MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Learning optimal nonlinearities for iterative thresholding algorithms

Kamilov, U. S.; Mansour, H.

TR2016-039 May 2016

Abstract

Iterative shrinkage/thresholding algorithm (ISTA) is a well-studied method for finding sparse solutions to illposed inverse problems. In this letter, we present a data-driven scheme for learning optimal thresholding functions for ISTA. The proposed scheme is obtained by relating iterations of ISTA to layers of a simple feedforward neural network and developing a corresponding error backpropagation algorithm for fine-tuning the thresholding functions. Simulations on sparse statistical signals illustrate potential gains in estimation quality due to the proposed data adaptive ISTA.

2016 IEEE Signal Processing Letters

Learning optimal nonlinearities for iterative thresholding algorithms

Ulugbek S. Kamilov, *Member, IEEE* and Hassan Mansour, *Member, IEEE*

***Abstract*—Iterative shrinkage/thresholding algorithm (ISTA) is a well-studied method for finding sparse solutions to ill-posed inverse problems. In this letter, we present a data-driven scheme for learning optimal thresholding functions for ISTA. The proposed scheme is obtained by relating iterations of ISTA to layers of a simple feedforward neural network and developing a corresponding error backpropagation algorithm for fine-tuning the thresholding functions. Simulations on sparse statistical signals illustrate potential gains in estimation quality due to the proposed data adaptive ISTA.**

***Index Terms*—Compressive sensing, sparse recovery, ISTA, neural networks, error backpropagation**

I. Introduction

THE problem of estimating an unknown signal from noisy linear observations is fundamental in signal processing. The estimation task is often formulated as the linear inverse problem

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{e}, \tag{1}$$

where the objective is to recover the unknown signal $\mathbf{x} \in \mathbb{R}^N$ from the noisy measurements $\mathbf{y} \in \mathbb{R}^M$. The matrix $\mathbf{H} \in \mathbb{R}^{M\times N}$ models the response of the acquisition device and the vector $\mathbf{e} \in \mathbb{R}^M$ represents the measurement noise, which is often assumed to be independent and identically distributed (i.i.d.) Gaussian.

A standard approach for solving ill-posed linear inverse problems is the regularized least-squares estimator

$$\widehat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_{\ell_2}^2 + \lambda\mathcal{R}(\mathbf{x}) \right\}, \tag{2}$$

where $\mathcal{R}$ is a regularizer that promotes solutions with desirable properties and $\lambda > 0$ is a parameter that controls the strength of regularization. In particular, sparsity-promoting regularization, such as ℓ_1-norm penalty $\mathcal{R}(\mathbf{x}) \triangleq \|\mathbf{x}\|_{\ell_1}$, has proved to be successful in a wide range of applications where signals are naturally sparse. Regularization with the ℓ_1-norm is an essential component of compressive sensing theory [1], [2], which establishes conditions for accurate estimation of the signal from $M < N$ measurements.

The minimization (2) with sparsity promoting penalty is a non-trivial optimization task. The challenging aspects are the non-smooth nature of the regularization term and the massive quantity of data that typically needs to be processed. Proximal gradient methods [3] such as iterative shrinkage/thresholding algorithm (ISTA) [4]–[6] or alternating direction method of multipliers (ADMM) [7], [8] are standard approaches to circumvent the non-smoothness of the regularizer while simplifying the optimization problem into a sequence of computationally efficient operations.

U. S. Kamilov (email: kamilov@merl.com) and H. Mansour (email: mansour@merl.com) are with Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA 02140, USA.

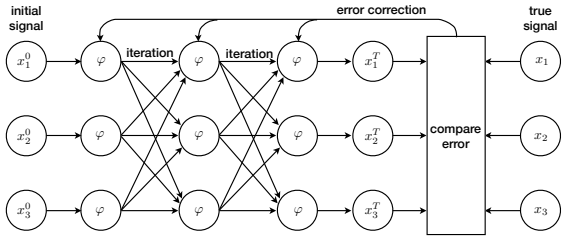

Fig. 1. Visual representation of the optimization scenario considered in this letter. ISTA with a pointwise nonlinearity φ is initialized with a signal $\mathbf{x}^0$ which results in the estimate $\mathbf{x}^T$ after T iterations. The algorithm proposed here allows to efficiently refine φ by comparing $\mathbf{x}^T$ against the true signal $\mathbf{x}$ from a set of training examples.

