ALISTA: Analytic Weights Are As Good As Learned Weights in LISTA

Jialin Liu* (UCLA Math) Xiaohan Chen* (Texas A&M CSE) Zhangyang Wang (Texas A&M CSE) Wotao Yin (UCLA Math)

Highlights

We propose a new learning algorithm for learned ISTA, that:
- reduces **MILLIONS** of parameters to only **32** scalars;
- shortens the training time from **1.5 HOURS** to **6 MINUTES**;
- achieves a provably **optimal linear** convergence rate.

Introduction to Learned ISTA (LISTA)

Problem: Recover a sparse vector $\mathbf{x}^*$ from its noisy measurements by $\mathbf{D} \in \mathbb{R}^{N\times M}$:

$$\mathbf{b} = \mathbf{D}\mathbf{x}^* + \varepsilon,$$

LASSO:

$$\underset{\mathbf{x}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{b}-\mathbf{D}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

Iterative shrinkage thresholding algorithm (ISTA)

$$\mathbf{x}^{k+1} = \eta_{\lambda/L}\Big(\mathbf{x}^k + \frac{1}{L}\mathbf{D}^T(\mathbf{b}-\mathbf{D}\mathbf{x}^k)\Big), \quad k=0,1,2,\dots \qquad \textbf{(ISTA)}$$

where η_θ is soft-thresholding, λ and L are selected by hand or cross-validation. ISTA converges **sublinearly** and **eventually-linearly** to a **LASSO solution**, **not** $\mathbf{x}^*$.

LISTA: unrolls ISTA with K total iterations to a neural network, replace $\lambda/L, \mathbf{D}, \mathbf{D}^T$ by free matrices (known as Learned ISTA or LISTA [1]):

$$\mathbf{x}^{k+1} = \eta_{\theta_k}(\mathbf{W}_1^k\mathbf{b} + \mathbf{W}_2^k\mathbf{x}^k), \quad k=0,1,\cdots,K-1, \qquad \textbf{(LISTA)}$$

Inputs are $\mathbf{x}^0$ and $\mathbf{b}$. Output $\mathbf{x}^K$ is our recovery.

Training (deciding θ_k, $\mathbf{W}_1^k$, $\mathbf{W}_2^k$) For a fixed $\mathbf{D}$ and almost all $(\mathbf{b},\mathbf{x}^*)$ following a certain distribution, obtain parameters $\Theta^K = \{(\mathbf{W}_1^k,\mathbf{W}_2^k,\theta_k)\}_{k=0}^{K-1}$ such that $\mathbf{x}^K$ approximates $\mathbf{x}^*$ (the ground truth).
In another word, given the distributions of $\mathbf{b}$ and $\mathbf{x}^*$, we

$$\underset{\Theta^K}{\text{minimize}}\ \frac{1}{2}\mathbb{E}_{\mathbf{b},\mathbf{x}^*}\|\mathbf{x}^K(\Theta^K,\mathbf{b},\mathbf{x}^0)-\mathbf{x}^*\|_2^2.$$

Stochastic gradient descent (SGD) can be applied to solve this minimization problem. The gradient w.r.t. $\mathbf{x}^K$ on Θ^K are obtained with the chain rule.

LISTA-CP: In [2], parameters are reduced by coupling: $\mathbf{W}_2^k = \mathbf{I} - \mathbf{W}_1^k\mathbf{D}$. With $\mathbf{W}^k := (\mathbf{W}_1^k)^T$, the formulation of LISTA-CP is

$$\mathbf{x}^{k+1} = \eta_{\theta_k}(\mathbf{x}^k - (\mathbf{W}^k)^T(\mathbf{D}\mathbf{x}^k - \mathbf{b})). \qquad \textbf{(LISTA-CP)}$$

Issues: With $O(KNM)$ trainable parameters, training takes 1.5 hours on a GTX 1080Ti. Can we reduce the number of trainable parameters and training time?

Our Contributions

- We show that the layer-wise weights $\mathbf{W}^k$ in (**LISTA-CP**) can be given by a data-free optimization problem. Our method is called ALISTA with the A for "analytic."
- The new scheme preserves the linear convergence proved in [2]. In addition, we develop a recovery error lower bound that shows the linear convergence rate is optimal w.r.t. the order of convergence.
- We design a robust ALISTA model that is robust to noises in the dictionary $\mathbf{D}$.
- We extend our algorithms and theories to the case where $\mathbf{D}$ is a convolutional operator. (See our paper [3] for details.)

