SURE-TISTA: A SIGNAL RECOVERY NETWORK FOR COMPRESSED SENSING

Mengcheng Yao[⋆] *Jian Dang*[⋆] *Zaichen Zhang*[⋆†] *Liang Wu*[⋆]

[⋆] National Mobile Communications Research Lab, Southeast University, Nanjing, China, 210096
[†] Corresponding E-mail: zczhang@seu.edu.cn

ABSTRACT

Deep neural network (DNN) has a wide range of applications in various fields, including solving sparse inverse problems. In this paper, we propose a novel network called the Stein's unbiased risk estimate based-trainable iterative thresholding algorithm (SURE-TISTA) for sparse signal recovery problems. Without prior information, SURE-TISTA outperforms TISTA, an algorithm based on the minimum mean squared error (MMSE) estimator. SURE-TISTA also shows a great robustness in many cases including large-scale and large-variance problems. Meanwhile, SURE-TISTA uses fewer learnable variables to achieve similar performance as learned approximate message passing (LAMP), which has more learnable parameters. Without any error measure estimator, SURE-TISTA achieves a near MMSE-based performance. Our numerical results indicate that SURE-TISTA is superior to TISTA and other traditional algorithms in many aspects, which can be promising in image denoising.

Index Terms— compressed sensing, deep learning, iterative thresholding algorithm, the Steins unbiased risk estimate, error measure

1. INTRODUCTION

Compressed sensing (CS) refers to a technique that recovers sparse signal accurately with a sampling ratio far below the Nyquist rate [1]. Consider the problem below of signal recovery from a noisy linear observation $\boldsymbol{y} \in \mathbb{R}^M$,

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{M\times N}(M \ll N)$ is a measurement matrix, $\boldsymbol{x} \in \mathbb{R}^N$ is the signal to be recovered and $\boldsymbol{n} \in \mathbb{R}^M$ is a vector of additive white Gaussian noise (AWGN) samples with zero mean and variance σ^2.

Many iterative algorithms have been developed for the sparse inverse problem. The iterative shrinkage thresholding algorithm (ISTA) [2] is one of the best-known algorithms. It aims to recover the original signal through solving Lasso problem [3]. Recently Donoho and Maleki proposed approximate message-passing algorithms (AMP), which can be asymptotically characterized by a simple scalar recursion called state evolution (SE) [1]. However, AMP can only be utilized when the measurement matrix has independent and identically distributed (i.i.d.) Gaussian entries. Orthogonal AMP (OAMP) is proposed to overcome the restrictions. In OAMP, Onsager corrected term, the key part of AMP, vanishes [4].

Recently, deep learning (DL) has been applied to various researches due to its powerful ability of solving complex problems. In addition to natural language processing (NLP) and pattern recognition [5,6], DL is gradually applied to communication systems and signal processing [7]. A promising method of DL, called deep unfolding, can "unfold" an existing algorithm into a signal-flow graph. Each layer contains an iteration and standard deep learning techniques can be used to find optimal parameters of the algorithm [8]. Learned ISTA (LISTA) [9], learned approximate message passing (LAMP) [10] and trainable ISTA (TISTA) [11] are derived from deep unfolding and iterative recovery algorithm. And they all outperform their original algorithms.

When probability density function of random variable x is known, the optimal estimator is the minimum mean squared error (MMSE) [12] estimator. For instance, the optimal structure of OAMP contains an MMSE estimator and TISTA has an MMSE estimator-based shrinkage unit [11]. However, such prior information may be unavailable in particular environment. SURE can achieve the Bayesian optimal performance without knowing prior information [13], which can be applicable to iterative recovery algorithm.

In this paper, we propose a novel algorithm called the Stein's unbiased risk estimate based-trainable iterative thresholding algorithm (SURE-TISTA) for sparse signal recovery. SURE-TISTA utilizes DL network and achieves great performance without error measure estimator and prior information. Using fewer learnable parameters, SURE-TISTA estimates the error measure terms correctly. Numerical results demonstrate that SURE-TISTA outperforms TISTA in many aspects, especially in large-scale and large-variance problems.