For the problem (2), ISTA can be written as

$$\mathbf{z}^t \leftarrow (\mathbf{I} - \gamma\mathbf{H}^{\mathrm{T}}\mathbf{H})\mathbf{x}^{t-1} + \gamma\mathbf{H}^{\mathrm{T}}\mathbf{y} \tag{3a}$$

$$\mathbf{x}^t \leftarrow \mathcal{T}(\mathbf{z}^t; \gamma\lambda), \tag{3b}$$

where $\mathbf{x}^t$ is the estimate at iteration t, $\mathbf{I}$ is the identity matrix, and $\gamma > 0$ is a step-size that can be set to $\gamma = 1/\lambda_{\max}(\mathbf{H}^{\mathrm{T}}\mathbf{H})$ to ensure convergence [9]. Here, the symbol $^{\mathrm{T}}$ denotes the matrix transpose operator, and $\lambda_{\max}(\mathbf{H}^{\mathrm{T}}\mathbf{H})$ denotes the largest eigenvalue of the matrix $\mathbf{H}^{\mathrm{T}}\mathbf{H}$. Iteration (3) combines the gradient descent step (3a) with a proximal operator (3b) that reduces to a pointwise nonlinearity

$$\mathcal{T}(z;\lambda) = \mathrm{prox}_{\gamma\mathcal{R}}(z) \tag{4a}$$

$$\triangleq \underset{x\in\mathbb{R}}{\arg\min} \left\{ \frac{1}{2}(x-z)^2 + \lambda\mathcal{R}(x) \right\}. \tag{4b}$$

for convex and separable regularizers such as the ℓ_1-norm penalty.

In this letter, we consider the problem of learning an optimal nonlinearity $\mathcal{T}$ for ISTA given a set of L training examples $\{\mathbf{x}_\ell, \mathbf{y}_\ell\}_{\ell\in[1,\ldots,L]}$. Specifically, as illustrated in Fig. 1, we interpret iteration (3) as a simple feedforward neural network [10] with T layers and develop an efficient algorithm that allows to determine optimal $\mathcal{T}$ directly from data. Simulations on sparse statistical signals show that data adaptive ISTA substantially improves over the ℓ_1-regularized reconstruction by approaching the performance of the minimum mean squared error (MMSE) estimator.

II. Related Work

Starting from the early works [4]–[6], iterative thresholding algorithms have received significant attention in the context of sparse signal estimation. Accelerated variants of ISTA were proposed by, among others, Bioucas-Dias and Figueiredo [11], and Beck and Teboulle [9]. Additional extensions were proposed by replacing soft-thresholding with alternative sparsity-promoting nonlinearities [12]–[16]. The method has also inspired the approximate message passing (AMP) algorithm by Donoho *et al.* [17], as well as its Bayesian extensions [18], [19]. In particular, it was shown that, in the compressive sensing setting, one can obtain an optimal estimation quality by adapting the thresholding function of AMP to the statistics of the signal [20], [21]. The primary difference of the work here, to the traditional approaches based on Bayesian signal modeling, is that the optimal thresholding functions are learned directly from independent realizations of the data, rather than being explicitly designed to the assumed statistics. Accordingly, the scheme presented here is particularly useful when the statistical distribution of the signals is not known.

More recently, several authors have considered relating iterative algorithms to neural networks. For example, in the context of sparse coding, Gregor and LeCun [22] proposed to accelerate ISTA by learning the matrix $\mathbf{H}$ from data. The idea was further refined by Sprechmann *et al.* [23] by considering an unsupervised learning approach and incorporating a structural sparsity model for the signal. In the context of the image deconvolution problem, Schmidt and Roth [24] proposed a scheme to jointly learn iteration dependent dictionaries and thresholds for ADMM. Similarly, Chen *et al.* [25] proposed to parametrize nonlinear diffusion models, which are related to the gradient descent method, and learned the parameters given a set of training images. A general application of deep learning to compressive sensing was presented in [26], while the idea of learning the activation functions was discussed in [27]. This letter extends those works by specifically learning separable thresholding functions for ISTA. Unlike the matrices $\mathbf{H}$, thresholding functions relate directly to the underlying statistical distributions of i.i.d. signals $\mathbf{x}$. Furthermore, by optimizing for the same nonlinearity across iterations, we obtain the MSE optimal ISTA for a specific statistical distribution of data, which, in turn, allows us to evaluate the best possible reconstruction achievable by ISTA.

