Interpretable LASSO

Ye Yuan

Xiaohan Chen

Our Problem

- We try to solve the inverse problem:

$$y = Ax$$

- Given signal y, dictionary A, recover x
- $\mathrm{y} \in R^m, \mathrm{x} \in R^n$: $\mathrm{m} \ll \mathrm{n}$ --- An underdetermined system
- However, we have a very strong prior knowledge:

  x is sparse --- most of entries of x are zero

LASSO

- LASSO: Least Absolute Shrinkage and Selection Operator
  First introduced by Robert Tibshirani in 1996:

$$x^{\text{lasso}} = \text{argmin}_x \ \frac{1}{2}||y - Ax||_2^2 \quad s.t. \quad ||x||_1 \leq t$$

- which is equivalent to the following formulation:

$$x^{\text{lasso}} = \text{argmin}_x \ \frac{1}{2}||y - Ax||_2^2 + \lambda||x||_1$$

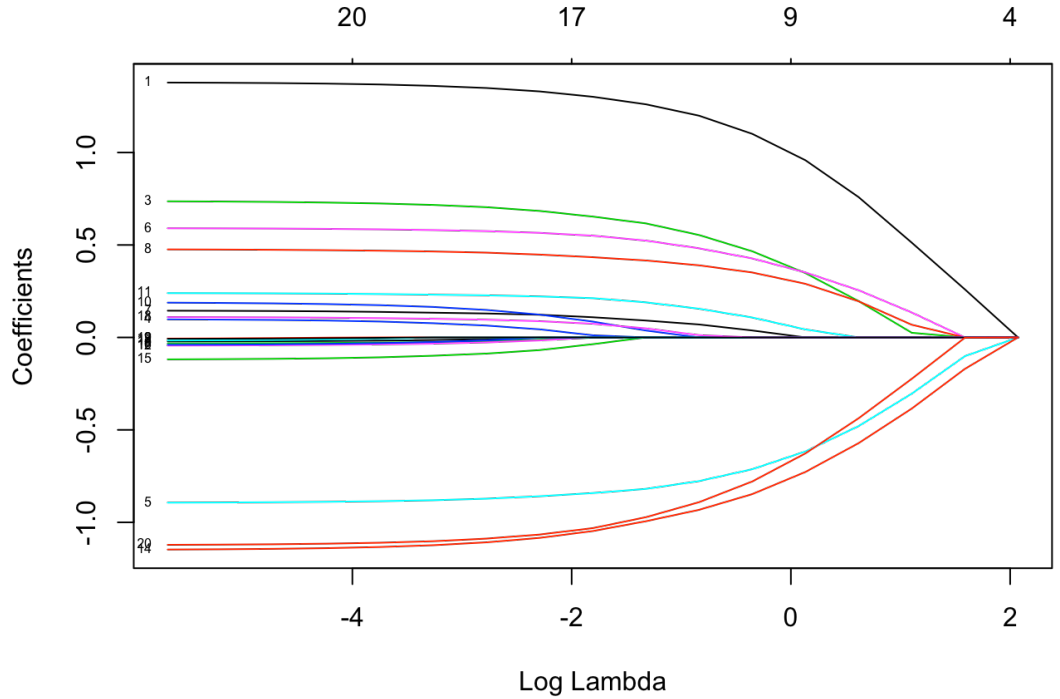
- t and λ have one-to-one correspondence

Interpretability of LASSO

- LASSO enjoys good interpretability --- feature selection

$$x^{\text{lasso}} = \text{argmin}_x \ \frac{1}{2}||y - Ax||_2^2 + \lambda||x||_1$$