2. ITERATIVE RECOVERY ALGORITHMS

We first review several recovery iterative algorithms to solve the original problem (1) in this section.

This work is supported by NSFC projects (61571105, 61501109, and 61601119).

2.1. ISTA

ISTA [2] is a well-known algorithm defined by the following simple iterations:

$$\begin{aligned} \boldsymbol{r}_t &= \hat{\boldsymbol{s}}_t + \beta \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{s}}_t) \\ \hat{\boldsymbol{s}}_{t+1} &= \eta_{st}(\boldsymbol{r}_t; \lambda) \end{aligned} \tag{2}$$

where β is a step size and $\hat{\boldsymbol{s}}_t$ is initialized by $\hat{\boldsymbol{s}}_0 = \boldsymbol{0}$. $\eta_{st}(\cdot; \lambda) : \mathbb{R}^N \to \mathbb{R}^N$ is a soft thresholding shrinkage function (Readers are referred to [3] for details).

2.2. AMP and OAMP

AMP is a new recovery algorithm and manifests as:

$$\begin{aligned} \boldsymbol{r}_t &= \hat{\boldsymbol{s}}_t + \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{s}}_t) + \boldsymbol{q}_t \\ \hat{\boldsymbol{s}}_{t+1} &= \eta_{st}(\boldsymbol{r}_t) \end{aligned} \tag{3}$$

Initialized with $\hat{\boldsymbol{s}}_0 = \boldsymbol{0}$ and $\boldsymbol{z}_0 = \boldsymbol{y}$, AMP has a key part $\boldsymbol{q}_t$ called 'Onsager correction term' (details in [1]). AMP outperforms ISTA in convergence speed, but it does not work well unless $\boldsymbol{A}_{i,j} \sim \mathcal{N}(0, M^{-1})$.

Based on de-correlated linear estimation and divergence-free non-linear estimation, OAMP is proposed as:

$$\begin{aligned} \boldsymbol{r}_t &= \hat{\boldsymbol{s}}_t + \boldsymbol{W}(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{s}}_t) \\ \hat{\boldsymbol{s}}_{t+1} &= \eta(\boldsymbol{r}_t) \end{aligned} \tag{4}$$

In this algorithm, two error measures can be estimated as,

$$\hat{\tau}_t^2 = \frac{1}{N} tr\left(\boldsymbol{B}\boldsymbol{B}^T\right) \cdot \hat{v}_t^2 + \frac{1}{N} tr\left(\boldsymbol{W}\boldsymbol{W}^T\right) \cdot \sigma^2 \tag{5}$$

$$\hat{v}_t^2 = max\{\frac{\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}_t\|^2 - M \cdot \sigma^2}{tr\left(\boldsymbol{A}^T\boldsymbol{A}\right)}, \varepsilon\} \tag{6}$$

where $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{W}\boldsymbol{A}$, and ε is a quite small positive constant. The de-correlated linear estimator $\boldsymbol{W}$ can be constructed from $\boldsymbol{A}$. η_t is a divergence-free estimator (i.e. a specific function) and its thresholding $\hat{\tau}_t$ can be estimated from $\hat{v}_t$. It has been proved that SE for OAMP is accurate for general unitarily-invariant matrices, which is advantageous over AMP.

3. PROPOSED SURE-TISTA

In this section, we propose a novel signal recovery network called SURE-TISTA based on ISTA, SURE framework and deep unfolding network.