III. Main Results

By defining a matrix $\mathbf{S} \triangleq \mathbf{I} - \gamma \mathbf{H}^{\mathrm{T}}\mathbf{H}$, vector $\mathbf{b} \triangleq \gamma \mathbf{H}^{\mathrm{T}}\mathbf{y}$, as well as nonlinearity $\varphi(\cdot) \triangleq \mathcal{T}(\cdot, \gamma\lambda)$, we can re-write ISTA using element-wise update steps as follows

$$z_m^t \leftarrow \sum_{n=1}^{N} S_{mn} x_n^{t-1} + b_m \tag{5a}$$

$$x_m^t \leftarrow \varphi(z_m^t), \tag{5b}$$

where $m \in [1, \ldots, N]$.

A. Problem Formulation

Our objective is now to design an efficient algorithm for adapting the function φ, given a set of L training examples $\{\mathbf{x}_\ell, \mathbf{y}_\ell\}_{\ell \in [1,\ldots,L]}$, as well as by assuming a fixed number of ISTA iterations T. In order to devise a computational approach for tuning φ, we adopt the following parametric representation for the nonlinearities

$$\varphi(z) \triangleq \sum_{k=-K}^{K} c_k \psi\left(\frac{z}{\Delta} - k\right), \tag{6}$$

where $\mathbf{c} \triangleq \{c_k\}_{k \in [-K,\ldots,K]}$, are the coefficients of the representation and ψ is a basis function, to be discussed shortly, positioned on the grid $\Delta[-K, -K+1, \ldots, K] \subseteq \Delta\mathbb{Z}$. Here, the constant $\Delta > 0$ denotes the distance between two grid points. We can reformulate the learning process in terms of coefficients $\mathbf{c}$ as follows

$$\widehat{\mathbf{c}} = \operatorname*{arg\,min}_{\mathbf{c} \in \mathcal{C}} \left\{ \frac{1}{L} \sum_{\ell=1}^{L} \mathcal{E}(\mathbf{c}, \mathbf{x}_\ell, \mathbf{y}_\ell) \right\} \tag{7}$$

where $\mathcal{C} \subseteq \mathbb{R}^{2K+1}$ is a set that incorporates prior constraints on the coefficients and $\mathcal{E}$ is a cost functional that guides the learning. The cost functional that interests us in this letter is the MSE defined as

$$\mathcal{E}(\mathbf{c}, \mathbf{x}, \mathbf{y}) \triangleq \frac{1}{2}\|\mathbf{x} - \mathbf{x}^T(\mathbf{c}, \mathbf{y})\|_{\ell_2}^2, \tag{8}$$

where $\mathbf{x}^T$ is the solution of ISTA at iteration T, which depends on both the coefficients $\mathbf{c}$ and the given data vector $\mathbf{y}$. Given a large number of independent and identically distributed realizations of the signals $\{\mathbf{x}_\ell, \mathbf{y}_\ell\}$, the empirical MSE is expected to approach the true MSE of ISTA for nonlinearities of type (6). Thus, by solving the minimization problem (7) with the cost (8), we are seeking the MMSE variant of ISTA for a given statistical distribution of the signal $\mathbf{x}$ and measurements $\mathbf{y}$.

B. Optimization

For notational simplicity, we now consider the scenario of a single training example and thus drop the indices ℓ from the subsequent derivations. The generalization of the final formula to an arbitrary number of training samples L is straightforward.

