Understanding Trainable Sparse Coding with Matrix Factorization

Thomas Moreau ENS Cachan - CMLA

Work in collaboration with Joan Bruna

Plan

1. Sparse Coding and ISTA
2. Adaptive Iterative Soft Thresholding
3. Numerical Experiments

LASSO [Tibshirani, 1996]

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

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

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

Typically :

- x are the label associated to the data points D and we look for the best sparse linear model. sparse regression
- x is a data point and D is a dictionary and we look for a sparse representation of x on D. sparse coding

The LASSO problem (1) can be rewritten as a proximal problem :

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

where $B = D^\mathsf{T} D$ and $y = D^\dagger x$.

Surrogate function F_k associated with point z_k :

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

Properties

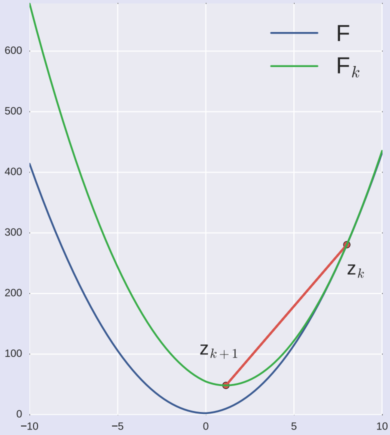
This surrogate function satisfies

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

ISTA (2/2)

Iterative procedure : proximal splitting

$$
\begin{aligned}
z_{k+1} &= \operatorname*{argmin}_z F_k(z) \\
&= \operatorname{prox}_{\lambda\|\cdot\|_1}\left(z_k - \frac{1}{L}\nabla E(z_k)\right)
\end{aligned} \tag{2}
$$



Properties

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

Why does it work ?

- **Guaranteed descent**

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

  $$F(z_{k+1}) \leq F_k(z_{k+1}) \leq F_k(z_k) = F(z_k)$$

- **Efficient computation :**

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

Plan

Solving multiple optimization problems

- Solve more than one instance of a problem :
  - Regularization path (Multiple λ)
  - Multi-objective regression (Multiple labels x)
  - Sparse coding (Multiple data points x)
  - ...
- Regular optimization techniques does not leverage the common structure of these problems :
  - Designed to solve the worst case,
  - Fix updates, without using information from previous resolutions.

Can we use the global structure of the problem to accelerate its resolution ?

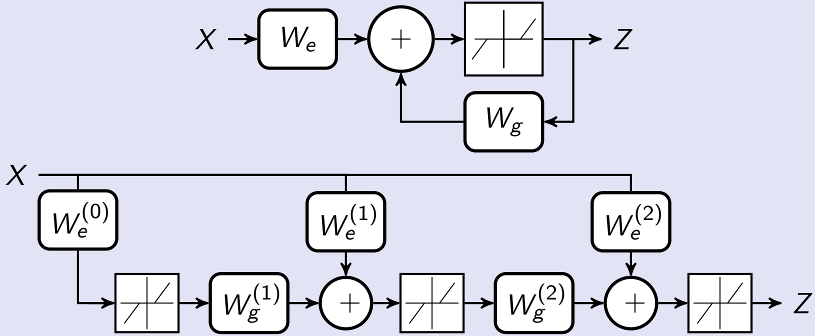


FIGURE – Network architecture for ISTA/LISTA. LISTA is the unfolded version of the RNN of ISTA, trainable with back-propagation.

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

Toward an adaptive procedure

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

$$\begin{aligned} F_k(z) &= E(z_k) + \langle B(z_k - y), z - z_k\rangle + Q_{L\mathbf{I}}(z, z_k) \ , \\ &\to \min_z Q_{L\mathbf{I}}(z, z_k - \frac{1}{L}B(z_k - y)) \end{aligned}$$

$\Longrightarrow$ Replace B with an upperbound $L\mathbf{I}$

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

$$\begin{aligned} \widetilde{F}_k(z) &= E(z_k) + \langle B(z_k - y), z - z_k\rangle + Q_S(Az, Az_k) \ , \\ &\to \min_z Q_S(Az, Az_k - S^{-1}AB(z_k - y)) \end{aligned}$$

$\Longrightarrow$ Replace B with an approximation $A^\mathsf{T} SA$

How can we choose A, S to accelerate the optimization ?

Toward an adaptive procedure

$$\widetilde{F_k}(z) = F(z) + (z - z_k)^{\mathsf{T}} R(z - z_k) + \delta_A(z)$$

Similar iterative procedure with steps adapted to the problem topology.