3.1. Deep Unfolding Network Applied to SURE-TISTA

Using a T-layer structure, deep neural network (DNN) can approximate functions applied to a certain algorithm that maps the input $\boldsymbol{x}_0 \in \mathbb{R}^N$ to the output $\boldsymbol{x}_T \in \mathbb{R}^N$, i.e.

$$\boldsymbol{x}_T = f_{T-1}(\cdots f_t(\cdots f_2(f_1(\boldsymbol{x}_0; \theta_1); \theta_2) \cdots; \theta_t); \cdots; \theta_{T-1}) \tag{7}$$

where $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_{T-1}]$ denotes the parameters leading to the best approximation mapping function in this feed-forward networks. It is learned during training process to minimize the following loss function:

$$\mathcal{L}(\boldsymbol{\theta}) = \langle |\boldsymbol{x}_T(\boldsymbol{\theta}) - \boldsymbol{x}_0|^2 \rangle \tag{8}$$

where $\langle \boldsymbol{x} \rangle = \frac{1}{N}\sum_i x_i$ as $\boldsymbol{x} \in \mathbb{R}^N$. When $f_1, f_2, ..., f_{T-1}$ turn to be certain functions, the final output $x_T(\boldsymbol{\theta})$ and loss function $\mathcal{L}(\boldsymbol{\theta})$ are only correlated to the chosen $\boldsymbol{\theta}$.

The main idea of deep unfolding [8] is to map a specific algorithm into a DNN layers. In this new architecture, belief propagation (BP) is used as the inference algorithm, which is a message passing (MP) algorithm based on graph factor. Thus, we can propose a new network tuning its parameters based on MSE optimization.

3.2. Recursive Formula of SURE-TISTA

The recursive formula of SURE-TISTA is presented below to solve problem (1), which contains linear estimation (LE) and non-linear estimation (NLE):

$$LE: \qquad \boldsymbol{r}_t = \boldsymbol{s}_t + \boldsymbol{W}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}_t) \tag{9}$$

$$NLE: \qquad \boldsymbol{s}_{t+1} = \hat{I}(\boldsymbol{r}_t; \alpha) \tag{10}$$

where $\boldsymbol{W} = \frac{N}{tr(\hat{\boldsymbol{W}}\boldsymbol{A})}\hat{\boldsymbol{W}}$, $\hat{\boldsymbol{W}} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}$ and α is a trainable variable. The initial condition is $\boldsymbol{s_0} = \boldsymbol{0}$. And the estimator $\hat{I}(\cdot)$ is a denoiser function and is divergence-free, which is inspired by [4].

Definition 1 *A divergence-free denoiser $\hat{I}(\cdot)$ is constructed as [4]*

$$\hat{I}(r_t) = C(I(r_t) - \mathrm{div}\{I(r_t)\} \cdot r_t) \tag{11}$$

where $I(\cdot)$ is an arbitrary function and C is a constant.

It can be proved that $\mathrm{div}\{\hat{I}(r_t)\} = 0$ in high-dimension.

3.3. Parameter Optimization Based on SURE Framework

In this subsection, we will show how SURE framework can be utilized over TISTA and how we can achieve an optimal C in the denoiser.

Consider the problem below to find a specific function $\vartheta(\cdot)$ that can minimize the following MSE:

$$MSE = \langle |\vartheta(y) - x|^2 \rangle \tag{12}$$

We use x to denote the sparse vector to be recovered. In practice, we can only obtain a "noisy" version observation $y = x + n$. Since x is a random variable, we define an expectation function $E(\cdot)$. In order to achieve an unbiased estimation of (12), we leverage the SURE framework: according to [14], an unbiased estimation of MSE is:

$$\varepsilon = \langle \vartheta(y)^2 - 2y\vartheta(y) + 2\tau^2 \,\mathrm{div}\{\vartheta(y)\} \rangle + \langle x^2 \rangle \tag{13}$$

where τ^2 is the estimation of the effective noise variance. Thus, $E(\varepsilon) = E(MSE)$.