We would like to minimize the following cost

$$\mathcal{E}(\mathbf{c}) \triangleq \frac{1}{2}\|\mathbf{x} - \mathbf{x}^T(\mathbf{c})\|_{\ell_2}^2 = \frac{1}{2}\sum_{m=1}^{N}(x_m - x_m^T(\mathbf{c}))^2, \tag{9}$$

where we dropped the explicit dependence of $\mathbf{x}^T$ on $\mathbf{y}$ for notational convenience. The optimization of the coefficients is performed via the projected gradient iterations

$$\mathbf{c}^i = \mathrm{proj}_{\mathcal{C}}(\mathbf{c}^{i-1} - \mu \nabla \mathcal{E}(\mathbf{c}^{i-1})), \tag{10}$$

where $i = 1, 2, 3, \ldots,$ denotes the iteration number of the training process, $\mu > 0$ is the step-size, which is also called the learning rate, and $\mathrm{proj}_{\mathcal{C}}$ is an orthogonal projection operator onto the convex set $\mathcal{C}$.

We now devise an efficient error backpropagation algorithm for computing the derivatives of $\mathcal{E}$ with respect to coefficients

Algorithm 1 Backpropagation for evaluating $\nabla\mathcal{E}(\mathbf{c})$

input: measurements $\mathbf{y}$, signal $\mathbf{x}$, current value of coefficients $\mathbf{c}$, and number of ISTA iterations T.
output: the gradient $\nabla\mathcal{E}(\mathbf{c})$.
algorithm:

1) Run T iterations of ISTA in eq. (5) by storing intermediate variables $\{\mathbf{z}^t\}_{t\in[1,\dots,T]}$ and the final estimate $\mathbf{x}^T$.
2) *Initialize:* Set $t \leftarrow T$, $\mathbf{r}^T \leftarrow \mathbf{x}^T - \mathbf{x}$, and $\mathbf{g}^T \leftarrow 0$.
3) *Compute:*
$$\mathbf{g}^{t-1} \leftarrow \mathbf{g}^t + [\mathbf{\Psi}^t]^{\mathrm{T}}\mathbf{r}^t \tag{11a}$$
$$\mathbf{r}^{t-1} \leftarrow \mathbf{S}^{\mathrm{T}}\mathrm{diag}(\varphi'(\mathbf{z}^t))\mathbf{r}^t \tag{11b}$$
4) If $t = 0$, return $\nabla\mathcal{E}(\mathbf{c}) = \mathbf{g}^0$, otherwise, set $t \leftarrow t-1$ and proceed to step 3).

Algorithm 2 Online learning for solving (7)

input: set of L training examples $\{\mathbf{x}_\ell, \mathbf{y}_\ell\}_{\ell\in[1,\dots,L]}$, learning rate $\mu > 0$, and constraint set $\mathcal{C} \subseteq \mathbb{R}^{2K+1}$.
output: nonlinearity φ specified by learned coefficients $\widehat{\mathbf{c}}$.
algorithm:

1) *Initialize:* Set $i \leftarrow 1$ and select $\mathbf{c}^0 \in \mathcal{C}$.
2) Select a small subset $\{\mathbf{x}_\ell\}$ and $\{\mathbf{y}_\ell\}$ with an equal probability, from the set of L training examples.
3) Using the selected training examples, update φ as follows
$$\mathbf{c}^i \leftarrow \mathrm{proj}_{\mathcal{C}}(\mathbf{c}^{i-1} - \mu\nabla\mathcal{E}(\mathbf{c}^{i-1})) \tag{19}$$
4) Return $\widehat{\mathbf{c}} = \mathbf{c}^i$ if a stopping criterion is met, otherwise set $i \leftarrow i+1$ and proceed to step 2).

$\mathbf{c}$. First, note that we can write the iteration (5) with the nonlinearity (6) as follows

$$x_m^t = \varphi(z_m^t) = \sum_{k=-K}^{K} c_k \psi\left(\frac{z_m^t}{\Delta} - k\right), \tag{12}$$

for all $m \in [1,\dots,N]$. The gradient can be obtained by evaluating

$$\nabla\mathcal{E}(\mathbf{c}) = \left[\frac{\partial}{\partial\mathbf{c}}\mathbf{x}^T(\mathbf{c})\right]^{\mathrm{T}}(\mathbf{x}^T(\mathbf{c}) - \mathbf{x}), \tag{13}$$

where we define the Jacobian

$$\frac{\partial}{\partial\mathbf{c}}\mathbf{x}^t(\mathbf{c}) \triangleq \begin{bmatrix} \frac{\partial x_1^t}{\partial c_{-K}} & \cdots & \frac{\partial x_1^t}{\partial c_K} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_N^t}{\partial c_{-K}} & \cdots & \frac{\partial x_N^t}{\partial c_K} \end{bmatrix}. \tag{14}$$

