

# Understanding Trainable Sparse Coding with Matrix Factorization

Thomas Moreau CMLA - ENS Paris-Saclay

---

Work in collaboration with Joan Bruna

école  
normale  
supérieure  
paris-saclay

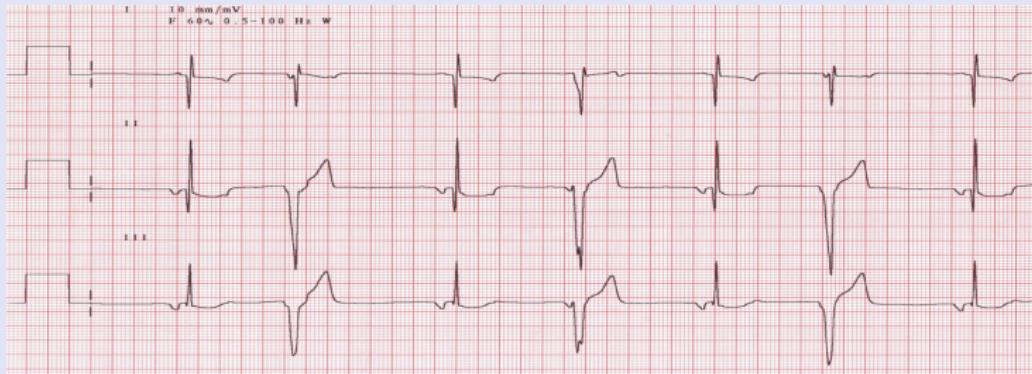


- 1 Physiological signals
- 2 End-to-end Approaches for Time Series
- 3 Post-training for Deep Learning
- 4 Adaptive Iterative Soft Thresholding
- 5 Numerical Experiments

- 1 Physiological signals
- 2 End-to-end Approaches for Time Series
- 3 Post-training for Deep Learning
- 4 Adaptive Iterative Soft Thresholding
- 5 Numerical Experiments

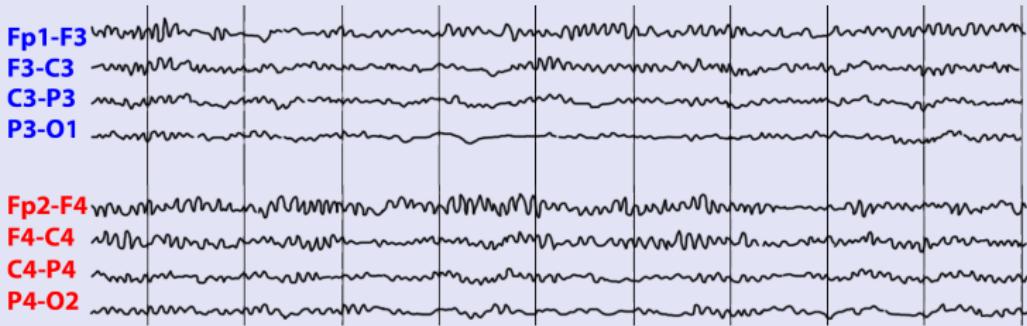
# Physiological signals

## ECG

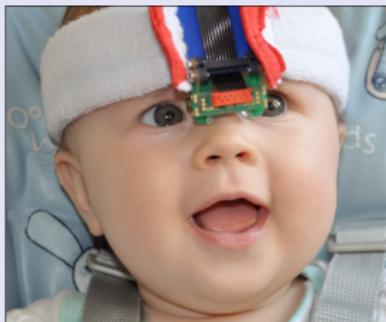
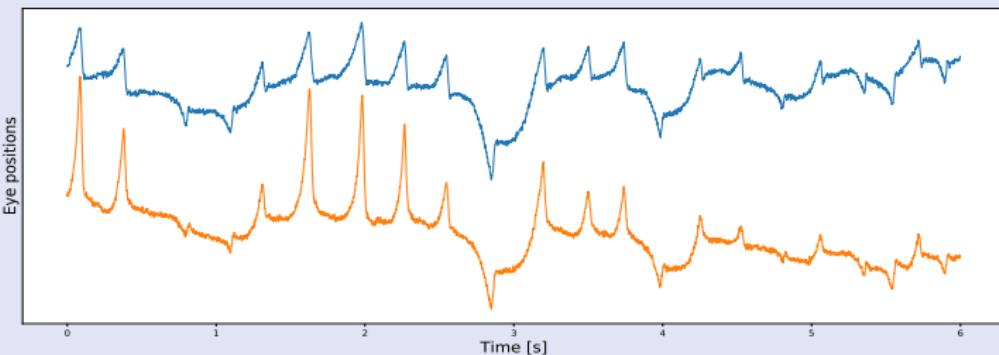


