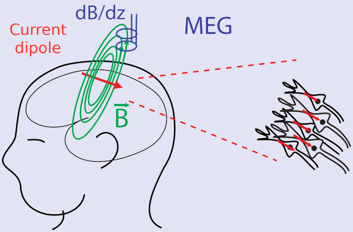
# Learning step sizes for unfolded sparse coding

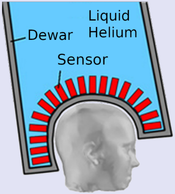
Thomas Moreau INRIA Saclay

Joint work with Pierre Ablin; Mathurin Massias; Alexandre Gramfort

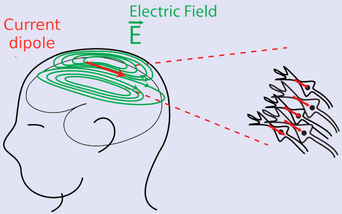
# Electrophysiology

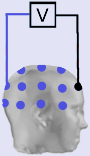
## Magnetoencephalography



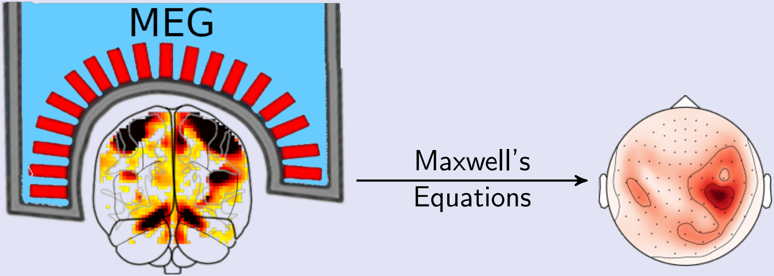




## Electroencephalography



# Inverse problems



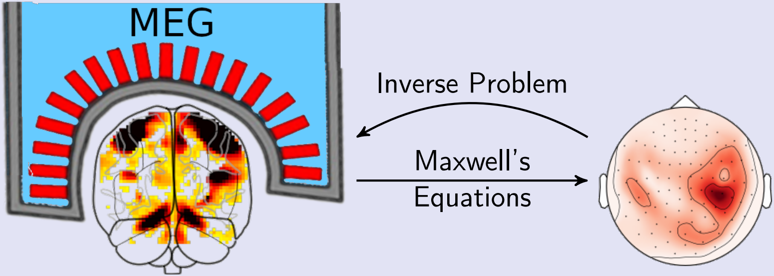

$\boxed{z}$ Electrical activity

$\boxed{D}$

$\boxed{x}$ Observed signal

Forward model: $x = Dz$

# Inverse problems




$\boxed{z}$ Electrical activity

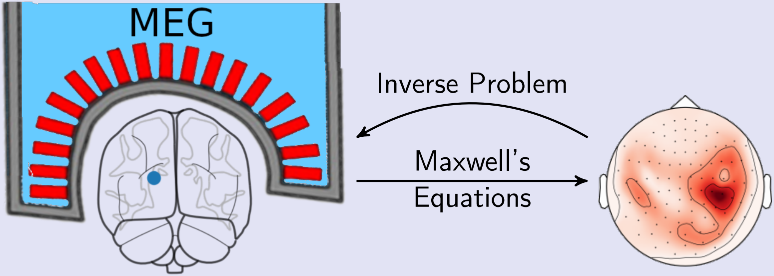
$\boxed{D}$

$\boxed{x}$ Observed signal

**Forward model:** $x = Dz$

**Inverse problem:** ill-posed

# Inverse problems




$\boxed{z}$ **Electrical activity**

$\boxed{D}$

$\boxed{x}$ **Observed signal**

**Forward model:** $x = Dz$

**Inverse problem:** ill-posed

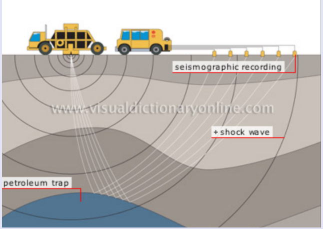
Optimization with a regularization $\mathcal{R}$ encoding prior knowledge