According to [1], we consider a signal recovery problem as a denoising problem and assume the observation at each iteration t, i.e. the original signal with some Gaussian perturbation is $\boldsymbol{r}_t = \boldsymbol{x}_0 + \tau_t \boldsymbol{w}_t$, where $w_t \in \mathcal{N}(0,1)$ and τ_t^2 is the estimation of the effective noise variance. Our goal is to optimize C in (11),

$$C = \arg\min \langle |\hat{I}(\boldsymbol{r}_t) - \boldsymbol{x}_0|^2 \rangle \tag{14}$$

We denote an unbiased estimation of $\langle |\hat{I}(\boldsymbol{r}_t) - \boldsymbol{x}_0|^2 \rangle$ as χ, according to (13),

$$\begin{aligned} E\{\chi\} &= E\{\langle \hat{I}(\boldsymbol{r}_t)^2 - 2\boldsymbol{r}_t \hat{I}(\boldsymbol{r}_t) + 2\tau^2 \operatorname{div} \hat{I}(\boldsymbol{r}_t) \rangle + \langle \boldsymbol{x}_0^2 \rangle\} \\ &= E\{\langle (\hat{I}(\boldsymbol{r}_t) - \boldsymbol{r}_t)^2 \rangle\} - \left(E\{\langle (\boldsymbol{x}_0 + \sqrt{\tau_t}\boldsymbol{w}_t)^2 \rangle\} - \langle \boldsymbol{x}_0^2 \rangle\right) \\ &= \frac{1}{N} E\left(|\hat{I}(\boldsymbol{r}_t) - \boldsymbol{r}_t)|^2 \right) - \tau_t^2 \end{aligned}$$

Leveraging least square (LS) ,we can solve the problem $C = \arg\min E\{\chi\}$ to achieve the optimal C:

$$C_{optimal} = \left(K(\boldsymbol{r}_t) K(\boldsymbol{r}_t)^T\right)^{-1} K(\boldsymbol{r}_t) \boldsymbol{r}_t \tag{15}$$

where $K(y) = I(y) - \operatorname{div}\{I(y)\} \cdot y$.

In this paper, we introduce a pointwise exponential shrinkage function to construct the denoiser:

$$\eta(y) = \sum_{k=1}^{K} a_k \theta_k(y), \quad \theta_k(y) = y e^{-(k-1)\frac{y^2}{2T^2}} \tag{16}$$

where k is the number of parameters, a_k is the coefficient of θ_k. T is a decision factor and is correlated to τ_t^2. We recommend $K = 2$ [14] and define a family of shrinkage function to minimize MSE:

$$\begin{cases} \eta_1(y) = a_1 \theta_1(y) \\ \eta_2(y) = a_2 \theta_2(y) \end{cases} \tag{17}$$

Then we concatenate them together and calculate the divergence,

$$\boldsymbol{I}(\boldsymbol{r}_t) \equiv \begin{pmatrix} \eta_1(\boldsymbol{r}_t) \\ \eta_2(\boldsymbol{r}_t) \end{pmatrix} = \boldsymbol{\mathcal{A}} \begin{pmatrix} \theta_1(\boldsymbol{r}_t) \\ \theta_2(\boldsymbol{r}_t) \end{pmatrix} \tag{18}$$

$$\operatorname{div} \boldsymbol{I}(\boldsymbol{r}_t) \equiv \boldsymbol{\mathcal{A}} \begin{pmatrix} \operatorname{div}\{\eta_1(\boldsymbol{r}_t)\} & \operatorname{div}\{\eta_2(\boldsymbol{r}_t)\} \end{pmatrix}^T \tag{19}$$

where $\boldsymbol{\mathcal{A}} = \operatorname{diag}(a_1, a_2)$. In this paper, we consider $\boldsymbol{E}$ as an identical matrix. Then we can deduce that