# Physiological signals

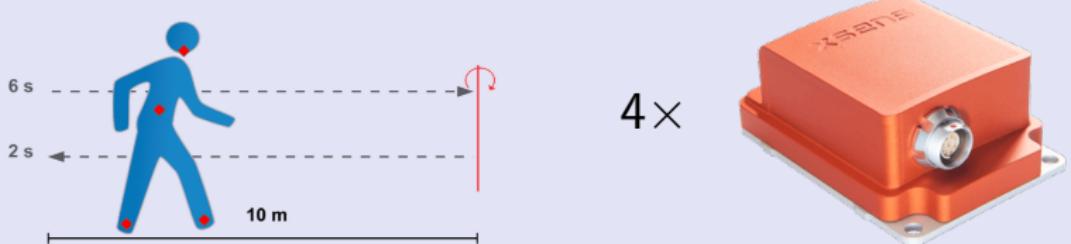
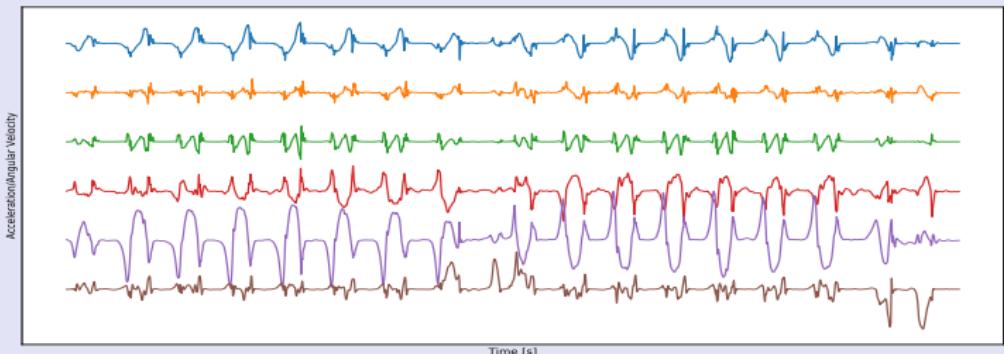
## EEG



## Oculometric signals



## Accelerometers



- ▶ Failure of the vectorial distances
  - ▶ Alignment issues, different lengths  
(can be solved with DTW)
  - ▶ "Curse of dimensionality"
- ▶ Different approaches which can be classified in 2 categories:
  - ▶ Model based methods:  
feature extraction + vectorial method, ...
  - ▶ Data driven methods  
End-to-end model, Neural networks, ...

- ▶ Failure of the vectorial distances
  - ▶ Alignment issues, different lengths  
(can be solved with DTW)
  - ▶ "Curse of dimensionality"
- ▶ Different approaches which can be classified in 2 categories:
  - ▶ Model based methods:  
feature extraction + vectorial method, ...
  - ▶ Data driven methods  
End-to-end model, Neural networks, ...

- ① Physiological signals
- ② End-to-end Approaches for Time Series
- ③ Post-training for Deep Learning
- ④ Adaptive Iterative Soft Thresholding
- ⑤ Numerical Experiments

### Neural Networks:

- ▶ Raw signal as input,  
No feature-engineering
- ▶ Internally select the data representation,  
Adaptive
- ▶ Representation adapted to the task,  
Performant
- ▶ Simple training algorithms,  
Scalable

### Split between risk error 3 terms:

[Bottou and Bousquet, 2008]

- ▶ Approximation error: Universal approximation,  
[Hornik, 1991]
- ▶ Estimation error: Generalization bound,  
[Kawaguchi et al., 2017]
- ▶ Optimization error: Learning convexification,  
[Haeffele and Vidal, 2017]

## Main drawback:

Lack of interpretability. It is often seen as a black box.

How can we bring interpretability in the internal representation?

### Task-driven Dictionary Learning Networks:

[Mairal et al., 2012]

- ▶ Raw signal as input,  
No feature-engineering
- ▶ Representation adapted to the task,  
Performant
- ▶ Complex training algorithms,  
Scalable
- ▶ Highlight local structures,  
Interpretable

**Can we study the links between  
these two models to bring more  
interpretability in neural networks?**

- ① Physiological signals
- ② End-to-end Approaches for Time Series
- ③ Post-training for Deep Learning
- ④ Adaptive Iterative Soft Thresholding
- ⑤ Numerical Experiments

# Post-training for Deep Learning

Paper with J. Audiffren: arxiv:1611.04499

Use the idea to split the representation learning and the task resolution:

- ▶ Post-training step: only train the last layer,
- ▶ Easy problem: this problem is often convex,
- ▶ Link with kernel: close form solution for optimal last layer.
- ▶ Experiments: consistent performance boost with multiple architecture.

- ① Physiological signals
- ② End-to-end Approaches for Time Series
- ③ Post-training for Deep Learning
- ④ Adaptive Iterative Soft Thresholding
- ⑤ Numerical Experiments

Accelerate the LASSO resolution using a neural network.