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

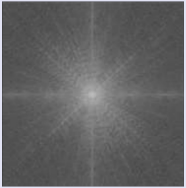
Example: sparsity with $\mathcal{R} = \lambda \| \cdot \|_1$

# Inverse problem: Other domains

## Ultra sound

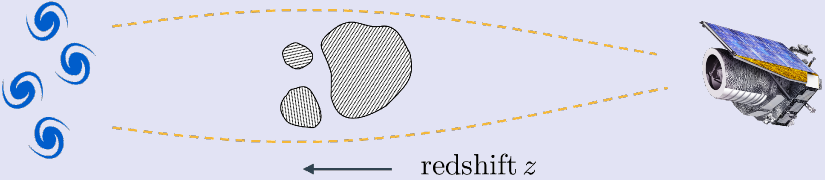



## fMRI - compress sensing



## Astrophysic

galaxies here ...tell us about... structures here



# 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?

Part of my research topic

**Modelization:** how to better account for the image structure,

- $\ell_2$ reconstruction evaluation does not account for localization
- Optimal transport could help in this case?

Hicham and Quentin projects

**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 talk topic!

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

# Sparse Coding

For a dictionary $D \in \mathbb{R}^{n \times m}$ and $\lambda > 0$, sparse coding 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.* Lasso, sparse linear regression, ...

We are interested in the 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 $\lambda \geq \lambda_{\text{max}} \doteq \|D^\top x\|_\infty$

# Iterative Shrinkage-Thresholding Algorithm [Daubechies et al. 2004]

Proximal gradient descent algorithm

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

where $L = \|D^\top D\|_2$ is the largest eigen-value of $D^\top D$.
Here, $1/L$ play the role of a step size.

## Convergence rates

If $f_x$ is $\mu$-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, $F_x(z^{(t)}) - F_x(z^*) \leq \frac{L\|z^*\|_2}{t}$

# ISTA: Majoration-Minimization

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

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

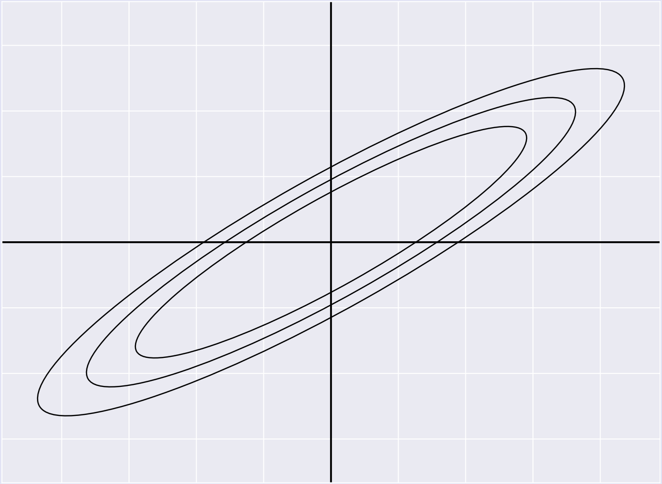
Replace the Hessian $D^\top D$ by $L$ **Id**.

