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

Learning MMSE Optimal Thresholds for FISTA

Kamilov, U.; Mansour, H.

TR2016-111 August 2016

Abstract

Fast iterative shrinkage/thresholding algorithm (FISTA) is one of the most commonly used methods for solving linear inverse problems. In this work, we present a scheme that enables learning of optimal thresholding functions for FISTA from a set of training data. In particular, by relating iterations of FISTA to a deep neural network (DNN), we use the error backpropagation algorithm to find thresholding functions that minimize mean squared error (MSE) of the reconstruction for a given statistical distribution of data. Accordingly, the scheme can be used to computationally obtain MSE optimal variant of FISTA for performing statistical estimation.

International Traveling Workshop on Interactions Between Sparse Models and Technology (iTWIST)

Learning MMSE Optimal Thresholds for FISTA

Ulugbek S. Kamilov and Hassan Mansour.
Mitsubishi Electric Research Laboratories (MERL)
201 Broadway, Cambridge, MA 02139, USA
email: {kamilov, mansour}@merl.com.

***Abstract*— Fast iterative shrinkage/thresholding algorithm (FISTA) is one of the most commonly used methods for solving linear inverse problems. In this work, we present a scheme that enables learning of optimal thresholding functions for FISTA from a set of training data. In particular, by relating iterations of FISTA to a deep neural network (DNN), we use the error backpropagation algorithm to find thresholding functions that minimize mean squared error (MSE) of the reconstruction for a given statistical distribution of data. Accordingly, the scheme can be used to computationally obtain MSE optimal variant of FISTA for performing statistical estimation.**

1 Introduction

We consider a linear inverse problem $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{e}$, where the goal is to recover the unknown signal $\mathbf{x} \in \mathbb{R}^N$ from the noisy measurements $\mathbf{y} \in \mathbb{R}^M$. The matrix $\mathbf{H} \in \mathbb{R}^{M \times N}$ is known and models the response of the acquisition device, while the vector $\mathbf{e} \in \mathbb{R}^M$ represents unknown errors in the measurements.

Many practical inverse problems are ill-posed, which means that measurements $\mathbf{y}$ cannot explain the signal $\mathbf{x}$ uniquely. One standard approach for solving such problems is the regularized least-squares estimator

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

where $\mathcal{R}$ is a regularizer that imposes prior constraints in order to promote more meaningful solutions.

Two common approaches for solving the optimization problem (1) is the *iterative shrinkage/thresholding algorithm (ISTA)* [1–3] and its accelerated variant called *fast ISTA (FISTA)* [4]. Both algorithms can be expressed as

$$\mathbf{s}^t \leftarrow \mathbf{x}^{t-1} + ((1 - q_{t-1})/q_t)\left(\mathbf{x}^{t-1} - \mathbf{x}^{t-2}\right) \tag{2a}$$

$$\mathbf{x}^t \leftarrow \mathrm{prox}_{\gamma\mathcal{R}}\left(\mathbf{s}^t - \gamma\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{s}^t - \mathbf{y})\right), \tag{2b}$$

with the initial condition $\mathbf{x}^0 = \mathbf{x}^{-1} = \mathbf{x}_{\text{init}} \in \mathbb{R}^N$. The parameter $\gamma > 0$ is a step-size that is often set to $\gamma = 1/L$ with $L \triangleq \lambda_{\max}(\mathbf{H}^{\mathrm{T}}\mathbf{H})$ to ensure convergence and parameters $\{q_t\}_{t\in[0,1,\dots]}$ are called relaxation parameters [4]. For a fixed $q_t = 1$, iteration (2) corresponds to ISTA, which has $O(1/t)$ global rate of convergence; however, for an appropriate selection of $\{q_t\}_{t\in[0,1,\dots]}$ as in [4] one obtains FISTA, which has a faster $O(1/t^2)$ convergence rate. When the regularizer $\mathcal{R}$ is separable and acts in an identical manner in every data dimension, the proximal operator in (2b) reduces to a scalar nonlinearity

$$\mathcal{T}_\gamma(z) = \mathrm{prox}_{\gamma\mathcal{R}}(z) \tag{3a}$$

$$\triangleq \underset{x\in\mathbb{R}}{\arg\min} \left\{ \frac{1}{2}(x - z)^2 + \gamma\mathcal{R}(x) \right\}, \tag{3b}$$

applied individual to each component of the input vector.