By differentiating (12) with respect to c_k and simplifying the resulting expression, we obtain

$$\frac{\partial x_m^t}{\partial c_k} = \Psi_{mk}^t + \varphi'(z_m^t)\sum_{n=1}^{N} S_{mn}\left[\frac{\partial x_n^{t-1}}{\partial c_k}\right], \tag{15}$$

where we defined a matrix $\Psi_{mk}^t \triangleq \psi(z_m^t/\Delta - k)$, and φ' is the derivative of φ with respect to z_m^t. Then, for any vector $\mathbf{r} \in \mathbb{R}^N$, we obtain

$$\sum_{m=1}^{N}\left[\frac{\partial x_m^t}{\partial c_k}\right] r_m = \sum_{m=1}^{N}\Psi_{mk}^t r_m + \sum_{n=1}^{N}\left[\frac{\partial x_n^{t-1}}{\partial c_k}\right]\sum_{m=1}^{N} S_{mn} r_m \varphi'(z_m^t), \tag{16}$$

which translates to the following vector equation

$$\left[\frac{\partial\mathbf{x}^t}{\partial\mathbf{c}}\right]^{\mathrm{T}}\mathbf{r} = [\mathbf{\Psi}^t]^{\mathrm{T}}\mathbf{r} + \left[\frac{\partial\mathbf{x}^{t-1}}{\partial\mathbf{c}}\right]^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathrm{diag}(\varphi'(\mathbf{z}^t))\mathbf{r}, \tag{17}$$

where the operator $\mathrm{diag}(\mathbf{g})$ creates a matrix and places the vector $\mathbf{g}$ into its main diagonal. Note that since the initial estimate $\mathbf{x}^0$ does not depend on $\mathbf{c}$, we have that

$$\frac{\partial\mathbf{x}^0}{\partial\mathbf{c}} = 0. \tag{18}$$

By applying the equation (17) recursively starting from $t = T$ and using (18), we obtain

$$\begin{aligned}
\left[\frac{\partial\mathbf{x}^T}{\partial\mathbf{c}}\right]^{\mathrm{T}}\mathbf{r}^T &= \underbrace{[\mathbf{\Psi}^T]^{\mathrm{T}}\mathbf{r}^T}_{\triangleq\,\mathbf{g}^{T-1}} + \left[\frac{\partial\mathbf{x}^{T-1}}{\partial\mathbf{c}}\right]^{\mathrm{T}}\underbrace{\mathbf{S}^{\mathrm{T}}\mathrm{diag}(\varphi'(\mathbf{z}^T))\mathbf{r}^T}_{\triangleq\,\mathbf{r}^{T-1}} \\
&= \mathbf{g}^{T-1} + \left[\frac{\partial\mathbf{x}^{T-1}}{\partial\mathbf{c}}\right]^{\mathrm{T}}\mathbf{r}^{T-1} = \underbrace{\mathbf{g}^{T-1} + [\mathbf{\Psi}^{T-1}]^{\mathrm{T}}\mathbf{r}^{T-1}}_{\triangleq\,\mathbf{g}^{T-2}} \\
&\quad + \left[\frac{\partial\mathbf{x}^{T-2}}{\partial\mathbf{c}}\right]^{\mathrm{T}}\underbrace{\mathbf{S}^{\mathrm{T}}\mathrm{diag}(\varphi'(\mathbf{z}^{T-1}))\mathbf{r}^{T-1}}_{\triangleq\,\mathbf{r}^{T-2}} \\
&= \mathbf{g}^{T-2} + \left[\frac{\partial\mathbf{x}^{T-2}}{\partial\mathbf{c}}\right]^{\mathrm{T}}\mathbf{r}^{T-2} = \cdots = \mathbf{g}^0 + \left[\frac{\partial\mathbf{x}^0}{\partial\mathbf{c}}\right]^{\mathrm{T}}\mathbf{r}^0 = \mathbf{g}^0.
\end{aligned}$$

This suggests the error backpropagation algorithm summarized in Algorithm 1 that allows one to obtain (13).