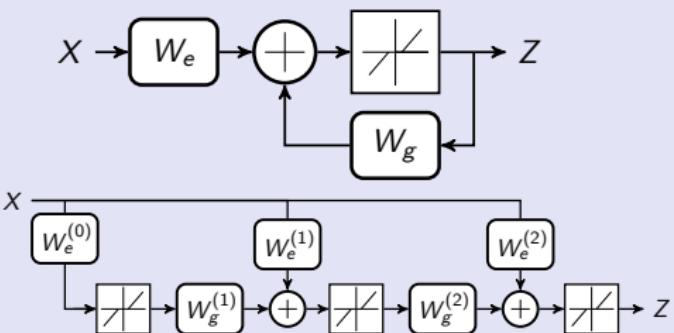
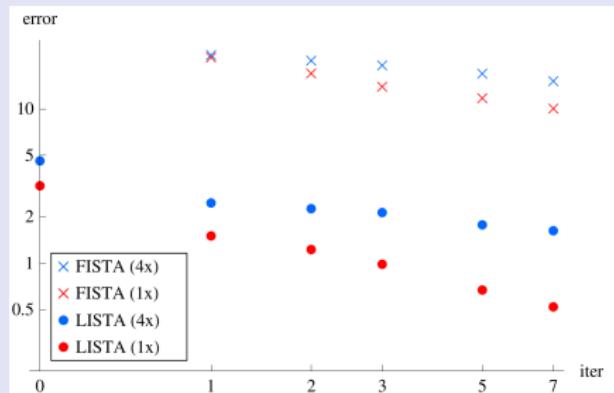


Figure: Adapted from [Gregor and Lecun, 2010]

Links dictionary learning model and sparse representation.

Why does it work?

One core block of today large scale ML is sparsity and particularly, LASSO. Want to solve the problem:

$$\operatorname{argmin}_z F(z) := \underbrace{\|x - Dz\|_2^2}_{E(z)} + \lambda \|z\|_1 , \quad (1)$$

where  $x \in \mathbb{R}^P$ ,  $D \in \mathbb{R}^{P \times K}$  and  $z \in \mathbb{R}^K$ .

(1) can be rewritten as a proximal problem:

$$\operatorname{argmin}_z \underbrace{(y - z)^T B(y - z)}_{E(z)} + \lambda \|z\|_1 \quad (= F(z))$$

where  $B = D^T D$  is the Gram matrix of  $D$  and  $y = D^\dagger x$ .

Surrogate function  $F_q$  associated with point  $z^{(q)}$  :

$$F_q(z) = E(z^{(q)}) + \langle B(z^{(q)} - y), z - z^{(q)} \rangle + \frac{\|B\|_2}{2} \|z - z^{(q)}\|_2^2 + \lambda \|z\|_1 ,$$

## Properties

This surrogate function satisfies

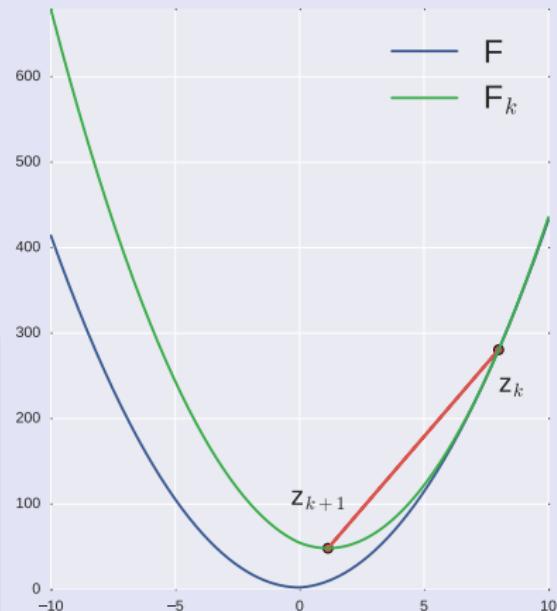
- ①  $F_q(z^{(q)}) = F(z^{(q)})$
- ② for all  $z$ ,  $F_q(z) \geq F(z)$ ,
- ③ solving  $\operatorname{argmin}_z F_q(z)$  is computationally efficient.

Iterative procedure: proximal splitting

$$\begin{aligned} z^{(q+1)} &= \underset{z}{\operatorname{argmin}} F_q(z) \\ &= \operatorname{prox}_{\lambda \|\cdot\|_1} \left( z^{(q)} - \frac{1}{L} \nabla E(z^{(q)}) \right) \end{aligned} \quad (2)$$

### Properties

- ①  $z^*$  is a fix point of (2),
- ② Efficient computation for  $z^{(q+1)}$  as the problem is separable,
- ③ Convergence in  $\mathcal{O}\left(\frac{1}{q}\right)$  in general.



# Why does it work?

---

## ► Guaranteed descent

The construction of the next point guarantees the cost function is decreasing:

$$F(z^{(q+1)}) \leq F_q(z^{(q+1)}) \leq F_q(z^{(q)}) = F(z^{(q)})$$

## ► Efficient computation:

With the isotropic quadratic form  $\frac{L}{2}\mathbf{I}_K$ , the function  $F_q$  is separable.  
The computation are linear in  $K$ .