Traditionally, the regularizer $\mathcal{R}$ and the corresponding proximal operator (3) are manually designed to preserve or promote certain properties in the solution. For example, ℓ_1-norm penalty $\mathcal{R}(\mathbf{x}) \triangleq \|\mathbf{x}\|_{\ell_1}$ is known to promote sparse solutions in (1), and has proved to be successful in a wide range of applications where signals are naturally sparse [5, 6]. One popular approach for designing regularizers comes from Bayesian theory, where $\mathcal{R}$ is selected according to the prior statistical distribution $p_{\mathbf{x}}$ of $\mathbf{x}$ as $\mathcal{R}(\mathbf{x}) = -\log(p_{\mathbf{x}}(\mathbf{x}))$, with the resulting estimator called the *maximum a posteriori probability (MAP)* estimator. From this statistical perspective, ℓ_1-norm penalty is often interpreted as a MAP estimator corresponding to the Laplace distribution. However, it has been shown that the MAP-based approach for designing proximals is suboptimal due to surprisingly poor performance of the resulting estimators in terms of mean squared error (MSE) [7, 8]. On the other hand, recent results have also showed that minimum MSE (MMSE) statistical estimator can also be expressed as a solution of (1), where $\mathcal{R}$ does not necessarily correspond to the negative logarithm of $p_{\mathbf{x}}$ [9–11].

In this work, we propose a data-driven scheme for computationally learning MSE optimal nonlinearity $\mathcal{T}$ for FISTA from a set of L training examples of true signals $\{\mathbf{x}_\ell\}_{\ell\in[1,\dots,L]}$ and measurements $\{\mathbf{y}_\ell\}_{\ell\in[1,\dots,L]}$. Specifically, we interpret iterations of FISTA as layers of a simple deep neural network (DNN) [12] and develop an efficient error backpropagation algorithm that allows to recover optimal $\mathcal{T}$ directly from data. Thus, for a large number of independent and identically distributed (i.i.d.) realizations of $\{\mathbf{x}_\ell, \mathbf{y}_\ell\}_{\ell\in[1,\dots,L]}$, the trained algorithm can be interpreted as MMSE variant of FISTA for a given statistical distribution of the signal and measurements.

Several other works have considered relating iterative algorithms to deep neural networks. For example, the learning scheme presented here extends the one in our recent paper [13] to FISTA, and thus improves the convergence properties of the trained algorithm. In the context of sparse coding, Gregor and LeCun [14] proposed to accelerate ISTA by learning the matrix $\mathbf{H}$ from data. The idea was further refined by Sprechmann *et al.* [15] who considered an unsupervised learning approach and incorporated a structural sparsity model for the signal. In the context of the image deconvolution problem, Schmidt and Roth [16] proposed a scheme to jointly learn iteration dependent dictionaries and thresholds for ADMM. Similarly, Chen *et al.* [17] proposed to parametrize nonlinear diffusion models, which are related to the gradient descent method, and learned the parameters given a set of training images. One distinction of our work is that we optimize for the same nonlinearity across iterations, which in turn allows us to interpret the algorithm as the MSE optimal FISTA for a given distribution of data.

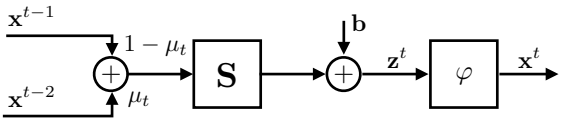

Figure 1: Visual representation of a single layer of the feedforward neural network, which also corresponds to a single iteration of FISTA. We use such layered representation of FISTA to obtain an error backpropagation algorithm for optimizing the scalar nonlinearity φ by comparing the outputs $\mathbf{x}^T$ after T iterations against the true signal $\mathbf{x}$ from a set of training examples.

2 Main Results

By define a matrix $\mathbf{S} \triangleq \mathbf{I} - \gamma \mathbf{H}^{\mathrm{T}}\mathbf{H}$, a vector $\mathbf{b} \triangleq \gamma \mathbf{H}^{\mathrm{T}}$, parameters $\mu_t \triangleq (1-q_{t-1})/q_t$, as well as nonlinearity $\varphi(\cdot) \triangleq \mathcal{T}_\gamma(\cdot)$, we can re-write FISTA as follows

$$\mathbf{z}^t \leftarrow \mathbf{S}\left((1-\mu_t)\mathbf{x}^{t-1} + \mu_t \mathbf{x}^{t-2}\right) + \mathbf{b} \tag{4a}$$
$$\mathbf{x}^t \leftarrow \varphi(\mathbf{z}^t). \tag{4b}$$