Separable function that can be minimized in close form

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

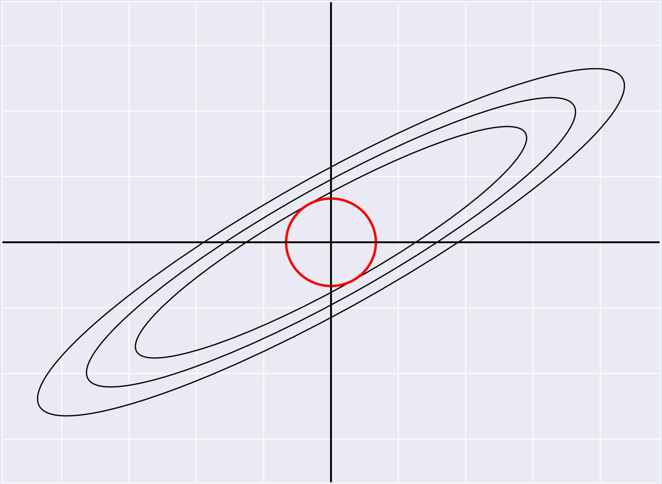
# ISTA: Majoration for the data-fit

- Hessian $D^\top D$



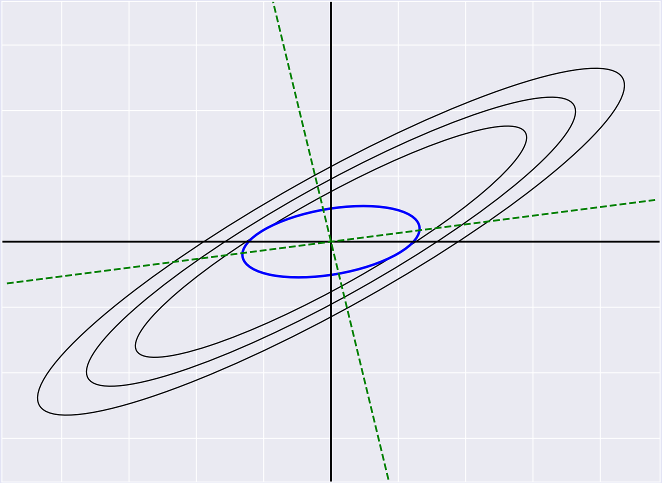
# ISTA: Majoration for the data-fit

- Hessian $D^\top D \prec L\ \mathbf{Id}$



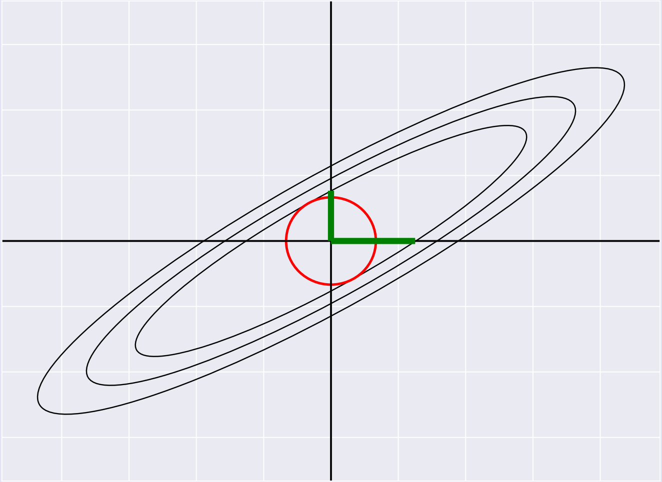
# ISTA: Majoration for the data-fit

- Hessian $D^\top D \prec A^\top \Lambda A$ [Moreau and Bruna 2017]