OpenReview Github

From LISTA to ALISTA

In compressive sensing, a dictionary $\mathbf{D}$ with smaller **mutual coherence** leads to the better recovery performance. Similarly, good weights $\mathbf{W}^k$ in (**LISTA-CP**) satisfy the following condition up to a scalar.

Coherence Minimization: Given $\mathbf{D} \in \mathbb{R}^{N\times M}$ (columns normalized), we take

$$\tilde{\mathbf{W}} \in \underset{\substack{\mathbf{W}\in\mathbb{R}^{N\times M} \\ (\mathbf{W}_{:,i})^T\mathbf{D}_{:,i}=1, 1\le i\le M}}{\arg\min} \Big\{ \max_{\substack{i\neq j \\ 1\le i,j\le M}} (\mathbf{W}_{:,i})^T\mathbf{D}_{:,j} \Big\}. \qquad (1)$$

We have proved that the optimization problem in (1) is feasible and attainable.

Theorem 1 (Recovery error upper bound) *Suppose $\varepsilon = 0$ and let $\{\mathbf{x}^k\}_{k=1}^\infty$ be generated by (**LISTA-CP**). There exists a sequence of parameters $\{\gamma_k,\theta_k\}_k$ such that, with*

$$\mathbf{W}^k = \gamma_k\tilde{\mathbf{W}}, \quad \tilde{\mathbf{W}} \text{ calculated by (1)}, \qquad (2)$$

we have the following error bound:

$$\|\mathbf{x}^k(\Theta^k,\mathbf{b},\mathbf{x}^0)-\mathbf{x}^*\|_2 \le C\exp(-ck) \qquad (3)$$

uniformly for all $\mathbf{x}^*$ satisfying some assumptions (see [3]), where $c, C > 0$ are constants that depend only on $\mathbf{D}$ and the distribution of $\mathbf{x}^*$.

Theorem 2 (Recovery error lower bound) *Suppose $\varepsilon = 0$ and let $\{\mathbf{x}^k\}_{k=1}^\infty$ be generated by (**LISTA-CP**). For all parameters $\{\mathbf{W}^k,\theta_k\}_{k=0}^\infty$ satisfying some mild conditions (see [3]) and any sufficient small $\epsilon > 0$, we have*

$$\|\mathbf{x}^k(\Theta^k,\mathbf{b},\mathbf{x}^0)-\mathbf{x}^*\|_2 \ge \epsilon\|\mathbf{x}^*\|_2\exp(-\bar{c}k), \quad \text{with proability } (1-p\epsilon), \qquad (4)$$

where $\bar{c}, p > 0$ are constants that depend only on $\mathbf{D}$ and the distribution of $\mathbf{x}^*$.

Theorem 1 shows that (2) significantly simplifies the model without compromising the linear convergence rate of (**LISTA-CP**). Theorem 2 shows that, with high probability, this rate is optimal w.r.t. the order of convergence.
Applying (2) to (LISTA-CP), we propose:

TiLISTA: $\mathbf{x}^{k+1} = \eta_{\theta_k}(\mathbf{x}^k - \gamma_k\mathbf{W}^T(\mathbf{D}\mathbf{x}^k - \mathbf{b}))$, trainable parameters: $\{\gamma_k,\theta_k\}_k$ and $\mathbf{W}$.

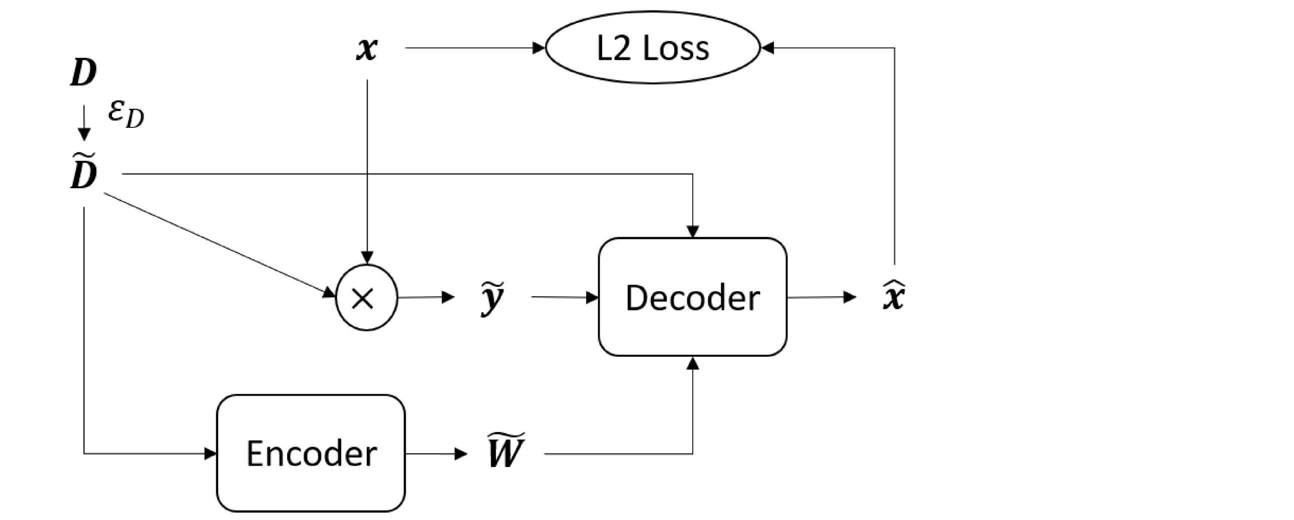
ALISTA: $\mathbf{x}^{k+1} = \eta_{\theta_k}(\mathbf{x}^k - \gamma_k\tilde{\mathbf{W}}^T(\mathbf{D}\mathbf{x}^k - \mathbf{b}))$, trainable parameters: $\{\gamma_k,\theta_k\}_k$.