The remarkable feature of Algorithm 1 is that it allows one to efficiently evaluate the gradient of ISTA with respect to the nonlinearity φ. Its computational complexity is equivalent to running a single instance of ISTA, which is a first-order method, known to be scalable to very large scale inverse problems. Finally, equipped with Algorithm 1, nonlinearity φ can easily be optimized by using an online learning approach [28] summarized in Algorithm 2.

C. Representation with B-Splines

In our implementation, we represent the nonlinearity φ in terms of its expansion with polynomial B-Splines (see an extensive review of B-Spline interpolation by Unser [29], [30]). The main advantage of the B-Spline representation is that it can approximate any nonlinearity with an arbitrary precision for a sufficiently small Δ. Accordingly, our basis function corresponds to $\psi = \beta^d$, where β^d refers to a B-Spline of degree $d \geq 0$. Within the family of polynomial splines, cubic B-Splines

$$\beta^3(z) = \begin{cases} \frac{2}{3} - |z|^2 + \frac{|z|^3}{2} & \text{when } 0 \leq |z| \leq 1 \\ \frac{1}{6}(2-|z|)^3 & \text{when } 1 \leq |z| \leq 2 \\ 0 & \text{when } 2 \leq |z|, \end{cases} \tag{20}$$

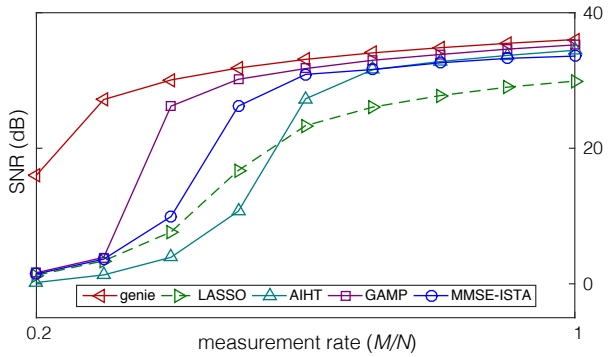

Fig. 2. Quantitative evaluation on sparse signals. Average SNR is plotted against the measurement rate M/N when recovering $N = 512$ Bernoulli-Gaussian signal $\mathbf{x}$ from measurements $\mathbf{y}$ under i.i.d. $\mathbf{H}$.

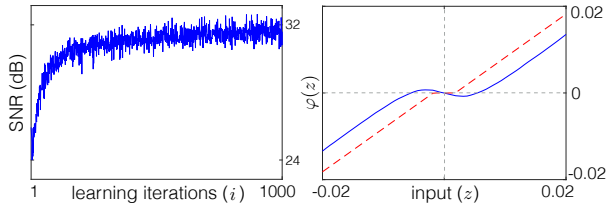

Fig. 3. Illustration of the learning process for $M/N = 0.7$. Left: SNR of training is plotted for each training iteration. Right: the final learned shrinkage (solid) is compared to the standard soft-thresholding under optimal λ.

tend to be the most popular in applications—perhaps due to their minimum curvature property [29]. B-Splines are very easy to manipulate. For instance, their derivatives are computed through the following formula

$$\frac{\mathrm{d}}{\mathrm{d}z}\beta^d(z) = \beta^{d-1}(z+\tfrac{1}{2}) - \beta^{d-1}(z-\tfrac{1}{2}), \qquad (21)$$

which simply reduces the degree by one. By applying this formula to the expansion of φ, we can easily obtain a closed form expression for φ' in terms of quadratic B-Splines

$$\beta^2(z) = \begin{cases} \frac{3}{4} - |z|^2 & \text{when } 0 \le |z| \le \frac{1}{2} \\ \frac{9}{8} - \frac{1}{2}|z|(3-|z|) & \text{when } \frac{1}{2} \le |z| \le \frac{3}{2} \\ 0 & \text{when } \frac{3}{2} \le |z|. \end{cases} \qquad (22)$$