## Toward an adaptive procedure

We define  $Q_S(u, v) = \frac{1}{2}(u - v)^T S(u - v) + \lambda \|u\|_1$ .

### ISTA:

$$\begin{aligned} F_q(z) &= E(z^{(q)}) + \langle B(z^{(q)} - y), z - z^{(q)} \rangle + Q_{L\mathcal{I}_K}(z, z^{(q)}) , \\ &\rightarrow \min_z Q_{L\mathcal{I}_K}(z, z^{(q)} - \frac{1}{L}B(z^{(q)} - y)) \end{aligned}$$

$\Rightarrow$  Replace  $B$  with an upperbound  $L\mathcal{I}_K$

FacNet: For any matrix  $S$  diagonal, and  $A$  unitary we define :

$$\begin{aligned} \tilde{F}_q(z) &= E(z^{(q)}) + \langle B(z^{(q)} - y), z - z^{(q)} \rangle + Q_S(Az, Az^{(q)}) , \\ &\rightarrow \min_z Q_S(Az, Az^{(q)} - S^{-1}AB(z^{(q)} - y)) \end{aligned}$$

$\Rightarrow$  Replace  $B$  with an approximation  $A^T SA$

Can we choose  $A, S$  to accelerate the optimization compared to ISTA?

## Toward an adaptive procedure

Similar iterative procedure with steps adapted to the problem topology.

$$\widetilde{F}_q(z) = F(z) + (z - z^{(q)})^T R(z - z^{(q)}) + \delta_A(z)$$

Tradeoff between:

- ▶ Rotation to align the norm  $\|\cdot\|_B$  and the norm  $\|\cdot\|_1$  , Computation

$$R = A^T S A - B$$

- ▶ Deformation of the  $\ell_1$ -norm with the rotation  $A$  . Accuracy

$$\delta_A(z) = \lambda \left( \|Az\|_1 - \|z\|_1 \right)$$

## Proposition

Suppose that  $R = A^T S A - B \succ 0$  is positive definite, and define

$$z^{(q+1)} = \arg \min_z \widetilde{F}_q(z) ,$$

Then

$$F(z^{(q+1)}) - F(z^*) \leq \frac{1}{2}(z^{(q)} - z^*)^T R(z^{(q)} - z^*) + \delta_A(z^*) - \delta_A(z^{(q+1)}) .$$

We are interested in factorization  $(A, S)$  for which  $\|R\|_2$  and  $\delta_A$  are small.

# Adaptive Iterative Soft thresholding - Convergence rate

## Theorem

Let  $A_q, S_q$  be the pair of unitary and diagonal matrices corresponding to iteration  $q$ , chosen such that  $R_q = A_q^T S_q A_q - B \succ 0$ . It results that

$$F(z^{(q)}) - F(z^*) \leq \frac{(z^* - z^{(0)})^T R_0 (z^* - z^{(0)}) + 2L_{A_0}(z^{(1)})\|(z^* - z^{(1)})\|_2}{2q} + \frac{\alpha_q - \beta_q}{2q},$$

$$\alpha_q = \sum_{i=1}^{q-1} \left( 2L_{A_i}(z^{(i+1)})\|(z^* - z^{(i+1)})\| + (z^* - z^{(i)})^T (R_{i-1} - R_i)(z^* - z^{(i)}) \right),$$

$$\beta_q = \sum_{i=0}^{q-1} (i+1) \left( (z^{(i+1)} - z^{(i)})^T R_i (z^{(i+1)} - z^{(i)}) + 2\delta_{A_i}(z^{(i+1)}) - 2\delta_{A_i}(z^{(i)}) \right),$$

where  $L_A(z)$  denote the local Lipschitz constant of  $\delta_A$  at  $z$ .

- ▶ For  $A_q = \mathbf{I}_K$  and  $S_q = \|B\|_2 \mathbf{I}_K$ , the procedure is equivalent to ISTA, with the same rate of convergence.
- ▶ If  $\|R_0\|_2 + 2 \frac{L_{A_0}(z_1)}{\|z^* - z_0\|_2} \leq \frac{\|B\|_2}{2}$  and  $A_q = \mathbf{I}_K$  and  $S_q = \|B\|_2 \mathbf{I}_K$  for  $k > 0$ , then the procedure get a head start compare to ISTA
- ▶ **Phase transition :**  
The upper bound is improved when  $\|R_q\|_2 + 2 \frac{L_{A_q}(z^{(q+1)})}{\|z^* - z^{(q)}\|_2} \leq \frac{\|B\|_2}{2}$ , it is thus harder to gain as  $\|z^{(q)} - z^*\|_2 \rightarrow 0$

## Generic Dictionaries

A dictionary  $D \in \mathbb{R}^{p \times K}$  is a generic dictionary when its columns  $D_i$  are drawn uniformly over the  $\ell_2$  unit sphere  $\mathcal{S}^{p-1}$ .

