A FAST ITERATIVE SHRINKAGE-THRESHOLDING ALGORITHM WITH APPLICATION TO WAVELET-BASED IMAGE DEBLURRING

Amir Beck

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology
Haifa 32000, Israel

Marc Teboulle

School of Mathematical Sciences
Tel-Aviv University
Ramat-Aviv 69978, Israel

ABSTRACT

We consider the class of Iterative Shrinkage-Thresholding Algorithms (ISTA) for solving linear inverse problems arising in signal/image processing. This class of methods is attractive due to its simplicity, however, they are also known to converge quite slowly. In this paper we present a Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) which preserves the computational simplicity of ISTA, but with a global rate of convergence which is proven to be significantly better, both theoretically and practically. Initial promising numerical results for wavelet-based image deblurring demonstrate the capabilities of FISTA.

Index Terms— iterative shrinkage-thresholding algorithm, least squares and l_1 regularization problems, optimal gradient method, two steps iterative algorithms, image deblurring.

1. INTRODUCTION

A basic linear inverse problem is to estimate an unknown signal $\mathbf{x}$ known to satisfy the relation

$$\mathbf{Ax} = \mathbf{b} + \mathbf{w}. \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$ are known and $\mathbf{w}$ is an unknown noise (or perturbation) vector.

A classical approach to this estimation problem is the least squares (LS) approach in which the estimator is chosen to minimize the least squares term $\|\mathbf{Ax} - \mathbf{b}\|^2$. In many applications, such as image deblurring, it is often the case that $\mathbf{A}$ is ill-conditioned and in these cases the LS solution usually has a huge norm and is thus meaningless. To overcome this difficulty, regularization methods are required to stabilize the solution. One regularization method that attracted a revived interest and considerable amount of attention in the signal processing literature is l_1 regularization in which one seeks to find the solution of

$$\min_{\mathbf{x}}\{F(\mathbf{x}) \equiv \|\mathbf{Ax} - \mathbf{b}\|^2 + \lambda\|\mathbf{x}\|_1\}, \tag{2}$$

This research is partially supported by the Israel Science Foundation, ISF grant #489-06.

where $\|\mathbf{x}\|_1$ stands for the sum of the absolute values of the components of $\mathbf{x}$, see e.g., [1, 2, 3, 4]. The presence of the l_1 term is used to induce sparsity in the optimal solution, of (2), see e.g., [5, 6]. In image deblurring for example, $\mathbf{A}$ is often chosen as $\mathbf{A} = \mathbf{RW}$ where $\mathbf{R}$ is the blurring matrix and $\mathbf{W}$ contains a wavelet basis. The underlying philosophy here in dealing with the l_1 norm regularization criterion is that most images have a sparse representation in the wavelet domain.

In many applications, e.g., in image deblurring, the problem is not only large scale (can reach millions of decision variables), but also involves dense matrix data, which often precludes the use and potential advantage of sophisticated interior point methods. This motivated the search of simpler gradient-based algorithms for solving (2), where the dominant computational effort is a relatively cheap matrix-vector multiplications involving $\mathbf{A}$ and $\mathbf{A}^T$. One of the most popular methods to solve problem (2) is in the class of *iterative shrinkage/thresholding* algorithms (ISTA), see e.g. [7, 1, 3, 8]. Specifically, the general step of ISTA is

$$\mathbf{x}_{k+1} = \mathcal{T}_{\lambda t_k}\left(\mathbf{x}_k - 2t_k\mathbf{A}^T(\mathbf{Ax}_k - \mathbf{b})\right) \tag{3}$$

where t_k is an appropriate stepsize and $\mathcal{T}_\alpha : \mathbb{R}^n \to \mathbb{R}^n$ is the shrinkage operator defined by

$$\mathcal{T}_\alpha(\mathbf{x})_i = (|x_i| - \alpha)_+\mathrm{sgn}\,(x_i). \tag{4}$$

Clearly, each iteration of ISTA is comprised of a gradient step of the smooth part followed by a shrinkage operation. The convergence analysis of ISTA has been well studied in the literature under various contexts and frameworks, including various modifications, see e.g., [1, 3, 9] and references therein. The advantage of ISTA is in its simplicity. However, ISTA has also been recognized as a slow method. Traditionally, the convergence analysis of iterative algorithms focuses on the asymptotic convergence of the sequence $\{\mathbf{x}_k\}$. Here, we focus on the *nonasymptotic* global rate of convergence and efficiency of methods measured through functions values, and we present a fast iterative shrinkage/thresholding algorithm (FISTA) for solving the general problem

$$\min_{\mathbf{x}}\{F(\mathbf{x}) \equiv \{f(\mathbf{x}) + g(\mathbf{x})\}, \tag{5}$$

where f and g are convex functions, with g possibly nonsmooth (see Section 2 for precise description). This algorithm is shown to converge to the optimal function value with the faster[1], rate of $O(1/k^2)$, k being the iteration number. Moreover, FISTA shares the same simplicity and computational demand of ISTA.