# ISTA: Majoration for the data-fit

- Hessian $D^\top D \prec L_S\ \mathbf{Id}$ on support $S$

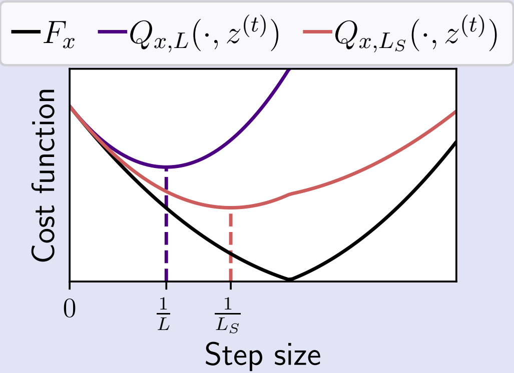


# OISTA: Majoration-Minimization

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



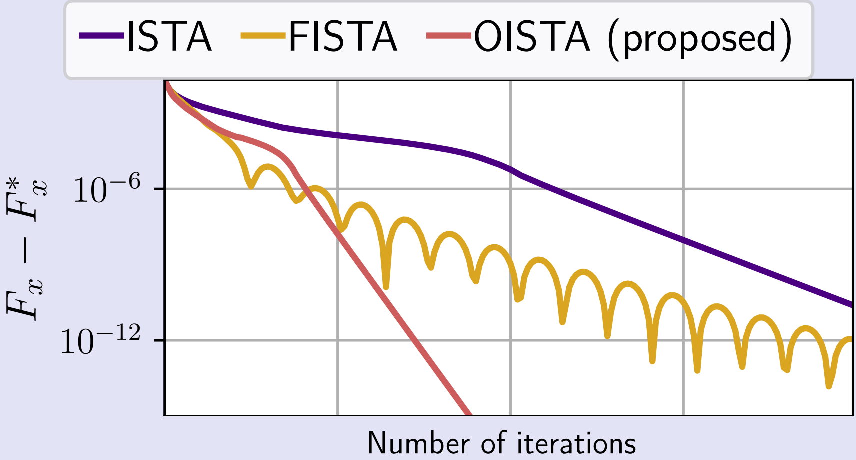
## Oracle ISTA:

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 $1/L_S$

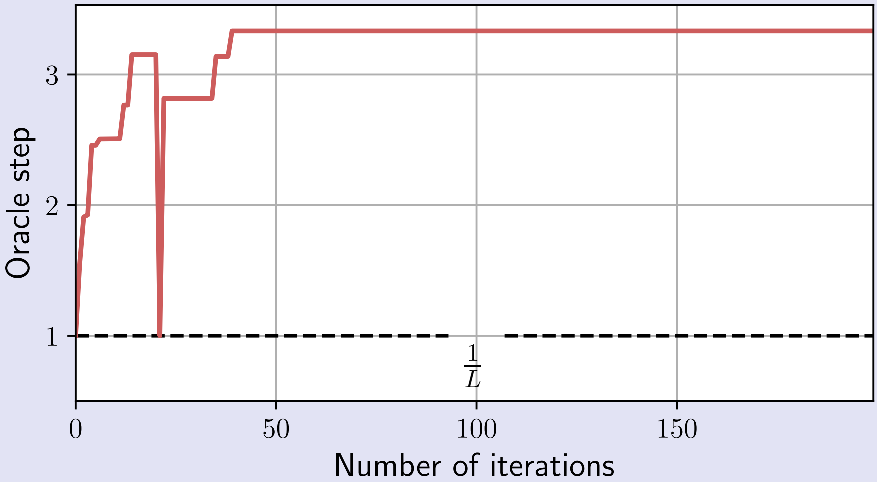
$$y^{(t+1)} = \mathrm{ST}\left(z^{(t)} - \frac{1}{L_S}D^\top(Dz^{(t)} - x), \frac{\lambda}{L_S}\right)$$

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 – 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 $t > T^*$ ,

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

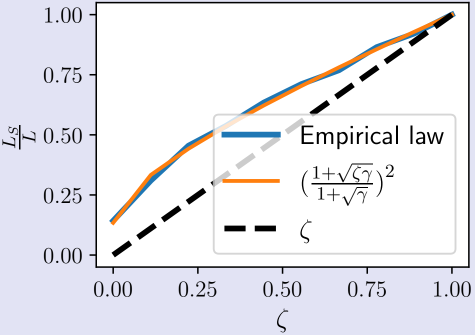
If moreover, $\lambda_{\min}(D_{S^*}^\top D_{S^*}) = \mu^* > 0$ , then

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

## 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

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



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

# Using deep learning to approximate OISTA

# Deep learning for inverse problem

For a direct operator $D$, the inverse problem 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?

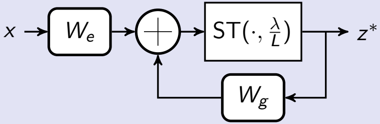
- **ISTA**: worst case algorithm, second order information is $L$.
- **OISTA**: adaptive algorithm, second order information is $L_S$ (NP-hard).
- **LISTA**: adaptive algorithm, use DL to learn second order information?

