# Theoretical Linear Convergence of Unfolded ISTA and Its Practical Weights and Thresholds

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

## Overview

- We design a deep neural network to drastically improve LASSO speed and quality.
- Our network has fewer parameters and is easier to train.
- Our network achieves global linear convergence, better than sublinear and eventual-linear convergence of ISTA/FISTA.

## Unfold ISTA to Neural Network

**Problem:** Recover a sparse vector $x^*$ from its noisy measurements:

$$b = Ax^* + \varepsilon,$$

**LASSO:**

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

**Iterative shrinkage thresholding algorithm (ISTA) or FPC:**

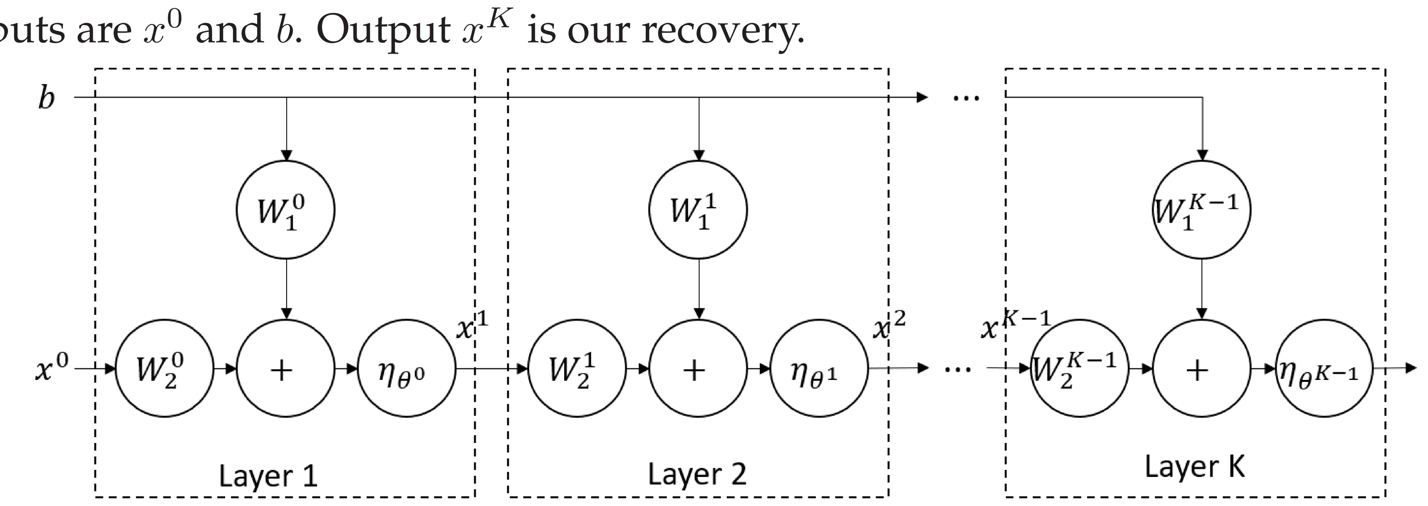
$$x^{k+1} = \eta_{\lambda/L}\Big(x^k + \frac{1}{L}A^T(b - Ax^k)\Big), \quad k = 0, 1, 2, \ldots \qquad \textbf{(ISTA)}$$

where $\eta_\theta$ is soft-thresholding, $\lambda$ and $L$ are selected by hand or cross-validation. ISTA converges **sublinearly** and **eventually-linearly** to a **LASSO solution, not** $x^*$.

**Neural network:** unrolls ISTA to a feed-forward neural network, replace $A, A^T$ in ISTA by free matrices, and truncates it to $K$ iterations (known as Learned ISTA or LISTA [1]):

$$x^{k+1} = \eta_{\theta^k}(W_1^k b + W_2^k x^k), \quad k = 0, 1, \cdots, K-1, \qquad \textbf{(LISTA)}$$

Inputs are $x^0$ and $b$. Output $x^K$ is our recovery.