- Regularization path

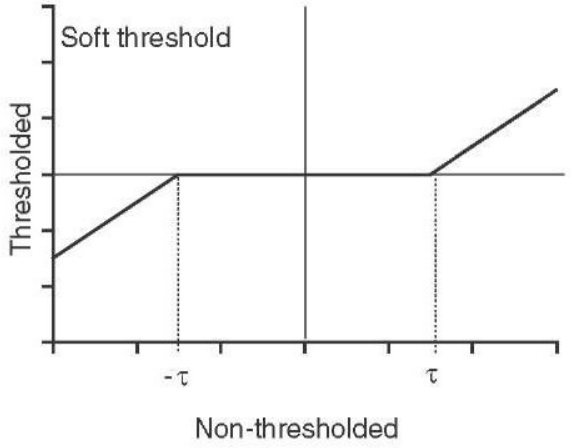


ISTA

- ISTA: Iterative Soft-Thresholding Algorithm

$$x^{k+1} = \eta_\tau(x^k + \alpha A^T(y - Ax^k)) = \eta_\tau(W x^k + B(y))$$

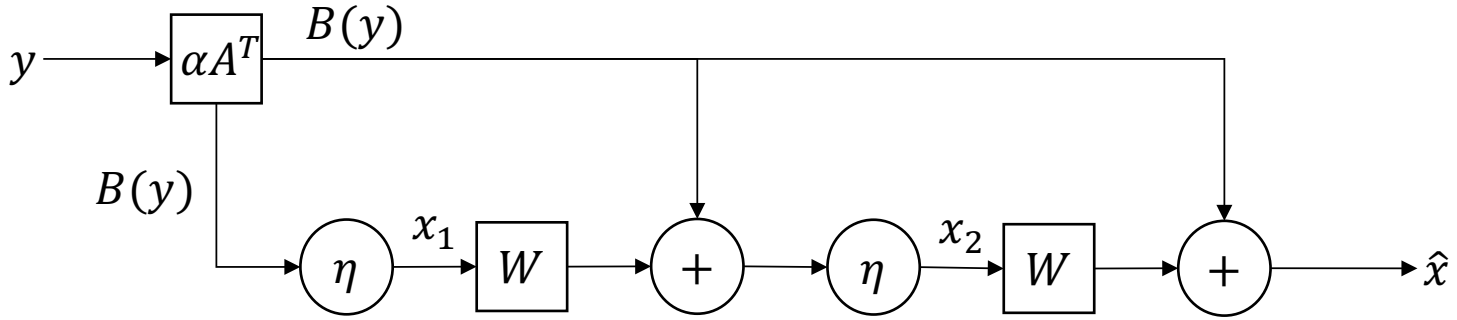
$$\eta_\tau(v) = \begin{cases} v - \tau, & v > \tau \\ 0, & |v| \leq \tau \\ v + \tau, & v < -\tau \end{cases}$$



- ISTA is a proximal gradient descent algorithm
  - First gradient descent with step length α
  - Perform proximal mapping considering the l_1 norm constraint

LISTA (LeCun, 2010)

- Unfold ISTA iteration



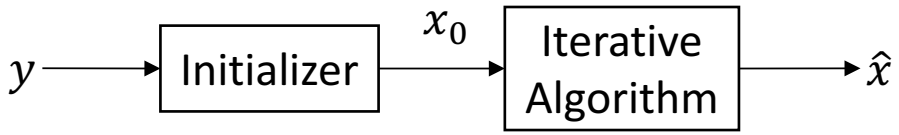
- Truncate iteration to T steps --- A T-layer neural network
- Initialize $W = I - \alpha A^T A$ and $B = \alpha A^T y$
- Feed Data to the model and learn the weights with back-prop

Our work

- LISTA is fast, however at the cost of interpretability
- LISTA's initial input is $\mathbf{0}$, which is a very poor initialization $A^T(AA^T)^{-1}y$ would be much better one

- So, can we just learn how to initialize our input x_0 but keep the structure of dictionary A and the good interpretability of LASSO?