Fig. 1 visually represents a single iteration of (4), and by stacking several of such iterations one can represent (4) as a feedforward neural network (see also [13]), whose adaptable parameters correspond to the nonlinearity φ. Our objective is then to design an efficient algorithm for adapting the function φ, given a set of L training examples $\{\mathbf{x}_\ell, \mathbf{y}_\ell\}_{\ell\in[1,\dots,L]}$, as well as by assuming a fixed number of FISTA iterations T. In order to devise a computational approach for tuning φ, we adopt the following parametric representation for nonlinearities

$$\varphi(z) \triangleq \sum_{k=-K}^{K} c_k \phi\left(\frac{z}{\Delta} - k\right), \tag{5}$$

where $\mathbf{c} \triangleq \{c_k\}_{k\in[-K,\dots,K]}$ are the coeficients of the representation, ϕ are the basis functions positioned on the grid $\Delta[-K, -K+1, \dots, K] \subseteq \Delta\mathbb{Z}$. We can formulate the learning process in terms of coefficients $\mathbf{c}$ as follows

$$\widehat{\mathbf{c}} = \arg\min_{\mathbf{c}\in\mathcal{C}} \left\{ \frac{1}{L}\sum_{\ell=1}^{L} \mathcal{E}_\ell(\mathbf{c}) \right\}, \tag{6}$$

where $\mathcal{C} \subseteq \mathbb{R}^{2K+1}$ is a set that incorporates prior constraints on the coefficients such as symmetry, monotonicity, and nonnegativity on $\mathbb{R}_+$ [18, 19], and $\mathcal{E}$ is a cost functional that guides the learning. The cost functional that interests us in this work is the MSE defined as

$$\mathcal{E}_\ell(\mathbf{c}) \triangleq \frac{1}{2}\|\mathbf{x}_\ell - \mathbf{x}^T(\mathbf{c}, \mathbf{y}_\ell)\|_{\ell_2}^2, \tag{7}$$

where $\mathbf{x}^T$ is the solution of FISTA at iteration T, which depends on both coefficients $\mathbf{c}$ and the given data vector $\mathbf{y}_\ell$. Given a large number of i.i.d. realizationf of the signals $\{\mathbf{x}_\ell, \mathbf{y}_\ell\}$, the empirical MSE is expected to approach the true MSE of FISTA for nonlinearities of type (5).

We perform optimization of the coefficients $\mathbf{c}$ in an online fashion with projected gradient iterations

$$\mathbf{c}^i \leftarrow \mathrm{proj}_{\mathcal{C}}(\mathbf{c}^{i-1} - \alpha \nabla \mathcal{E}_\ell(\mathbf{c}^{i-1})), \tag{8}$$

where $i = 1, 2, 3, \dots,$ denotes the iteration number of the training process, $\alpha > 0$ is the learning rate, and $\mathrm{proj}_{\mathcal{C}}$ is an orthogonal projection operator on the set $\mathcal{C}$. Note that at each iteration i, we select a training pair $(\mathbf{x}_\ell, \mathbf{y}_\ell)$ uniformly at random. By

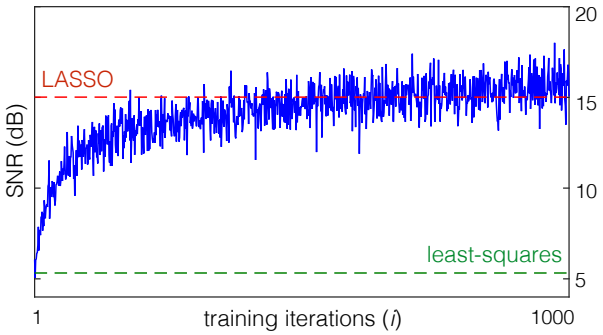

Figure 2: Illustration of the learning process for sparse image deconvolution problem. Top: SNR of training is plotted for each training iteration. Bottom: Top 8×8 pixels of (a) original image; (b) blurry and noisy (SNR = 0.86 dB); (c) LASSO (SNR = 13.36 dB); (d) Proposed (SNR = 14.48 dB).