$$\boldsymbol{C}_{optimal} = \left(\boldsymbol{K}(\boldsymbol{r}_t) \boldsymbol{K}(\boldsymbol{r}_t)^T + \gamma \boldsymbol{E}\right)^{-1} \boldsymbol{K}(\boldsymbol{r}_t) \cdot \boldsymbol{r}_t \tag{20}$$

where kernel function $\boldsymbol{K}(\boldsymbol{r}_t) = \boldsymbol{I}(\boldsymbol{r}_t) - \operatorname{div} \boldsymbol{I}(\boldsymbol{r}_t) \cdot \boldsymbol{r}_t^T$. We introduce an ℓ_2 regularization here, where γ is a small positive constant. Then according to (11), we can achieve the divergence-free denoiser $\hat{\boldsymbol{I}}(\boldsymbol{r}_t)$.

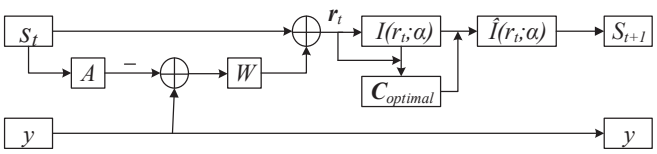


Fig. 1. A specific layer structure of SURE-TISTA with trainable parameter α.

3.4. General Structure of SURE-TISTA

The structure of SURE-TISTA consists of three parts: recursive formula (9, 10), a divergence-free denoiser and deep unfolding network. The optimal $\boldsymbol{C}$ of the denoiser can be worked out based on SURE framework. Here we treat T as the only learnable parameter and redenote it using α. Owing to the close correlation between T and τ_t^2, when we treat α (i.e. T) to be learnable, α itself can "carry" the information of noise variance. Thus we no longer need any noise variance estimator. A specific layer structure of SURE-TISTA is demonstrated in Fig. 1. Unfolding each iteration into DNN layers, we can find the layer-dependent parameter $\alpha_{optimal}$ of each layer.

4. NUMERICAL RESULTS

In this section, we provide numerical results on our work. For a Bernoulli-Gaussian sparse data $\boldsymbol{x} \in \mathbb{R}^N$, each entry is a realization drawn i.i.d from

$$p_{BG}(x) = (1-\lambda)\delta(x) + \lambda \mathcal{N}(x; 0, 1) \tag{21}$$

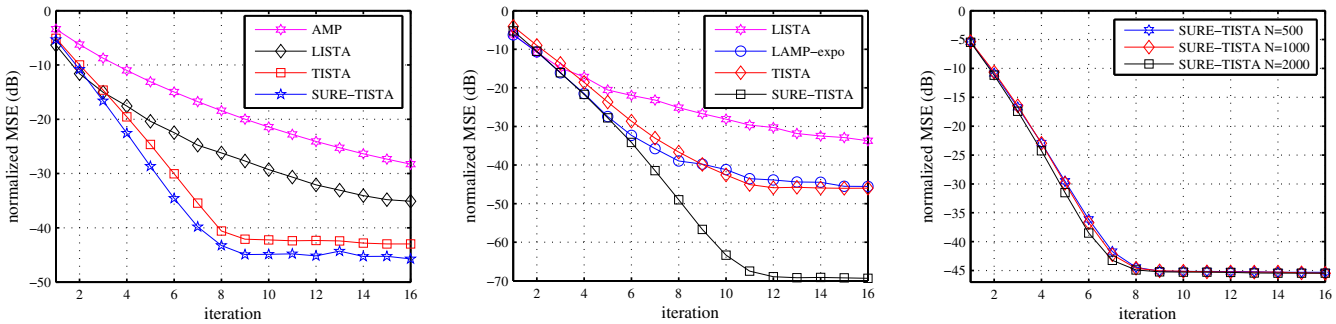
where λ decides the sparse level of x, we set $\lambda = 0.1$. SNR is set to be 40dB. The mini-batches are of size-1000 for training and testing respectively. Training and testing set are independent but from the same distribution. We implement our experiment over Tensorflow [15] with Adam optimizer [16]. SNR is defined as $SNR = E\{\|Ax\|^2\}/E\{\|n\|^2\}$.