Our Model

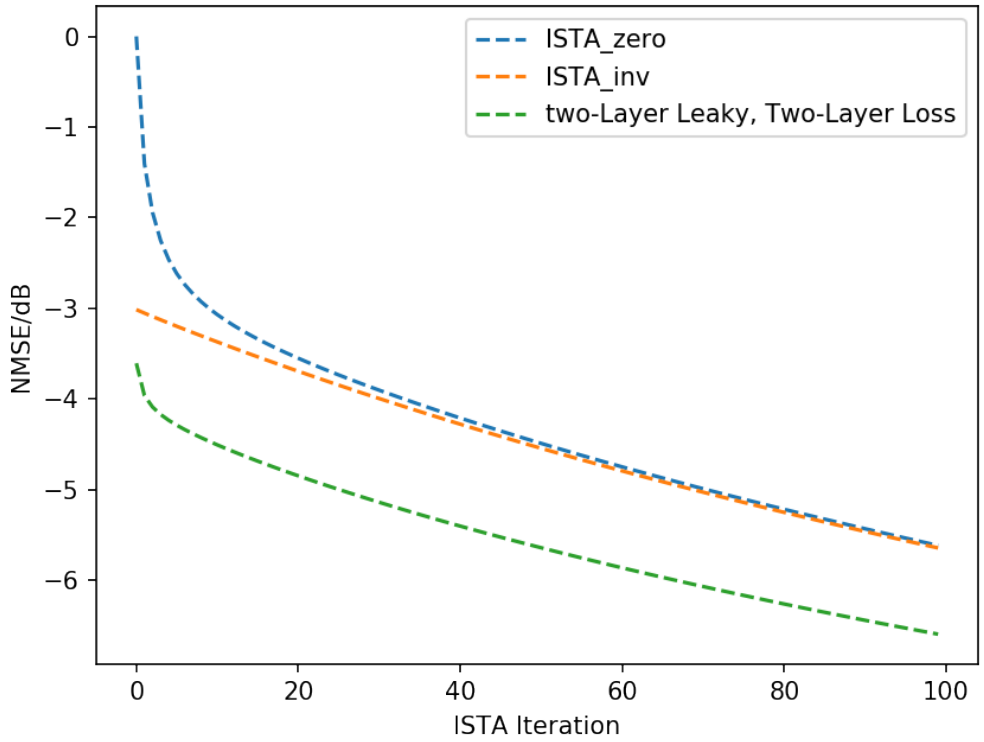


- Loss function: $Loss = \left|\left|\hat{x}(y) - x\right|\right|_2^2$
  Use Adam Optimizer

- The design of the network here is the main task
- Currently we tried the simplest one-layer and two-layer fully connected with sigmoid or leaky ReLU activation