defining $\Phi^t_{mk} = \phi(z^t_m/\Delta - k)$, the gradient $\nabla\mathcal{E}_\ell$ can be computed using the following error backpropagation algorithm for $t = T, T-1, \dots, 2,$

$$\mathbf{g}^{t-1} \leftarrow \mathbf{g}^t + [\mathbf{\Phi}^t]^{\mathrm{T}}\mathbf{r}_1^t \tag{9a}$$
$$\mathbf{r}^{t-1} \leftarrow [\mathbf{S}^{\mathrm{T}}\mathrm{diag}(\varphi'(\mathbf{z}^t))]\mathbf{r}_1^t \tag{9b}$$
$$\mathbf{r}_1^{t-1} \leftarrow \mathbf{r}_2^t + (1-\mu_t)\mathbf{r}^{t-1} \tag{9c}$$
$$\mathbf{r}_2^{t-1} \leftarrow \mu_t \mathbf{r}^{t-1}, \tag{9d}$$

where $\mathbf{g}^T = 0$, $\mathbf{r}_1^T = \mathbf{r}^T = \mathbf{x}^T(\mathbf{c}, \mathbf{y}_\ell) - \mathbf{x}_\ell$, and $\mathbf{r}_2^T = 0$. Finally, we return $\nabla\mathcal{E}_\ell(\mathbf{c}) = \mathbf{g}^1 + [\mathbf{\Phi}^1]^{\mathrm{T}}\mathbf{r}_1^1$. Note that (9) is backward compatible with the scheme in [13]; in particular, when $\mu_t = 0$ for all t, we recover the error backpropagation algorithm for the standard ISTA.

In Fig. 2, we illustrate results of a simple image deblurring problem, where a 3×3 Gaussian blur of variance 2 was applied to a 32×32 Bernoulli-Gaussian (BG) image with sparsity ratio 0.2. The mean and variance of the Gaussian component of BG were set 0 and 1, respectively. The blurry image was further contaminated with additive white Gaussian noise (AWGN) corresponding to 20 dB SNR. We plot per-training-iteration SNR of the reconstruction where training samples were generated in i.i.d. fashion. In all cases, FISTA was initialized with zero and run for 100 iterations. The nonlinearity φ was represented with 201 B-Spline basis functions on the interval $[-6, 6]$, and initialized with an identity operator, which means that initially the algorithm acted like a simple least-squares estimator. The plot illustrates that the learning procedure in (9) deviates the shape of φ from identity, which leads to a significant increase in the SNR of the solution, which eventually surpasses that of ℓ_1-based FISTA estimator denoted with LASSO. In the bottom of Fig. 2, we give an example reconstructed images by showing its top 8×8 corner.

To conclude, we proposed a scheme, summarized in eq. (9), to computationally learn shrinkage functions for FISTA. By using this scheme, it is possible to benchmark the best possible reconstruction achievable by FISTA in terms of MSE and for i.i.d. signals. Since the shrinkage functions are kept constant across the layers of our network, the number of parameters the algorithms needs to learn is small, which means that the scheme can be implemented on a simple desktop machine without extensive computations.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] R. Gribonval, V. Cevher, and M. E. Davies, "Compressible distributions for high-dimensional statistics," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5016–5034, August 2012.

[8] U. S. Kamilov, P. Pad, A. Amini, and M. Unser, "MMSE estimation of sparse Lévy processes," *IEEE Trans. Signal Process.*, vol. 61, no. 1, pp. 137–147, January 2013.

[9] R. Gribonval, "Should penalized least squares regression be interpreted as maximum a posteriori estimation?" *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2405–2410, May 2011.

[10] A. Kazerouni, U. S. Kamilov, E. Bostan, and M. Unser, "Bayesian denoising: From MAP to MMSE using consistent cycle spinning," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 249–252, March 2013.

[11] R. Gribonval and P. Machart, "Reconciling "priors" & "priors" without prejudice?" in *Proc. Advances in Neural Information Processing Systems 26*, Lake Tahoe, NV, USA, December 5-10, 2013, pp. 2193–2201.

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

[13] U. S. Kamilov and H. Mansour, "Learning optimal nonlinearities for iterative thresholding algorithms," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 747–751, May 2016.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] A. Antoniadis, "Wavelet methods in statistics: Some recent ddevelopment and their applications," *Statistical Surveys*, vol. 1, pp. 16–55, 2007.

[19] M. Kowalski, "Thresholding rules and iterative shrinkage/thresholding algorithm: A convergence study," in *Proc. IEEE Int. Conf. Image Process (ICIP 2014)*, Paris, France, October 27-30, 2014, pp. 4151–4155.