# Trainable ISTA for Sparse Signal Recovery

Dept. of Computer Science, Nagoya Institute of Technology
Satoshi Takabe

Daisuke Ito, Satoshi Takabe, and Tadashi Wadayama, 2018 IEEE ICC workshops;
arXiv: 1801.01978 (2018)

## Motivation

- Compressed sensing:
  Infer sparse signal $x \in \mathbb{R}^N$ from observation $y = Ax + w$
  ($A \in \mathbb{R}^{M\times N}$: sensing matrix ($M < N$), $w \in \mathbb{R}^M$: noise)
- LASSO as convex optimization problem with sparseness:
  $\min 1/2\|y - Ax\|_2^2 + \lambda\|x\|_1$
  $\rightarrow$ various algorithms; ISTA, AMP, etc.
- Deep neural network (DNN) framework:
  application to iterative algorithms such as ISTA, AMP.
  $\rightarrow$ tune parameters in algorithms by standard DNN techniques
- Our goal: trainable algorithm with fewer parameters and faster convergence

## Existing algorithm; ISTA and AMP

### Iterative Soft Thresholding Algorithm [1]

$$r_t = s_t + \beta A^T(y - As_t)$$
$$s_{t+1} = \eta(r_t;\tau)$$

Soft thresholding func. (element-wise): $\eta(r;\tau) = \mathrm{sign}(r)\max\{|r| - \tau, 0\}$
$\beta, \tau$: tuning parameters
$\rightarrow$ Learning version (LISTA) [2] tunes parameters with excellent performance.

### Approximate Message Passing (AMP) [3]

$$r_t = y - As_t + b_t r_{t-1}$$
$$s_{t+1} = \eta(s_t + A^T r_t; \tau_t)$$
$$b_t = M^{-1}\|s_t\|_0, \quad \tau_t = \theta M^{-1/2}\|r_t\|_2$$

- Faster convergence than ISTA in standard situations.
- Many valiants (GAMP, OAMP, VAMP, etc.)
- Learning version (LAMP) [4] successfully works for harder situation.

## Proposal: Trainable ISTA (TISTA)

Using the structure of OAMP [5], TISTA is written down as follows:

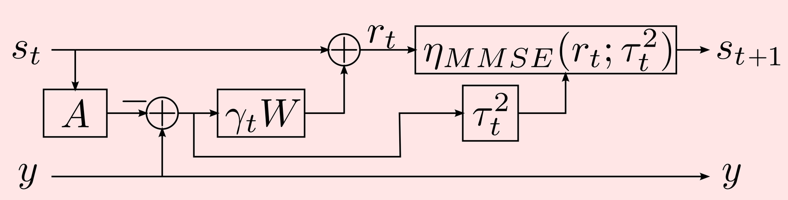
$$r_t = s_t + \gamma_t W(y - As_t) \quad (1)$$
$$s_{t+1} = \eta_{MMSE}(r_t;\tau_t^2) \quad (2)$$
$$v_t^2 = \max\left\{\frac{\|y - As_t\|_2^2 - M\sigma^2}{\mathrm{tr}(A^TA)}, \epsilon\right\} \quad (3)$$
$$\tau_t^2 = N^{-1}v_t^2(N + \gamma_t(\gamma_t - 2)M) + N^{-1}\gamma_t^2\sigma^2\mathrm{tr}(WW^T) \quad (4)$$

Output after $T$ iterations/layers: $\hat{x} = s_T$

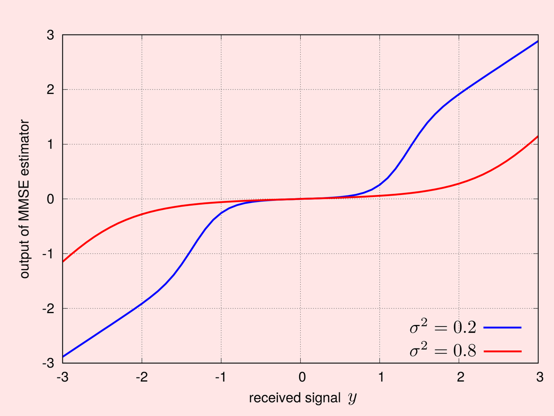
- $\{\gamma_t\}_{t=0}^{T-1}$: trainable parameter controlling shrinkage
  $\rightarrow$ much fewer than LISTA and LAMP ($O(MNT)$)
- Computational cost: $O(N^2)$ equivalent to ISTA and AMP
  (with one $O(N^3)$ pre-computation for $W$)
- $W = A^T(AA^T)^{-1}$ as linear estimator (1)
- $\eta_{MMSE}$: element-wise MMSE estimator as nonlinear estimator (2)
  e.g.) Bernoulli-Gaussian case: $x \sim (1-p)\delta(x) + pF(x;\alpha^2)$
  $(F(x;\alpha^2) = e^{-x^2/\alpha^2}/\sqrt{2\pi\alpha^2})$, $y = x + F(w;\tau^2)$
  $$\eta_{MMSE}(y;\tau^2) = \left(\frac{\alpha^2 y}{\alpha^2+\tau^2}\right)\frac{pF(y;\alpha^2+\tau^2)}{(1-p)F(y;\tau^2) + pF(y;\alpha^2+\tau^2)}$$
- $\sigma^2$: variance of Gaussian noise, $\epsilon$: small const. (e.g. $10^{-9}$)
- (3), (4): error-variance estimator (derived under some assumptions)