### Theorem (Acceleration conditions)

In **expectation over the generic dictionary**  $D$ , the factorization algorithm using a diagonally dominant matrix  $A \subset \mathcal{E}_\delta$ , has better performance for iteration  $q + 1$  than the normal ISTA iteration – which uses the identity – when

$$\lambda \mathbb{E}_z \left[ \|z^{(q+1)}\|_1 + \|z^*\|_1 \right] \leq \sqrt{\frac{K(K-1)}{p}} \underbrace{\mathbb{E}_z \left[ \|z^{(q)} - z^*\|_2^2 \right]}_{\text{expected resolution at iteration } q}$$

## Corollary (Acceleration conditions)

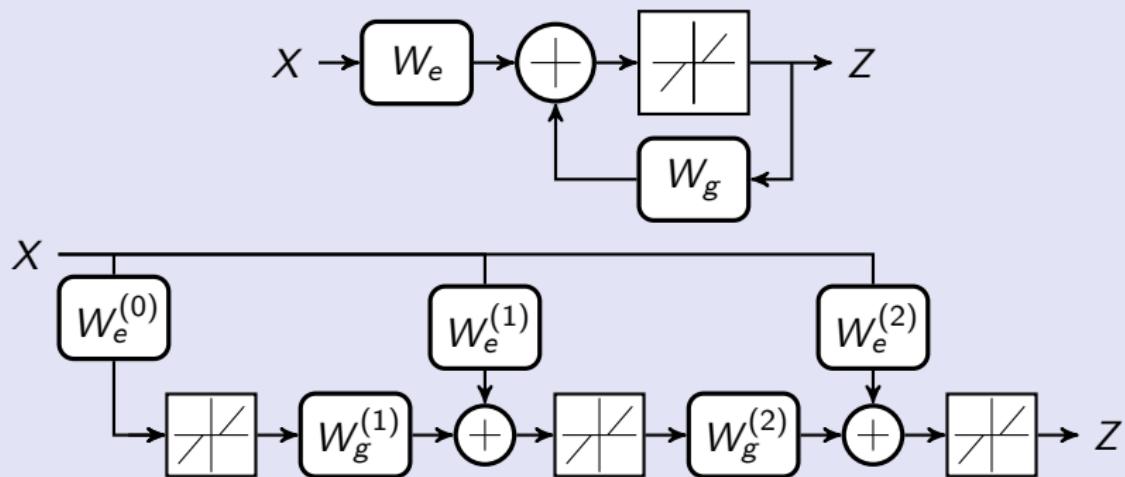
If the input distribution and the regularization parameter  $\lambda$  verify

$$\frac{\lambda\sqrt{p}}{8} \leq \mathbb{E}_z \left[ \|z^*\|_1 \right],$$

Then for any resolution  $\mathbb{E}_z \left[ \|z^{(q)} - z^*\|_2 \right] = \epsilon > 0$  at iteration  $q$ , the performance of our factorization algorithm is better than the performance of ISTA, in expectation over the generic dictionaries.

FacNet can improve the performances compared to ISTA when this is verified.

- ① Physiological signals
- ② End-to-end Approaches for Time Series
- ③ Post-training for Deep Learning
- ④ Adaptive Iterative Soft Thresholding
- ⑤ Numerical Experiments



**Figure:** Network architecture for ISTA/LISTA. LISTA is the unfolded version of the RNN of ISTA, trainable with back-propagation.

If  $W_e = \frac{D^T}{L}$  and  $W_g = I - \frac{B}{L}$ , this network is exactly 2 iterations of ISTA.

## Specialization of LISTA

$$z^{(q+1)} = A^T \underset{S}{\text{prox}}(Az^{(q)} - S^{-1}AB(z^{(q)} - y)) ,$$

with  $A$  unitary and  $S$  diagonal.

Same architecture with more constraints on the parameter space:

$$\begin{cases} W_e &= S^{-1}AD^T \\ W_g &= A^T - S^{-1}ABA^T \end{cases}$$

⇒ LISTA can be at least as good as this model.

# Learned FISTA

The same ideas can also be applied to FISTA to obtain similar procedures:

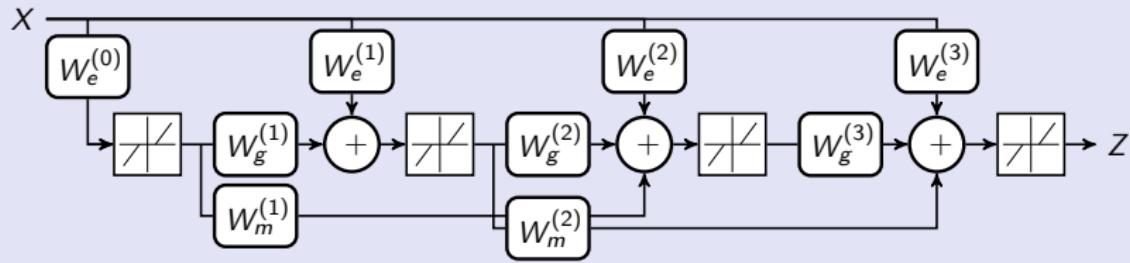


Figure: Network architecture for L-FISTA.