**Training (deciding $\theta^k$, $W_1^k$, $W_2^k$)** For fixed $A$, we want to obtain parameters $\Theta^K = \{(W_1^k, W_2^k, \theta^k)\}_{k=0}^{K-1}$ such that $x^K$ is close to $x^*$ **(the ground truth)** for input $b = Ax^* + \varepsilon$ for almost all $x^*, \varepsilon$ following certain distribution. In another word, given the distributions of $x^*$ and $\varepsilon$, we

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

Stochastic gradient descent (SGD) can be applied to solve the above minimization problem. The gradient of $x^K$ on $\Theta^K$ can be obtained by the chain rule (back propagation).

**Issues:** largely many free parameters, training is slow

## 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.


## Links

arXiv preprint: Source codes:

## Improve by Weight Coupling (CP)

**New idea:** exploit certain dependencies among $W_1^k, W_2^k, \theta^k$ to simplify the network and improve the recovery result.

**Theorem 1 (Necessary Condition)** *Suppose $K = \infty$ and there is no noise $\varepsilon = 0$. Let $\{x^k\}_{k=1}^\infty$ be generated by (LISTA). If $x^k(\Theta^k, b, x^0) \to x^*$ as $k \to \infty$ uniformly for all sparse $x^*$, then the parameters $\{W_1^k, W_2^k, \theta^k\}_{k=0}^\infty$ are **not independent to each other** but must satisfy*

$$W_2^k - (I - W_1^k A) \to 0, \quad \theta^k \to 0, \quad \text{as } k \to \infty. \qquad (1)$$

**Weight simplification:** Couple $W_2^k = I - W_1^k A$ and simplify LISTA to:

$$x^{k+1} = \eta_{\theta^k}\Big(x^k + W_1^k(b - Ax^k)\Big), \quad k = 0, 1, \cdots, K-1. \qquad \textbf{(LISTA-CP)}$$

Now, only $\bar{\Theta}^K = \{W_1^k, \theta^k\}_{k=0}^{K-1}$ need to be trained, yet recovery is still fast.

**Theorem 2 (LISTA-CP trainability)** *Suppose $K = \infty$ and let $\{x^k\}_{k=1}^\infty$ be generated by (LISTA-CP). There exists a sequence of parameters $\{W_1^k, \theta^k\}$ such that*

$$\|x^k(\bar{\Theta}^k, b, x^0) - x^*\|_2 \le C_1 \exp(-ck) + C_2\sigma, \quad \forall k = 1, 2, \cdots,$$

*holds for all $(x^*, \varepsilon)$ satisfying some assumptions (see [2]), where $c, C_1, C_2 > 0$ are constants that depend only on $A$ and the distribution of $x^*$, and $\sigma$ is the noise level.*

If $\sigma = 0$ (noiseless case), the $k$th layer output $x^k$ converges to $x^*$ linearly:

$$\|x^k - x^*\|_2 \le C_1 e^{-ck}.$$

## Improve by Support Selection (SS)

*Before* applying soft thresholding in each layer, trust a percentages of largest entries as "true support" to bypass thresholding.

$$x^{k+1} = \eta_{\mathrm{ss}\,\theta^k}^{p^k}\Big(x^k + W_1^k(b - Ax^k)\Big), \quad k = 0, 1, \cdots, K-1. \qquad \textbf{(LISTA-CPSS)}$$

**Theorem 3 (Convergence of LISTA-CPSS)** *Suppose $K = \infty$ and let $\{x^k\}_{k=1}^\infty$ be generated by (LISTA-CPSS). There exists a sequence of parameters $\{W_1^k, \theta^k\}$ such that*

$$\|x^k(\bar{\Theta}^k, b, x^0) - x^*\|_2 \le C_1 \exp\Big(-\sum_{t=0}^{k-1}\tilde{c}_{\mathrm{ss}}^t\Big) + \tilde{C}_{\mathrm{ss}}\sigma, \quad \forall k = 1, 2, \cdots,$$