Learning strategy:

- Implemented by TensorFlow
- Incremental training: train up to $t$-th layer with MSE loss func.
  $\rightarrow$ train up to $(t+1)$-th layer with new training data.
  # of mini-batches $= 200$,
  mini-batch size $1000$
- Adam optimizer
- Only 6 min. for $(N,M) = (500, 250)$ and $T = 7$ (using GPU)




Layer structure of TISTA

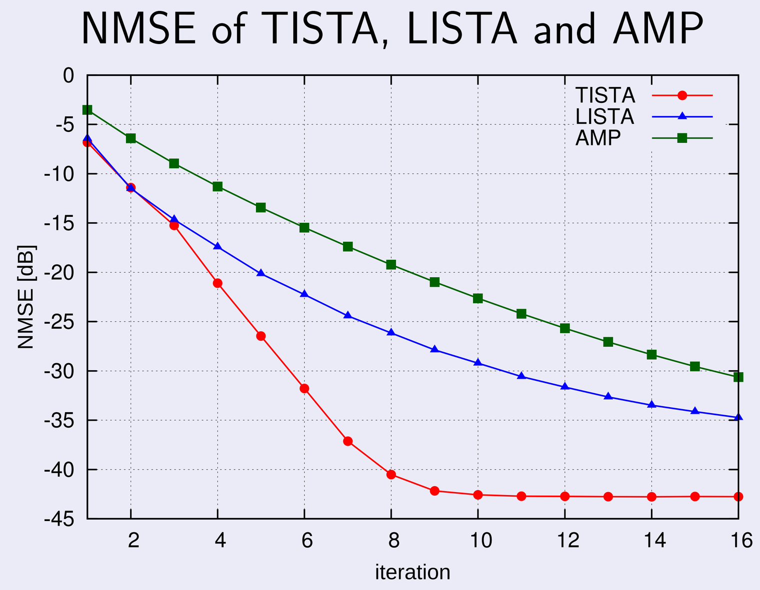


$\eta_{MMSE}$ for Bernoulli-Gaussian ($\alpha^2 = 1$)

## Main Results: Standard Setup

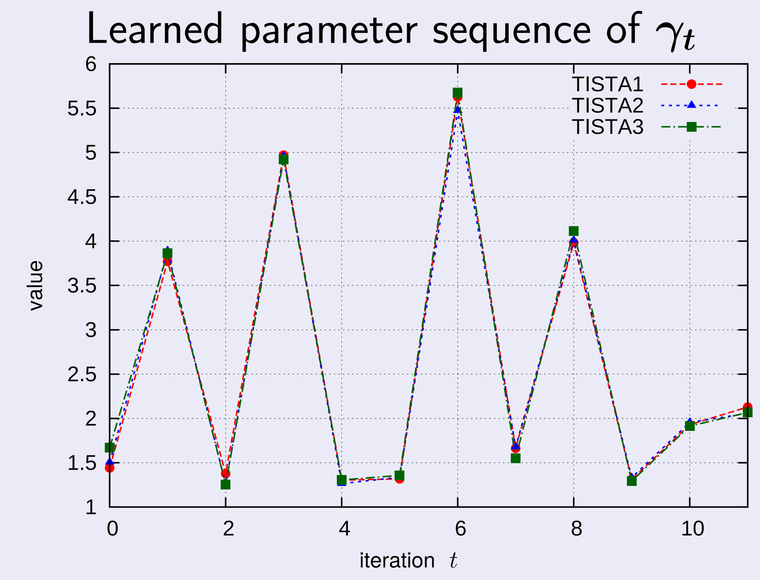
- For $x$, ratio of non-zero components: $p = 0.1$,
  std of non-zero components: $\alpha = 1$.
- For noise $w$, Gaussian dist. with zero mean and
  $\mathrm{SNR}(= \mathbb{E}[\|Ax\|_2^2]/\mathbb{E}[\|w\|_2^2]) = 40$ [dB].
- For $A$, $A_{ij} \sim \mathcal{N}(0, 1/M)$ (i.i.d.), $N = 250$, $M = 500$.
- Evaluate $\mathrm{NMSE} = 10\log_{10}(\|\hat{x} - x\|_2^2/\|x\|_2^2)$.