# Learned ISTA

[Gregor and Le Cun 2010]

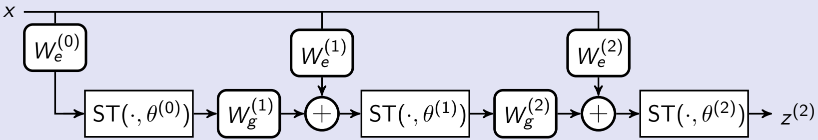
Recurrence relation of ISTA define a RNN

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




With $W_e = \frac{D^\mathsf{T}}{L}$ and $W_g = I - \frac{D^\top D}{L}$, this network is equivalent to ISTA.

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




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

**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]

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

Can be seen as learning

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

# 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]

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

Can be seen as learning

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

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

**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.

# LISTA – Training

**Training :** Given a distribution $p$ in the input space $\mathbb{R}^n$, the training solves

$$\tilde{\Theta}^{(T)} \in \arg\min_{\Theta^{(T)}} \mathbb{E}_{x \sim p}[\mathcal{L}_x(\Phi_{\Theta^{(T)}}(x))] \ .$$

for a given 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 directly.

**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 directly.

# LISTA – Training

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

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

Solving the inverse problem directly.

**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 directly.

# Interlude – regularization $\lambda$

Importance of the parameter $\lambda$

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

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

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

### Maximal value

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

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

### Equiregularization set

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

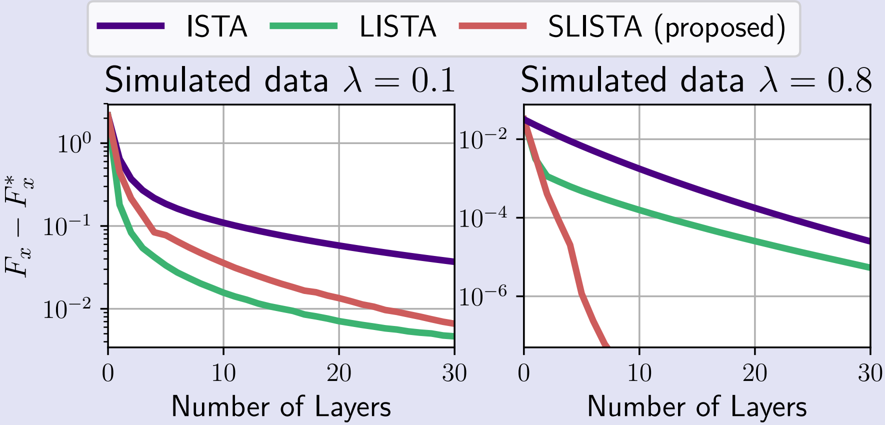
$$\mathcal{B}_\infty = \{x \in \mathbb{R}^n \; ; \quad \|D^\top x\|_\infty = 1\}$$

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

## Performances

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

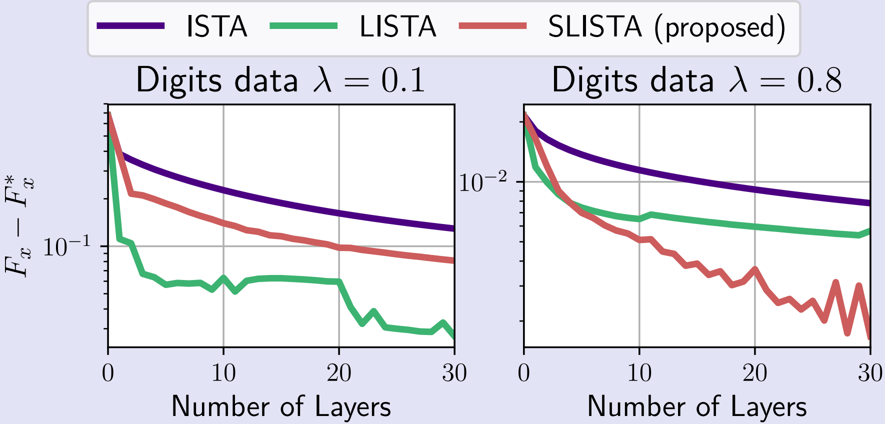
$$D_k \sim \mathcal{U}(\mathcal{S}^{n-1}) \text{ and } x = \frac{\widetilde{x}}{\|D^\top \widetilde{x}\|_\infty} \text{ with } \widetilde{x}_i \sim \mathcal{N}(0, 1)$$