Comparison of Parameter Spaces: All algorithms are truncated to K steps/layers.

LISTA[1]	LISTA-CP[2]	TiLISTA (this poster)	ALISTA (this poster)
$O(KM^2+K+MN)$	$O(KNM+K)$	$O(NM+K)$	$O(K)$

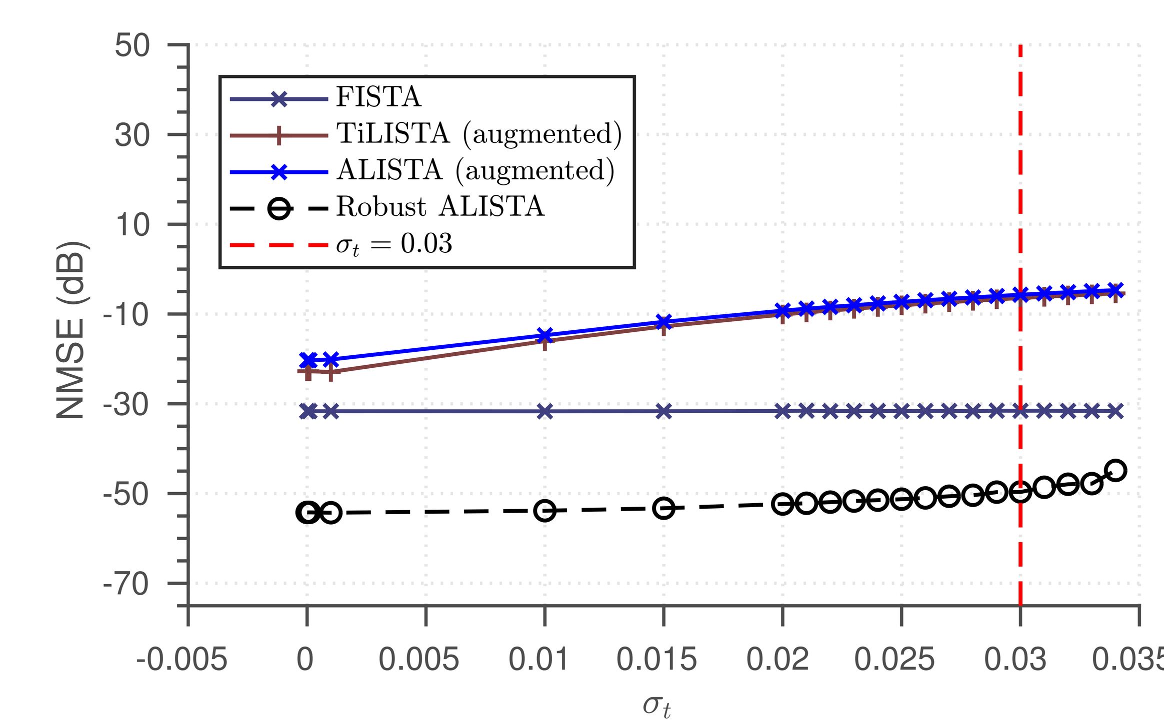
End-to-End Robust ALISTA

Motivation: If a noisy dictionary $\tilde{\mathbf{D}} = \mathbf{D} + \varepsilon_{\mathbf{D}}$ is observed, ALISTA performs bad.
Robust ALISTA: uses an **encoder** net to feed $\tilde{\mathbf{W}}$ into the **decoder**, an ALISTA net.