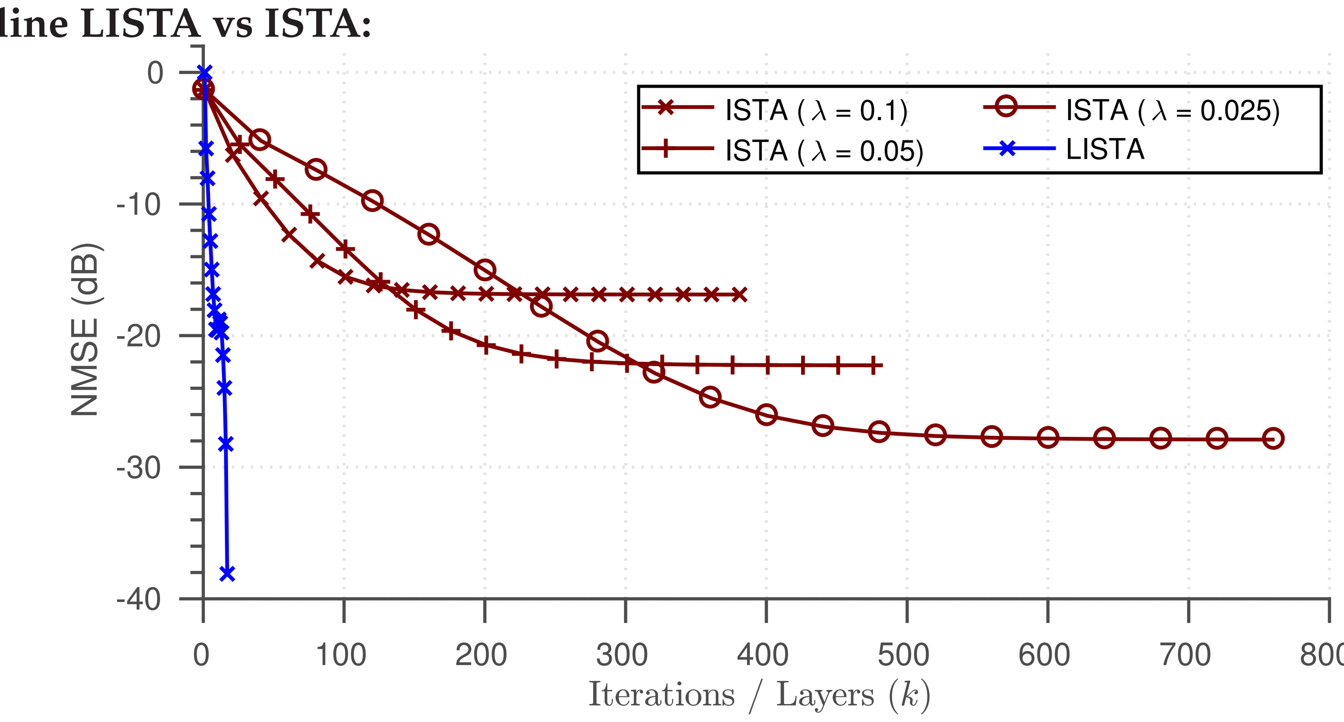
*holds for all $(x^*, \varepsilon)$ satisfying some assumptions. The convergence rate is better: $\tilde{c}_{\mathrm{ss}}^k > c$ for large enough $k$. The recovery error is better: $\tilde{C}_{\mathrm{ss}} < C_2$.*