IV. Experiments

To verify our learning scheme, we report results of ISTA with learned MSE optimal nonlinearities (denoted *MMSE-ISTA*) on the compressive sensing recovery problem. In particular, we consider the estimation of sparse Bernoulli-Gaussian signals $\mathbf{x}$ with an i.i.d. prior $p_x(x_n) = \rho\mathcal{N}(x_n, 0, 1) + (1-\rho)\delta(x_n)$, where $\rho \in (0, 1]$ is the sparsity ratio, $\mathcal{N}(\cdot, \mu, \sigma^2)$ is the Gaussian probability distribution function of mean μ and variance σ^2, and δ is the Dirac delta distribution. In our experiments, we fix the parameters to $N = 512$ and $\rho = 0.2$, and we numerically compare the signal-to-noise ratio (SNR) defined as SNR (dB) $\triangleq 10\log_{10}\left(\|\mathbf{x}\|_{\ell_2}^2/\|\mathbf{x}-\widehat{\mathbf{x}}\|_{\ell_2}^2\right)$, for the estimation of $\mathbf{x}$ from linear measurements of form (1), where $\mathbf{e}$ has variance set to achieve SNR of 30 dB, and where the measurement matrix $\mathbf{H}$ is drawn with i.i.d. $\mathcal{N}(0, 1/M)$ entries.

We compare results of MMSE-ISTA against four alternative methods. As the first reference method, we consider standard least absolute shrinkage and selection operator (*LASSO*) [31] estimator, which corresponds to solving (2) with an ℓ_1-norm regularizer. In addition to LASSO, we consider the accelerated iterative hard thresholding (*AIHT*) algorithm [14] that seeks minimum ℓ_0-norm solution to the linear inverse problem. We also consider the MMSE variant of the generalized AMP (*GAMP*) algorithm [19], which is known to be nearly optimal for recovery of sparse signals from random measurements. Finally, we consider a support-aware MMSE estimator (*genie*), which provides an upper bound on the reconstruction performance of any algorithm.

The regularization parameter λ of LASSO was optimized for the best SNR performance. Similarly, the parameters of AIHT and GAMP were set to their statistically optimal values. The implementation of LASSO is based on FISTA [9]. FISTA, AIHT, and GAMP were run for a maximum of 1000 iterations or until convergence that was measured using the relative change in the solution in two successive iterations $\|\mathbf{x}^t - \mathbf{x}^{t-1}\|_{\ell_2}/\|\mathbf{x}^{t-1}\|_{\ell_2} \le 10^{-4}$. The number of layers of MMSE-ISTA was set to $T = 200$. Learning was performed by using online learning in Algorithm 2 that was run for 1000 iterations with $\mu = 10^{-4}$. The nonlinearity φ was defined with 8000 basis functions that were spread uniformly over the dynamic range of the signal and was initialized to correspond to the soft-thresholding function with optimal λ.

Figure 2 reports the SNR performance of all algorithms under test after averaging the results of 1000 random trials. The results show that the quality of the estimated signal can be considerably boosted using the learnt nonlinearities φ that are adapted to the training data. In particular, the SNR performance of MMSE-ISTA is significantly better than that of LASSO and AIHT at the lower values of M/N, and is about 1 dB away from the SNR obtained by GAMP at higher values of M/N. Note that while the SNR performance of MMSE-ISTA is slightly inferior to that of GAMP, the former does not require randomness of $\mathbf{H}$ [32] and avoids any explicit assumptions on the statistics of the i.i.d. signal $\mathbf{x}$. Figure 3 illustrates the per-iteration evolution of SNR evaluated on the training sample during the learning process (left), as well as the final shape of the learned nonlinearity (right). Although the nonlinearity φ is initialized with the soft-thresholding function, the plots demonstrate that the learning procedure deviates the shape of φ from the soft-thresholding function, which leads to a significant increase in the SNR of the solution.

V. Conclusion

The scheme developed in this letter is useful for optimizing the nonlinearities of ISTA given a set of independent realizations of data samples. By using this scheme, we were able to benchmark the best possible reconstruction achievable by ISTA for i.i.d. sparse signals. Specifically, in the context of compressive sensing, we showed that by optimizing the nonlinearities, the performance of ISTA improves by several dBs and approaches that of the optimal estimator.