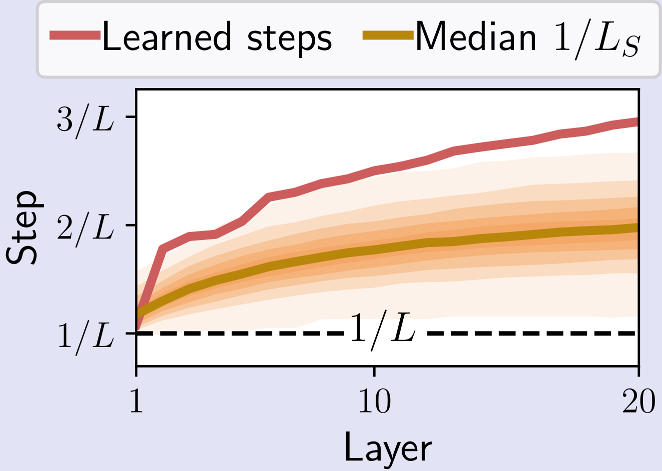


# Performance on semi-real datasets

**Digits:** $8 \times 8$ images [Pedregosa et al. 2011]

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






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

# Theoretical results

Hold on for 2 slides!

# Weights coupling

We denote $\theta = (W, \alpha, \beta)$ the parameters of a given layer $\phi_\theta$.

$$\phi_\theta(z, x) = \mathrm{ST}\left(z - \alpha D^\top(Dz - x), \lambda\alpha\right)$$

Assumption 1:
$D \in \mathbb{R}^{n\times m}$ is a dictionary with non-duplicated unit-normed columns.

**Lemma 4.3 – Weight coupling**

If for all the couples $(z^*(x), x) \in \mathbb{R}^m \times \mathcal{B}_\infty$ such that $z^*(x) \in \operatorname{argmin} F_x(z)$, it holds $\phi_\theta(z^*(x), x) = z^*(x)$. Then, $\frac{\alpha}{\beta} W = D$ .

The solution of the Lasso is a fixed point of a given layer $\phi_\theta$ if and only if $\phi_\theta$ is equivalent to a step of ISTA with a given step-size.

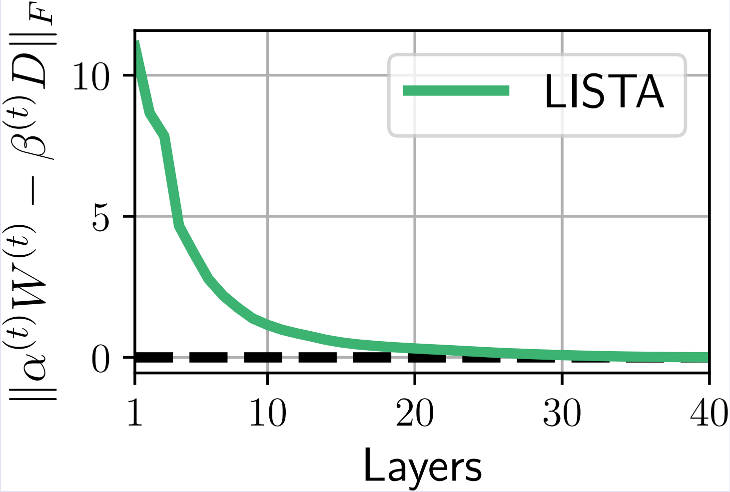
**Theorem 4.4 – Asymptotic convergence**

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

## Numerical verification




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

**The weights $W^{(t)}$ align with $D$ and $\alpha, \beta$ get coupled.**

# Conclusion

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

Take home message:

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

(except for step-sizes!)

Future work:

- Finding a good starting point (first layer)?
- Adversarial cases?

# Thanks!

Code available online:

**adopty** : github.com/tommoral/adopty

Slides are on my web page:

tommoral.github.io @tomamoral