4.1. Performance Comparison

We first consider the i.i.d Gaussian matrix $A_{i,j} \sim \mathcal{N}(0, M^{-1})$. Performance is described by the normalized MSE (NMSE), i.e.,

$$NMSE = \|\boldsymbol{s}_{t+1} - \boldsymbol{x}\|^2 / \|\boldsymbol{x}\|^2 \tag{22}$$

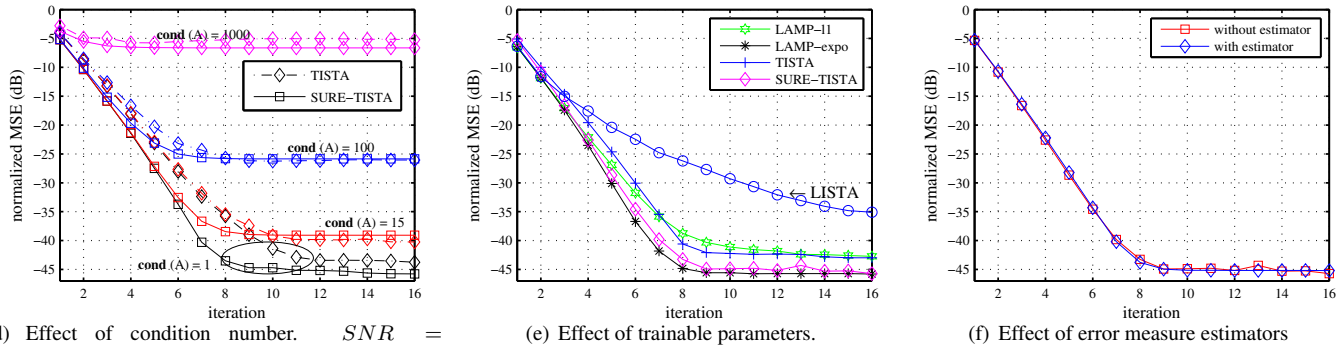
M and N are set to be 250 and 500 originally. Fig. 2(a) exhibits that SURE-TISTA has a much better curve than AMP and LISTA and achieves a better performance both in speed and convergence level than TISTA. We note that TISTA requires prior information, but SURE-TISTA can recover signals without a prior. Fig. 2(b) and Fig. 2(c) indicate a splendid adaptability of SURE-TISTA to large-scale and large-variance problems in practical environment.



(a) Comparison with AMP, LISTA and TISTA.

(b) Comparison with LISTA, LAMP-expo and TISTA in large variance. i.e. $\boldsymbol{A}_{i,j} \sim \mathcal{N}(0,1)$

(c) Effect of size of $\boldsymbol{A}$: $(250,500)(500,1000)(1000,2000)$



(d) Effect of condition number. $SNR = 40dB$.$\mathbf{cond}(A) = 1, 15, 100, 1000$.

(e) Effect of trainable parameters.

(f) Effect of error measure estimators

Fig. 2. Numerical results of SURE-TISTA. SNR=40dB.

Table 1. Numbers of trainable variables in each layer

LAMP-$\ell 1$	LAMP-expo	TISTA	SURE-TISTA
$NM+2$	$NM+3$	1	1

Then we consider $\boldsymbol{A}$ is ill-conditioned (Fig. 2(d)). Both SURE-TISTA and ISTA have degraded performance when condition number increases. In order to overcome the problem, we apply an improved method in [11]: define $\hat{\boldsymbol{W}} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T + \lambda \boldsymbol{E})^{-1}$, where λ is a real constant. Although brief explanation is provided in [11], we provide another explanation on this improvement: Since sparse inverse problems can be seen as convex optimization problems, we introduce an ℓ_2 regularization λ for a better performance of convergence, which is often considered in optimal solution with large condition numbers. λ can help changing the objective function into a strongly convex function.