NMSE of TISTA, LISTA and AMP




Faster convergence than LISTA and AMP.

Learned parameter sequence of $\gamma_t$

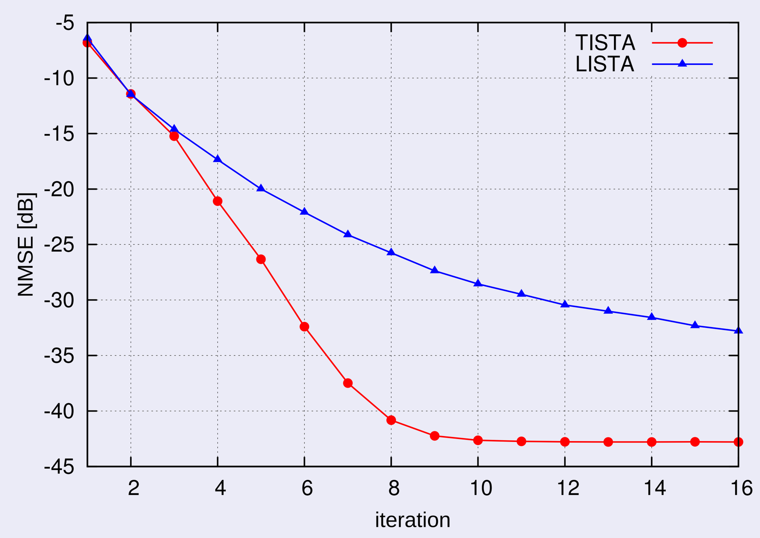



Coincide with each other (with 3 trials).

## Main Results: Harder Setup

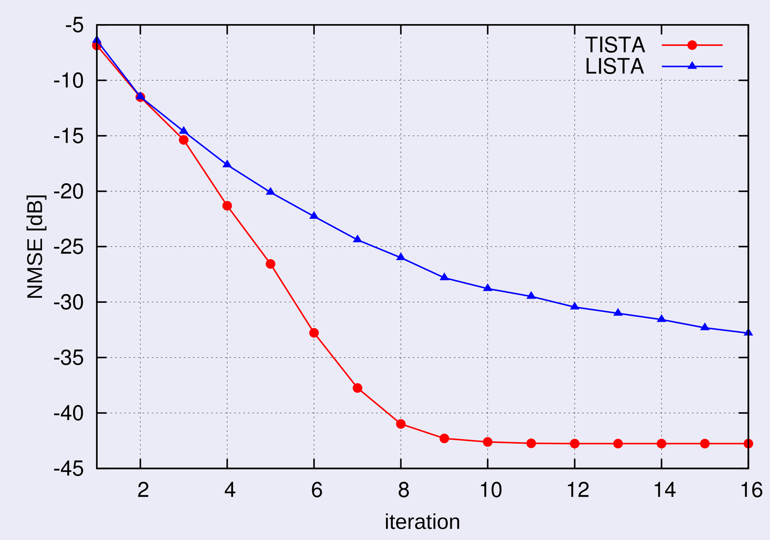
Harder setup of sensing matrices where AMP cannot converge.

Sensing matrix with large variance
$A_{ij} \sim \mathcal{N}(0,1)$




Fast convergence within 10 iterations.

