t - T^*)} \ .$$

If moreover, $\lambda_{\min}(D_{S^*}^\top D_{S^*}) = \mu^* > 0$, then

$$F_x(z^{(t)}) - F_x(z^*) \leq (1 - \frac{\mu^*}{L_{S^*}})^{t - T^*}(F_x(z^{(T^*)}) - F_x(z^*)) \ .$$

Interlude – regularization λ

Importance of the parameter λ

$$\mathcal{L}_x(z) = \frac{1}{2}\|x - Dz\|_2^2 + \lambda\|z\|_1$$

$$z^{(t+1)} = \mathrm{ST}\left(z^{(t)} - \alpha^{(\mathbf{t})} D^\top (Dz^{(t)} - x), \lambda\alpha^{(\mathbf{t})}\right)$$

Control the distribution of $z^*(x)$ sparsity.

Maximal value

$\lambda_{\max} = \|D^\top x\|_\infty$ is the minimal value of λ for which

$$z^*(x) = 0$$

Equiregularization set

Set in $\mathbb{R}^n$ for which $\lambda_{\max} = 1$

$$\mathcal{B}_\infty = \{x \in \mathbb{R}^n \; ; \quad \|D^\top x\|_\infty = 1\}$$

$\Rightarrow$ Training performed with points sampled in $\mathcal{B}_\infty$