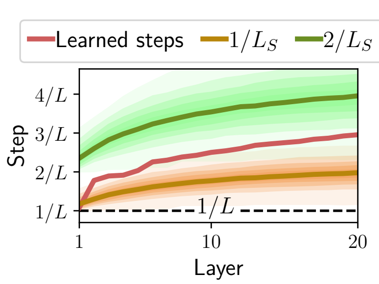
We thank the reviewers for their careful reading and their constructive comments. We clarify the main points raised in the reviews:

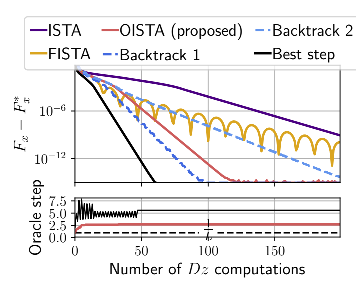
Rev#1 **1.1 More experiments**: With suggestion from Rev.#2, we propose to replace the randomly generated x in Figure.6(c/d) by images from the digits dataset. The curves do not change significantly but this use-case is close to a real application. Also, such method could be used to accelerate the resolution of inverse problems (*e.g.* see Adler et al 2017). We will emphasize such application in our introduction.

Rev#2 **2.1 Comparison with backtracking (BT)**: BT typically considers (e.g. Nesterov 2013) candidate steps of the form $\alpha_0\beta^k, k \geq 0$, where α_0 is an initial guess and $\beta < 1$ is a shrinking constant (Armijo line-search). It is indeed close to our work with two main differences: i) At test time, BT should be performed at each step of ISTA, making the cost of one iteration larger than ISTA/SLISTA. Also note that the hyperparameters α_0, β are critical and hard to tune properly. ii) BT finds greedily a suitable step-size at *a given iteration* for *a given sample*. The goal with SLISTA is to learn a sequence of step-sizes for *all iterations* jointly for the *input distribution*. We will clarify this relationship and core differences in the text and replace Fig.2 with Fig A.2, where we show the results of a perfect line-search algorithm (taking at each iteration a step-size exactly minimizing the loss function), and two backtracking with different parameters (α_0, β). One could also use BT to learn steps for the whole distribution, by picking step sizes in the set $(\alpha_0\beta^k)_{k\geq 0}$. This corresponds to a discretized version of the SLISTA problem, which we feel is out of the scope of the current article.
2.2 Results relative to [2]: We stress that Theorem 4.4 states that LISTA only learns step-sizes *asymptotically*. Thus, LISTA can still leverage the dictionary structure in the first layers and improve compared to SLISTA, as seen in Fig.6(a/c). Our results are consistent with those of [2], as their section 2.3.1 shows a phase transition where the structure of D leveraged by FacNet only improves the convergence *in the first layers*.
2.3 Clarify proposition 4.1: The proof will be rewritten as the part about uniform convergence of ISTA is indeed confusing. ISTA's rate of convergence writes: $F_x(z_T(x)) - F_x^* \leq \frac{L}{T}\|z^*(x)\|_2^2$. Further, $\|z^*\|_2 \leq \|z^*\|_1 \leq \frac{1}{\lambda}F_x^* \leq \frac{1}{\lambda}F_x(0) = \frac{1}{\lambda}\|x\|_2^2$. Since $\mathcal{B}_\infty$ is bounded, $\|z^*(x)\|_2$ is uniformly bounded. It yields an inequality of the form $F_x^* \leq F_x(z_T(x)) \leq F_x^* + K/T$ where K is independent of x. Taking expectations shows as advertised: $\mathbb{E}_{x\sim p}[F_x(\Phi_{\Theta^{(T)}_{\text{ISTA}}}(x))] \xrightarrow[T\to\infty]{} \mathbb{E}_{x\sim p}[F_x^*]$. This proposition shows the global (*theoretical*) minimizer of the unsupervised problem solves the LASSO. In practice, as any neural network, the computed solution faces optimization (non-convexity) and generalization (empirical risk minimization) errors.
2.4 Fig.4: We modify Fig.4 by adding the $2/L_S$ line (see Fig A.1). Learned steps are mainly included in $]0, 2/L_S[$ which guarantees the cost function decrease, if the support inclusion condition is verified. However, steps above $2/L_S$ may lead to greater decrease of the loss function as seen in Fig A.2.
2.5 Semi-real experiment: see 1.1.

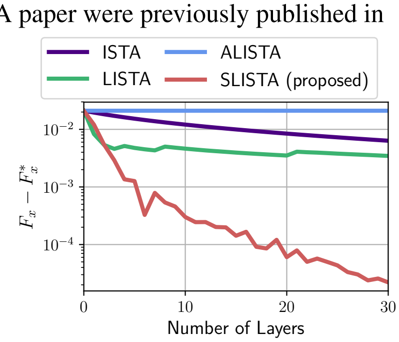
Rev#3 **3.1 Effect of different step-sizes**: Fig.2 shows that larger step-sizes can lead to faster convergence. See 2.1.
3.2 Is SLISTA better? ALISTA does not work In the experiments, we see that SLISTA is better when the iterates are very sparse (high λ), leveraging the same properties as OISTA. It is outperformed by LISTA for small λ. As stated in the text (l.229/252), following thm.4.4., ALISTA cannot converge in this unsupervised setting.
3.3 Unsupervised/Supervised: ISTA converges to a solution of the Lasso. In the supervised setting, a unique solution exists independently from the Lasso solution. In most practical cases, the Lasso solution and the supervised solution are different. Even if they match, it is for a specific λ, unkown *a piori*. Hence ISTA does not converge for the supervised problem. This is typically highlighted in Figure.1(a) from the ALISTA article (Liu et al 2019) where the MSE of ISTA plateaus. We will clarify this in the text.
3.4 Extension to convolutional cases/robustness: This is surely an interesting extension, which cannot unfortunately be inserted in the paper for lack of space. Also, the theoretical results from ALISTA paper were previously published in NeurIPS without the convolution/robustness part (Chen et al 2018).



(A.1) Step-sizes learned with SLISTA



(A.2) Performance of F/ISTA, OISTA and ISTA with Backtracking / oracle line search.



(A.3) Performances of SLISTA on digits with $\lambda = 0.8$.