Sensing matrix with (uniformly random) binary entries

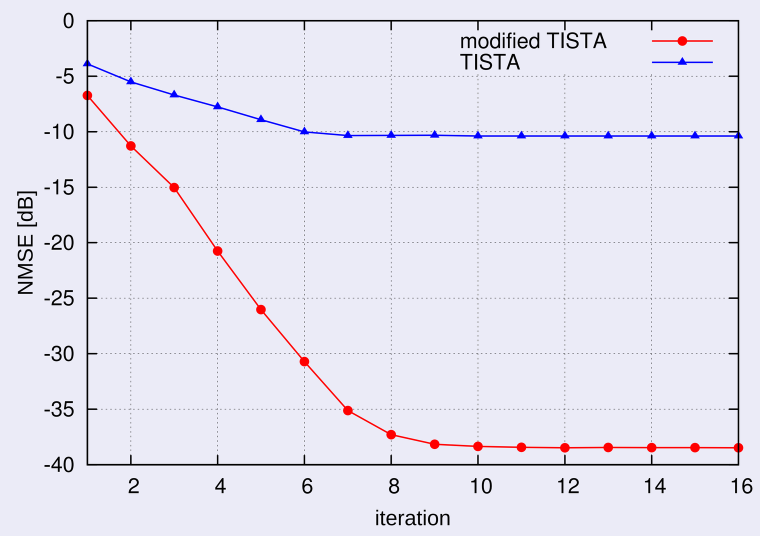



Well works even for non-Gaussian case.

## Extensions

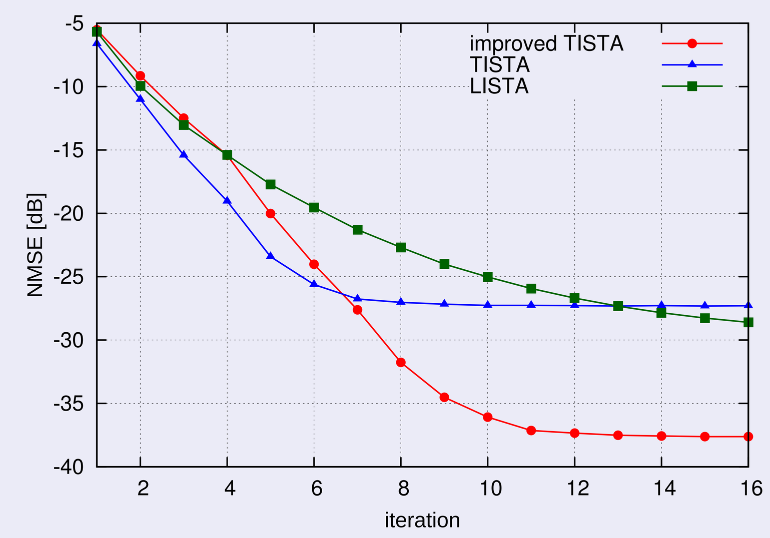
Small modifications of TISTA realizes more robustness and fast convergence.

Sensing matrix with non-zero mean entries
$A_{ij} \sim \mathcal{N}(1, 1/M)$, SNR = 60[db]
Modification: mean removal, i.e.,
$A'_{ij} = A_{ij} - \mu_A$ ($\mu_A$: mean of $A_{ij}$'s)




Modified TISTA successfully recovers signals.

Sensing matrix with large condition number
$\kappa = 1000$, SNR = 60[db]
Modification: $W = A^T(AAT + \beta I)^{-1}$
($\beta = 5 \times 10^{-4}$)




Modified TISTA shows nice signal recovery.

## Summary and future works

- TISTA for compressed sensing as sparse signal recovery
  - Trainable algorithm based on OAMP and MMSE estimator
  - Only contains $T$ trainable parameters $\{\gamma_t\}$
    $\rightarrow$ fast training process/high scalability
  - Same computational cost with AMP and ISTA
  - Training process with standard DNN techniques
  - Faster convergence than AMP and LISTA
  - High robustness: sensing matrices with large variance, large mean
    even applicable to binary sensing matrices or matrices with large condition number
- Possible future works
  - Discrete signal case, e.g., application to wireless communication
    Imanishi, Takabe, Wadayama, arXiv:1806.10827 for massive MIMO detection
  - Theoretical analysis $\rightarrow$ zig-zag shape of $\{\gamma_t\}$?

Reference:

[1] K. Gregor and Y. LeCun, ICML 2010, 399-406 (2010).
[2] A. Chambolle et. al, IEEE Trans. Image Process., 7, 319-335 (1998).
[3] D. L. Donoho, A. Maleki, and A. Montanari, PNAS, 106, 18914-18919 (2009).
[4] M. Borgerding and P. Schniter, 2016 IEEE GlobalSIP, 227-231 (2016).
[5] J. Ma and L. Ping, IEEE Access, 5, 2020-2033 (2017).