## Generating Model:

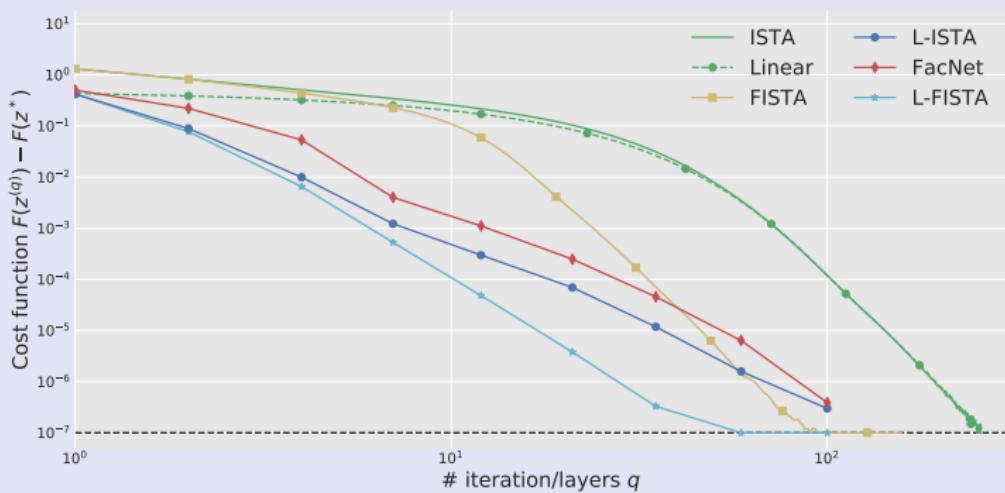
- ▶  $D = \left( \frac{d_1}{\|d_1\|_2}, \dots, \frac{d_K}{\|d_K\|_2} \right)$  with  $d_k \sim \mathcal{N}(0, I_P)$  for all  $k \in \llbracket 1, K \rrbracket$ ,
- ▶  $z = (z_1, \dots, z_K)$  are constructed following a bernouilli gaussian:

$$z_k = b_k a_k, \quad b_k \sim \mathcal{B}(\rho) \text{ and } a \sim \mathcal{N}(0, \sigma I_K)$$

with:  $K = 100$ ,  $P = 64$ , for the dimension,  $\sigma = 10$  and  $\lambda = 0.01$

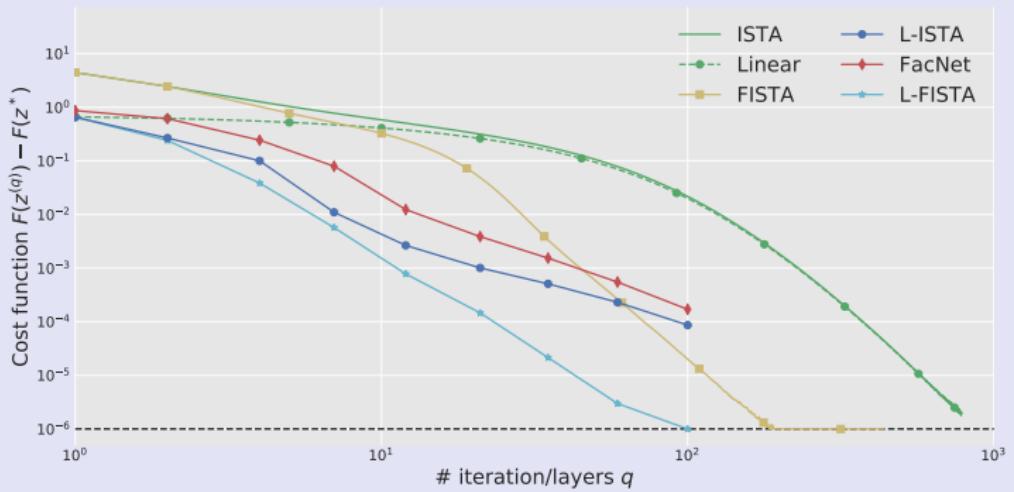
⇒ The sparsity patterns are uniformly distributed.

# Artificial simulation



Evolution of the cost function  $F(z^{(q)}) - F(z^*)$  with the number of layers/iterations  $q$  with a sparse model  $\rho = 1/20$ .

# Artificial simulation



Evolution of the cost function  $F(z^{(q)}) - F(z^*)$  with the number of layers/iterations  $q$  with a denser model  $\rho = \frac{1}{4}$ .