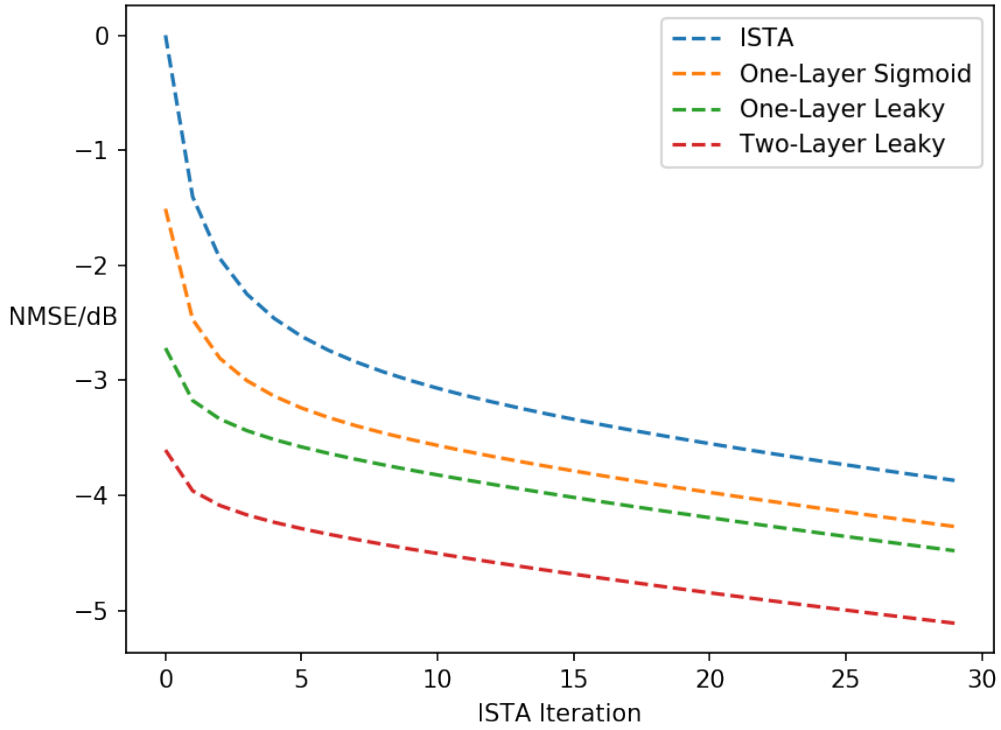
Some Results

- Compare different initial inputs to ISTA



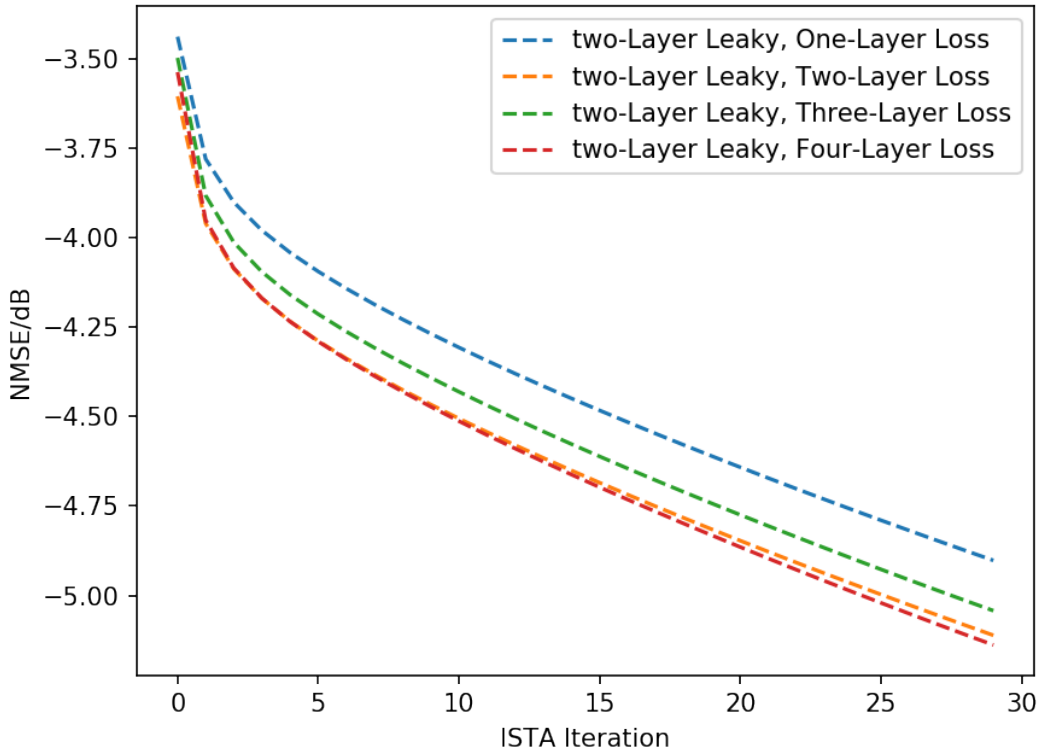
Some Results

- Use different initialization network and two-step ISTA after that



Some Results

- Use two-layer fully connected with leaky ReLU activation and different numbers of steps of ISTA after that



Conclusion and Future Work

- Our initialization model works.

- More experiments need to be done.
- Needs more explanation on interpretability.

Thanks!