Encoder: unfolds and truncates the optimization algorithm for (1).
Decoder: an ALISTA net.

Numerical Validation: Robust ALISTA is much robust to the noise on $\mathbf{D}$, shown in the plot of recovery NMSE v.s. testing noise standard deviation below. The red vertical line represents maximal noise in training stage.



References

[1] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *ICML*, 2010.

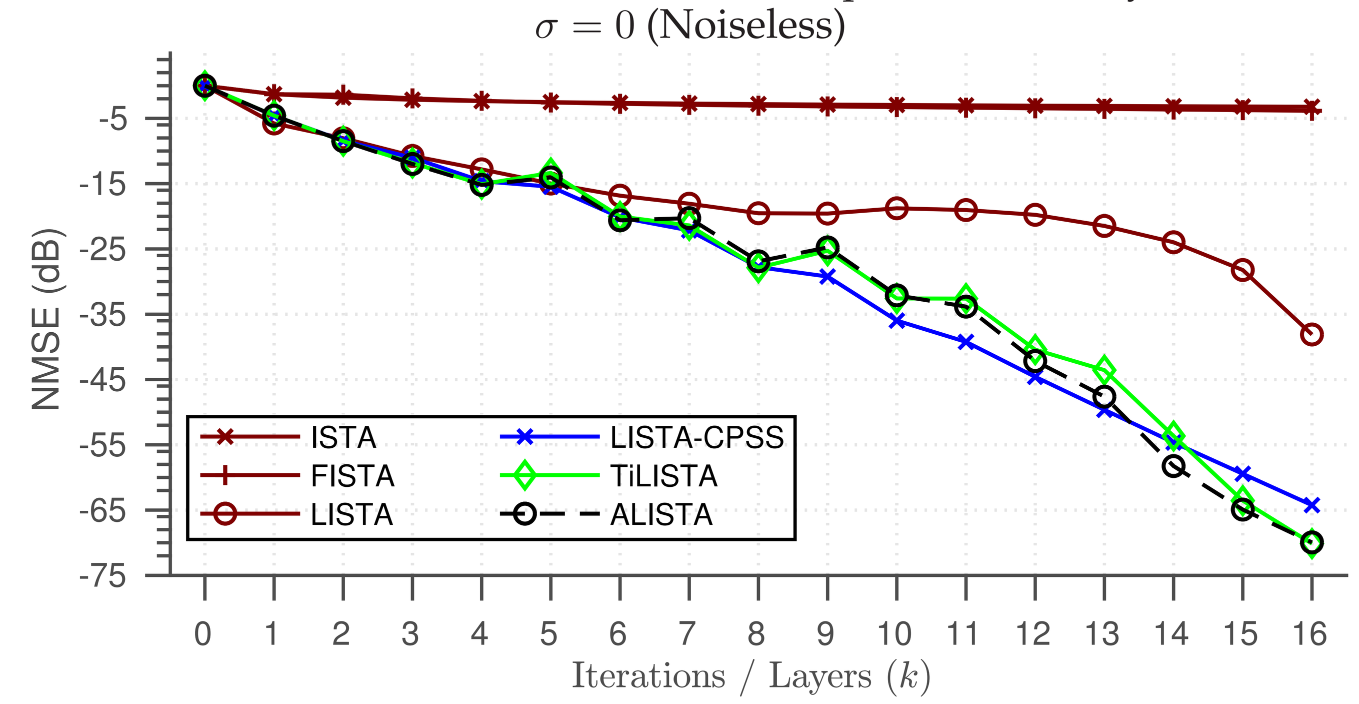
[2] X. Chen, J. Liu, Z. Wang, and W. Yin, "Theoretical linear convergence of unfolded ista and its practical weights and thresholds," in *NIPS*, 2018.

[3] J. Liu, X. Chen, Z. Wang, and W. Yin, "Alista: Analytic weights are as good as learned weights in lista," in *ICLR*, 2019.

Numerical Validation

Validation of Theorem 1: LISTA-CPSS, TiLISTA and ALISTA adopt the support selection technique developed in [2].
Noiseless Case: TiLISTA and ALISTA achieve even better NMSE compared to LISTA-CPSS in [2], with much fewer parameters and less training time. Training LISTA-CPSS takes 1.5 hours; training ALISTA takes only **6 minutes**.
Noisy Case: TiLISTA and ALISTA can still achieve comparable recovery error when high level noises exist.

$\sigma = 0$ (Noiseless)



$\sigma^2 \approx 2\times10^{-4}$ (SNR=30dB)