# Adversarial dictionary

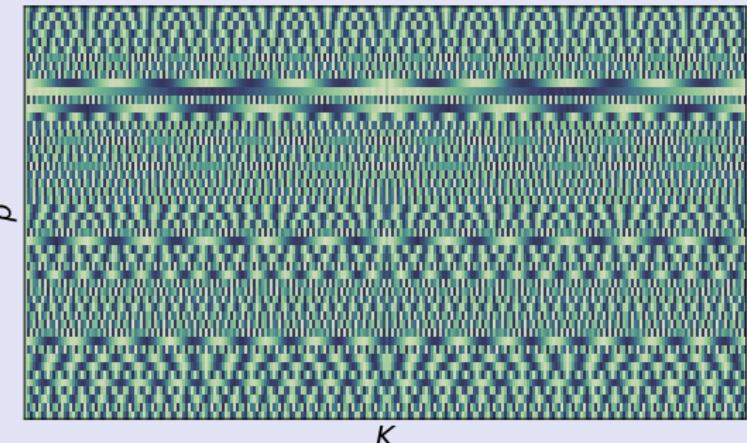
**Adversarial dictionary:**

$$D = [d_1 \dots d_K] \in \mathbb{R}^{K \times p},$$

with

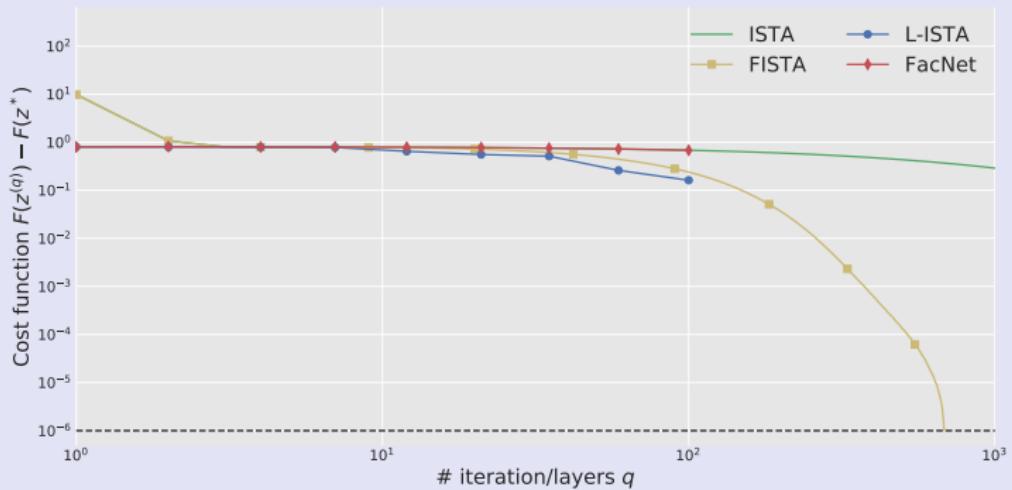
$$d_j = e^{-i \frac{2\pi j \zeta_q}{K}}$$

for a random subset of frequencies  $\{\zeta_i\}_{i \leq m}$



⇒ Eigenvectors of  $D$  are far from canonical basis.

# Adversarial dictionary

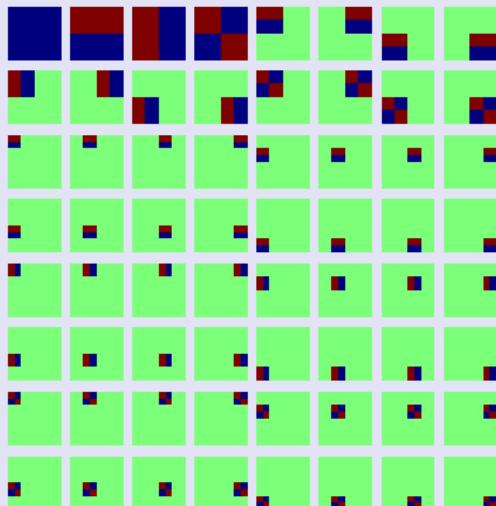


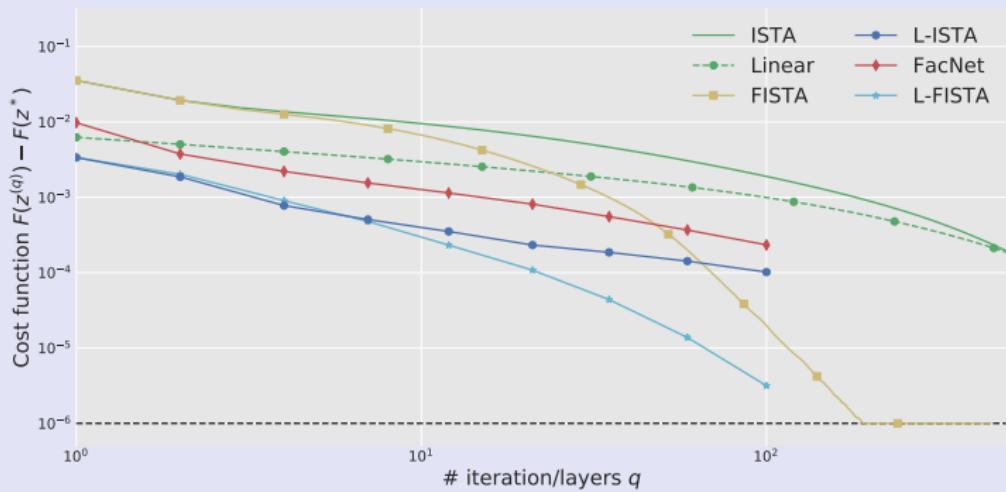
Evolution of the cost function  $F(z^{(q)}) - F(z^*)$  with the number of layers/iterations  $k$  with  $n$  adversarial dictionary.