REFERENCES

[1] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, February 2006.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[3] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2010.

[4] M. A. T. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, August 2003.

[5] J. Bect, L. Blanc-Feraud, G. Aubert, and A. Chambolle, "A ℓ_1-unified variational framework for image restoration," in *Proc. ECCV*, Springer, Ed., vol. 3024, New York, 2004, pp. 1–13.

[6] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, November 2004.

[7] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, pp. 293–318, 1992.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[9] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[10] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, 1995.

[11] J. M. Bioucas-Dias and M. A. T. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2992–3004, December 2007.

[12] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, pp. 1348–1360, 2001.

[13] T. Blumensath and M. Davies, "Iterative thresholding for sparse approximation," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 629–654, 2008.

[14] T. Blumensath, "Accelerated iterative hard thresholding," *Signal Process.*, vol. 92, no. 3, pp. 752–756, 2012.

[15] J. Zeng, S. Lin, Y. Wang, and Z. Xu, "$L_{1/2}$ regularization: Convergence of iterative half thresholding algorithm," *IEEE Trans. Signal Process.*, vol. 62, no. 9, pp. 2317–2329, 2014.

[16] Y. Wang, J. Zeng, X. Peng, Z. andChang, and Z. Xu, "Linear convergence of adadaptively iterative threhsolding algorithm for compressed sensing," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2957–2971, 2015.

[17] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, November 2009.

[18] ——, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *IEEE Inf. Theory Workshop*, Dublin, Ireland, January 6-8, 2010, pp. 1–5.

[19] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Information Theory*, St. Petersburg, Russia, July 31-August 5, 2011, pp. 2168–2172.

[20] U. S. Kamilov, S. Rangan, A. K. Fletcher, and M. Unser, "Approximate message passing with consistent parameter estimation and applications to sparse learning," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2969–2985, May 2014.

[21] U. S. Kamilov, "Sparsity-driven statistical inference for inverse problems," EPFL Thesis no. 6545 (2015), 198 p., Swiss Federal Institute of Technology Lausanne (EPFL), March 27, 2015.

[22] K. Gregor and Y. LeCun, "Learning fast approximation of sparse coding," in *Proc. 27th Int. Conf. Machine Learning (ICML)*, Haifa, Israel, June 21-24, 2010, pp. 399–406.

[23] P. Sprechmann, P. Bronstein, and G. Sapiro, "Learning efficient structured sparse models," in *Proc. 29th Int. Conf. Machine Learning (ICML)*, Edinburgh, Scotland, June 26-July 1, 2012, pp. 615–622.

[24] U. Schmidt and S. Roth, "Shrinkage fields for effective image restoration," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, June 23-28, 2014, pp. 2774–2781.

[25] Y. Chen, W. Yu, and T. Pock, "On learning optimized reaction diffuction processes for effective image restoration," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 8-10, 2015, pp. 5261–5269.

[26] A. Mousavi, A. B. Patel, and R. G. Baraniuk, "A deep learning approach to structured signal recovery," pp. 1–8, August 2015, arXiv:1508.04065 [cs.LG].

[27] F. Agostinelli, M. Hoffman, P. Sadowski, and P. Baldi, "Learning activation functions to improve deep neural networks," pp. 1–9, December 2014, arXiv:1412.6830 [cs.NE].

[28] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proc. 20th Int. Conf. Machine Learning (ICML)*, Washington DC, USA, August 21-24, 2003.

[29] M. Unser, "Splines: A perfect fit for signal and image processing," *IEEE Signal Process. Mag.*, vol. 16, no. 6, pp. 22–38, November 1999.

[30] ——, "Sampling - 50 Years After Shannon," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 569–587, 2000.

[31] R. Tibshirani, "Regression and selection via the lasso," *J. R. Stat. Soc. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[32] S. Rangan, A. K. Fletcher, P. Schniter, and U. S. Kamilov, "Inference for generalized linear models via alternating directions and Bethe free energy minimization," in *Proc. IEEE Int. Symp. Information Theory*, Hong Kong, June 14-19, 2015, pp. 1640–1644.