## Numerical Validation

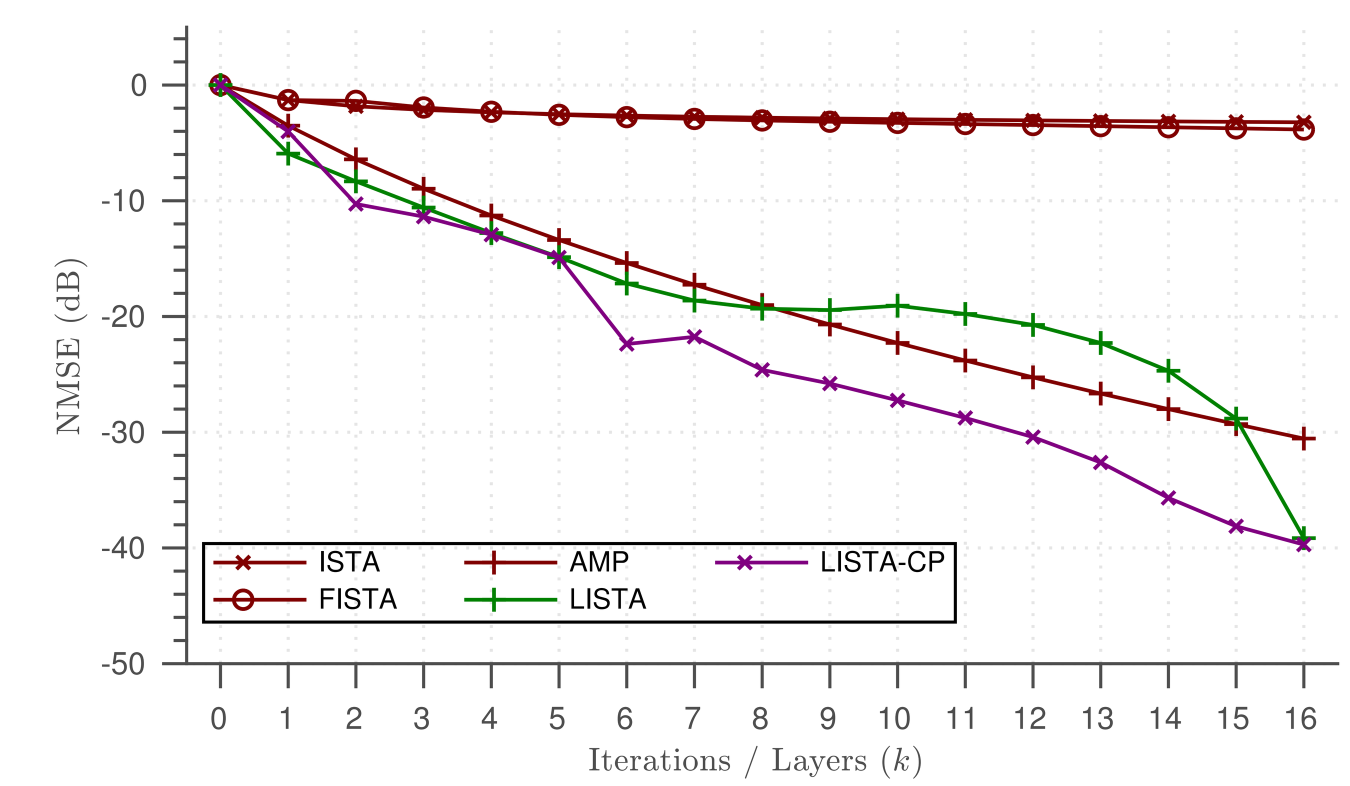
**Data:** fix $A \in \Re^{250\times500}$, $A_{ij} \sim N(0,1)$. Columns of $A$ are normalized. Sample $x^*$ with 10% nonzeros, each from the normal distribution. All plots below used the same 1000 samples.

**Recovery speeds:** ISTA≪LISTA[1]<LISTA-CP[2]<LISTA-CPSS[2]

**Baseline LISTA vs ISTA:**




**Validation of coupled LISTA ($\sigma$=0):** Coupled LISTA (LISTA-CP) achieves linear convergence and stabilize intermediate steps compared to LISTA.




## Numerical Validation — Support Selection

**Validation of support selection:**
LISTA with support selection (LISTA-SS) achieves linear convergence and better final performance.
Coupled LISTA with support selection (LISTA-CPSS) yields the best performance.