Sparse coding for the PASCAL 08 datasets over the Haar wavelets family.

The sparse coding is performed for patches of size  $8 \times 8$ .

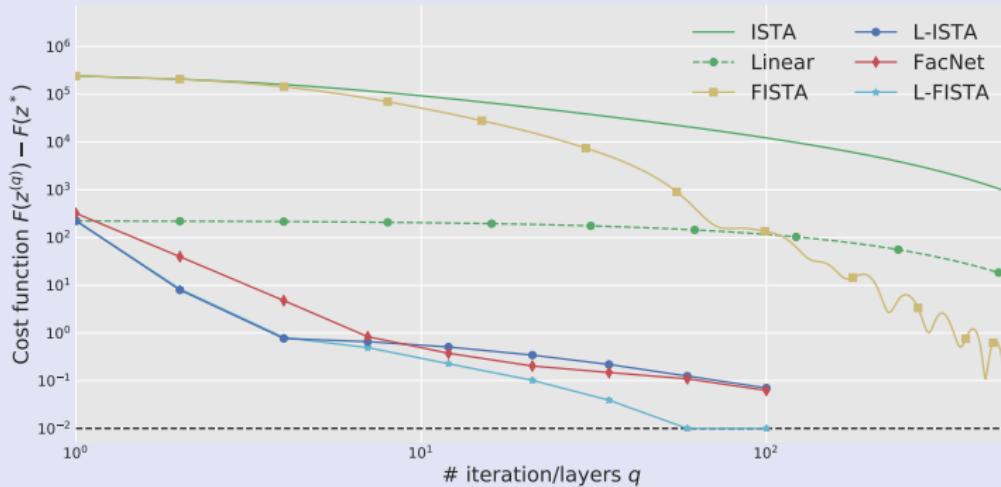
Train over 500 images and test over 100 images.





Evolution of the cost function  $F(z^{(q)}) - F(z^*)$  with the number of layers  
or the number of iteration  $q$  for Pascal VOC 2008.

Dictionary  $D$  with  $K = 100$  atoms learned on 10 000 MNIST samples (17x17) with dictionary learning. LISTA trained with MNIST training set and tested on MNIST test set.



Evolution of the cost function  $F(z^{(q)}) - F(z^*)$  with the number of layers or the number of iteration  $q$  for MNIST.

# Conclusion

---

- ▶ Non asymptotic acceleration is possible :  
Approximate matrix factorization of  $B = D^T D$ 
  - ▶ Nearly diagonalize the kernel,
  - ▶  $\ell_1$ -norm nearly invariant by this orthogonal transformation.
- ▶ Future work:
  - ▶ Improve the factorization formulation:

$$\min_{A^T A = I_K} f(\|DA\|_{1,2}) + \lambda_q \frac{\|A\|_{1,1}}{n},$$

- ▶ Give generic bounds for sub gaussian  $D$ ,
- ▶ Link to Sparse PCA.

# Questions?

Code:  tomMoral/AdaptiveOptim

Paper: <https://arxiv.org/abs/1706.01338>

More at  [tommoral.github.io](https://tommoral.github.io)  tomMoral

# References

---

- Beck, A. and Teboulle, M. (2009). A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. [SIAM Journal on Imaging Sciences](#), 2(1):183–202.
- Bottou, L. and Bousquet, O. (2008). Learning using large datasets. [Mining Massive DataSets for Security](#), 3.
- Daubechies, I., Defrise, M., and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. [Communications on Pure and Applied Mathematics](#), 57(11):1413–1457.
- Gregor, K. and LeCun, Y. (2010). Learning Fast Approximations of Sparse Coding Karol. In [International Conference on Machine Learning \(ICML\)](#), volume 152, pages 399–406, Haifa, Israel.
- Haefele, B. D. and Vidal, R. (2017). Global Optimality in Neural Network Training. In [Conference on Computer Vision and Pattern Recognition \(CVPR\)](#), pages 7331–7339, Honolulu, HI, USA.
- Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. [Neural Networks](#), 4(2):251–257.
- Kawaguchi, K., Pack Kaelbling, L., and Bengio, Y. (2017). Generalization in Deep Learning. [preprintq](#), arXiv:1710(05468).
- Mairal, J., Bach, F., and Ponce, J. (2012). Task-driven dictionary learning. [IEEE Transactions on Pattern Analysis and Machine Intelligence](#), 34(4):791–804.
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. [Journal of the royal statistical society. Series B \(methodological\)](#), 58(1):267—288.