Tradeoff between :

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

$$R = A^{\mathsf{T}} S A - B$$

- Deformation of the ℓ_1-norm with the rotation A . accuracy

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

One step improvement

Proposition

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

$$z_{k+1} = \arg\min_{z} \widetilde{F_k}(z) \ ,$$

Then

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

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

Theorem

Let A_k, S_k be the pair of unitary and diagonal matrices corresponding to iteration k, chosen such that $R_k = A_k^\mathsf{T} S_k A_k - B \succ 0$. It results that

$$F(z_k) - F(z^*) \leq \frac{(z^* - z_0)^\mathsf{T} R_0 (z^* - z_0) + 2L_{A_0}(z_1)\|(z^* - z_1)\|_2}{2k} + \frac{\alpha_k - \beta_k}{2k} \ , \text{ with} \quad (3)$$

$$\alpha_k = \sum_{i=1}^{k-1} \left(2L_{A_i}(z_{i+1})\|(z^* - z_{i+1})\| + (z^* - z_i)^\mathsf{T}(R_{i-1} - R_i)(z^* - z_i)\right) \ ,$$

$$\beta_k = \sum_{i=0}^{k-1}(i+1)\left((z_{i+1} - z_i)^\mathsf{T} R_i (z_{i+1} - z_i) + 2\delta_{A_i}(z_{i+1}) - 2\delta_{A_i}(z_i)\right) \ ,$$

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

Interpretation

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

Plan

FacNet

Specialization of LISTA

$$z_{k+1} = A^\mathsf{T} \operatorname{prox}_S(Az_k - S^{-1}AB(z_k - y)) \ ,$$

with A unitary and S diagonal.
Same architecture with more constraints on the parameter space :

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

$\Rightarrow$ 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 :

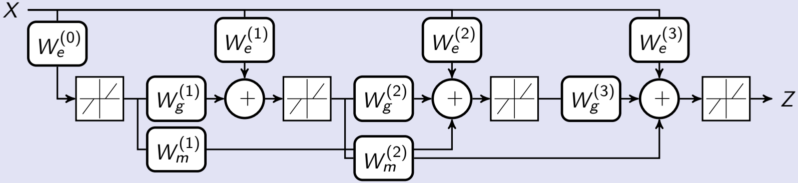


FIGURE – Network architecture for L-FISTA.

Artificial simulation

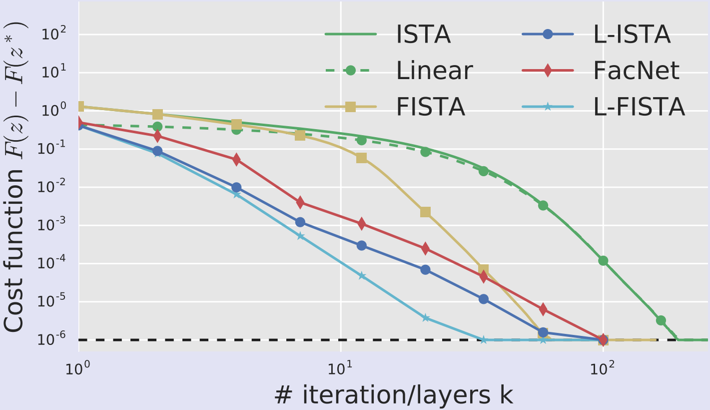
Generating Model :

- $D = \left(\frac{d_1}{\|d_1\|_2}, \dots \frac{d_m}{\|d_m\|_2}\right)$ with $d_i \sim \mathcal{N}(0, \mathbf{I}_n)$ for all $i = 1 \dots m$,
- $z = (z_1, \dots z_m)$ are constructed following a bernouilli gaussian :

$$z_i = b_i a_i, \qquad b_i \sim \mathcal{B}(\rho) \text{ and } a_i \sim \mathcal{N}(0, \sigma\,\mathbf{I}_m)$$

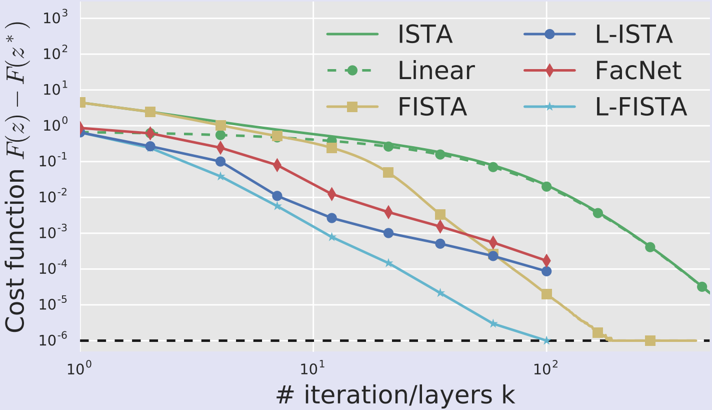
with : $m = 100$, $n = 64$, for the dimension, $\sigma = 10$ and $\lambda = 0.01$

$\Rightarrow$ The sparsity patterns are uniformely distributed.



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

Artificial simulation

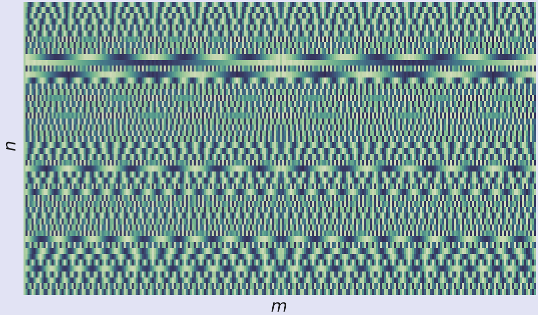


Evolution of the cost function $F(z_k) - F(z^*)$ with the number of layers/iterations k with a denser model $\rho = {}^1/_4$.