4.2. Analysis of Trainable Parameters and Error Measure Estimators

Table 1 summarizes the numbers of trainable variables of LAMP-ℓ_1 [10], LAMP-expo (LAMP with an exponential shrinkage function [10]), TISTA and SURE-TISTA in one layer. Their NMSEs are demonstrated in Fig. 2(e) in the case of $A_{i,j} \sim \mathcal{N}(0, M^{-1})$. Although SURE-TISTA has fewer trainable variables than LAMP-expo, they have similar curves. Moreover, $(NM+3)$ variables of LAMP indicate its dependence on M and N, but the number of trainable variables of SURE-TISTA is only related to the number of layers, which reduces training complexity significantly.

We also adopt error measure estimators (5) and (6) to SURE-TISTA and let $T = \beta\sqrt{\hat{\tau}_t^2}$ with learnable variable β. Fig. 2(f) shows that performance of two methods are quite similar. In the other words, our novel network is successful in estimation of error measures.

5. CONCLUSION

In this paper, we propose a novel network SURE-TISTA to recover signal for compressed sensing. Based on SURE framework, SURE-TISTA outperforms TISTA and does not require prior information nor error measure estimators. SURE-TISTA also exhibits a great robustness in different setups, especially in the case of large variance. Using fewer learnable variables, SURE-TISTA has a better performance than some other algorithms. We believe SURE-TISTA can be adapted for image denoising and wireless communication systems, which will be one of our future works.

6. REFERENCES

[1] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, pp. 18914–18919, Nov. 2009.

[2] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure and Appl. Math.*, vol. 57, pp. 1413–1457, Nov. 2004.

[3] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[4] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020–2033, Jan. 2017.

[5] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comp. Int. Mag.*, vol. 13, pp. 55–75, 2018.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Inf. Process. Sys.*, vol. 2012, pp. 1097–1105, Dec. 2012.

[7] T. Wang, CK. Wen, H. Wang, F. Gao, T. Jiang, S. Jin, "Deep learning for wireless physical layer: opportunities and challenges," *China Commun.*, vol. 14, no.11, pp. 92–111, 2017.

[8] J. R. Hershey, J. Le Roux, and F. Weninger, "Deep unfolding: model-based inspiration of novel deep architectures," *Tech. Rep.* TR2014–117, Mitsubishi Electric Research Labs, 2014.

[9] K. Gregor and Y. Lecun, "Learning fast approximations of sparse coding," in *Proc. Int. Conf. Mach. Learning*, pp. 399–406, 2010.

[10] M. Borgerding, P. Schniter, and S. Rangan, "AMP-inspired deep networks for sparse linear inverse problems," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4293C-4308, 2017.

[11] D. Ito, S. Takabe, and T. Wadayama, "Trainable ISTA for sparse signal recovery," IEEE Int. Conf. of Commun., 2018.

[12] D. Guo, Y. Wu, S. S. (Shitz), and S. Verdu, "Estimation in gaussian noise: properties of the minimum mean-square error," *IEEE Trans. Inform. Theory*, vol. 57, no. 4, pp. 2371C-2385, 2011.

[13] C. Guo and M. E. Davies, "Near optimal compressed sensing without priors: parametric sure approximate message passing," *IEEE Trans. Signal Process.*, vol. 63, no. 8, pp. 2130C-2141, 2015.

[14] F. Luisier, T. Blu, and M. Unser, "A new SURE approach to image denoising: interscale orthonormal wavelet thresholding," *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 593-C606, Mar. 2007.

[15] M. Abadi, A. Agarwal, P. Barham, et al. (2015) TensorFlow: large-scale machine learning on heterogeneous systems. [Online]. Available: https://www.tensorflow.org.

[16] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proc. Internat. Conf. on Learning Repres.*, 2015.