Adversarial dictionary :
$D = \left[d_1 \dots d_m\right] \in \mathbb{R}^{m \times n}$,
with

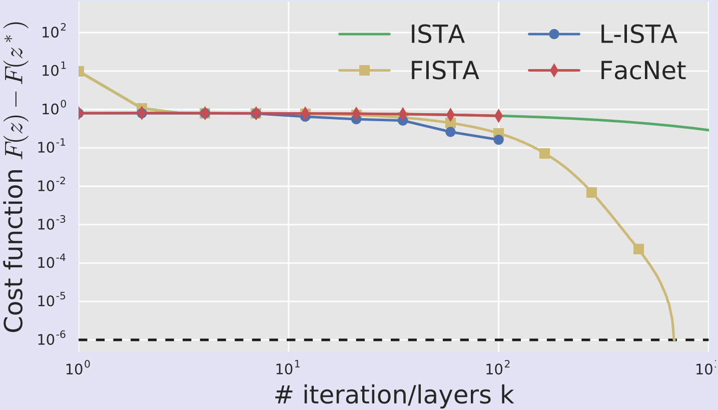
$$d_j = e^{-i\frac{2\pi j \zeta_k}{m}}$$

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



$\Rightarrow$ Eigenvectors of D are far from canonical basis.

Adversarial dictionary



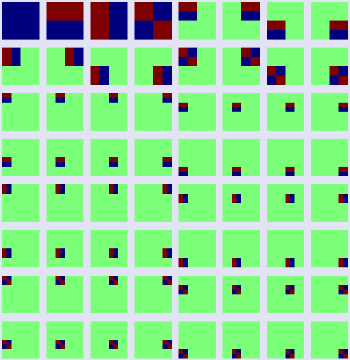
Evolution of the cost function $F(z_k) - F(z^*)$ with the number of layers/iterations k with n adversarial dictionary.

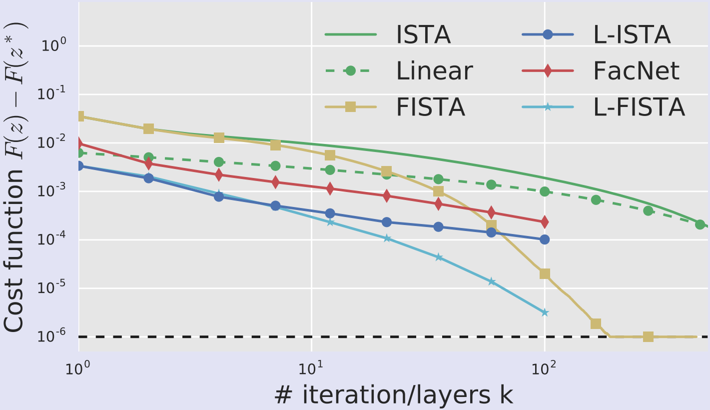
PASCAL 08

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

The sparse coding is performed for patches of size 8×8.

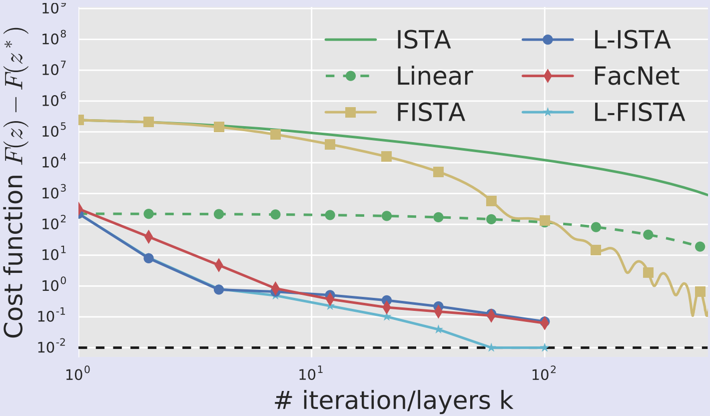
Train over 500 images and test over 100 images.





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

MNIST



Evolution of the cost function $F(z_k) - F(z^*)$ with the number of layers or the number of iteration k for MNIST with $m = 100$ *(dashed lines)* and $m = 289$ *(solid line)*.

Conclusion

- Non asymptotic acceleration is possible :
  Approximate matrix factorization of $B = D^\intercal D$
  - Nearly diagonalize the kernel,
  - ℓ_1-norm nearly invariant by this orthogonal transformation.

- *Future work :*
  - Improve the factorization formulation :

$$\min_{A^\intercal A = \mathbf{I}} f(\|DA\|_{1,2}) + \lambda_k \frac{\|A\|_{1,1}}{n} \ ,$$

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

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.

Gregor, K. and Le Cun, Y. (2010). Learning Fast Approximations of Sparse Coding. In *International Conference on Machine Learning (ICML)*, pages 399–406.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the royal statistical society. Series B (methodological)*, 58(1) :267–288.