Recently, several accelerations of ISTA have been proposed in the literature, e.g., [11, 12]. The recent scheme of [11], called TwIST, uses at each step the last two iterations and is also based on a "gradient" type step followed by a shrinkage operation. Within another line of analysis, the recent work of [12] uses sequential subspace optimization techniques to accelerate ISTA. The speedup gained by both of these methods over ISTA was demonstrated experimentally on various linear inverse problems. However, for both of these two recent methods [11, 12], global rate of convergence have not been established. More recently, a different speed-up of ISTA was introduced in [13] for problem (5)with a $O(1/k^2)$ global rate of convergence. Although the method in [13] and FISTA shares the same $O(1/k^2)$ complexity result, the two schemes are very much different. In particular, the method of [13] requires two projections at each iteration, as opposed to one in FISTA. Moreover, FISTA is a two-steps method, while the scheme of [13] is a multistep method. Finally, the analysis of the two methods is completely different, see [10, 13] for details.

The paper is organized as follows. In Section 2, we recall ISTA as applied to the general model (5) and present the convergence result in this general setting. In Section 3 we present the details of the algorithm FISTA for both the constant and non-constant stepsize rules and provide the promised faster rate of convergence. In Section 4 we describe some preliminary numerical results for image deblurring problems, which demonstrate that FISTA can be even faster than the proven theoretical rate, and can outperform existing algorithms by several orders of magnitude.

Omitted proofs and further details on FISTA can be found in [10].

2. ISTA FOR THE GENERAL MODEL

As mentioned in the introduction, for the purpose of our analysis, we consider the following general formulation which naturally extends the problem formulation (2):

$$(P) \qquad \min\{F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}, \qquad (6)$$

where $g : \mathbb{R}^n \to \mathbb{R}$ is a continuous convex function which is possibly *nonsmooth* and $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth convex function with gradient which is Lipschitz continuous. That is, there exists a constant $L(f)$ for which

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L(f)\|\mathbf{x} - \mathbf{y}\| \quad \text{for every } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

[1]Function values for ISTA converge sublineraly, i.e. with a rate of $O(1/k)$, see, [10] for details.

We also denote the optimal solution set by X_*. In this setting, the general step of ISTA is of the form

$$\mathbf{x}_{k+1} = \text{prox}_{t_k}(g)(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)),$$

where the prox operation is defined by

$$\text{prox}_t(g)(\mathbf{x}) := \underset{\mathbf{u}}{\text{argmin}} \left\{ g(\mathbf{u}) + \frac{1}{2t}\|\mathbf{u} - \mathbf{x}\|^2 \right\}. \qquad (7)$$

The simplicity of ISTA therefore depends on the ability to compute the prox operation. When $g(\mathbf{x}) := \lambda\|\mathbf{x}\|_1$, then the prox operation is the same as soft thresholding, thus rendering it easy to compute. In other cases, the prox operation might not be so easy to compute. For example when g is chosen as a total variation function, then the computation of the prox operation requires the solution a total-variation based denoising problem (see e.g. [11]), for which good algorithms exist, but still there seems to be no explicit formula for the prox in this case. In general, the prox operation is easy to compute when $g(\cdot)$ is separable, since in that case the computation of prox reduces to solving a one dimensional minimization problem.

It is also interesting to note that when $g(\mathbf{x}) := 0$, the prox operation is just the identity operator and therefore ISTA in this case is the gradient method. For the gradient method it is known that the sequence of function values $F(\mathbf{x}_k)$ converges to the optimal function value F_* at a rate of convergence that is no worse than $O(1/k)$ also called a "sublinear" rate of convergence. The same result also holds for ISTA, see [10].

Theorem 2.1 *Let $\{\mathbf{x}_k\}$be the sequence generated by ISTA with a constant stepsize $t_k = 1/L(f)$. Then for any $k \geq 1$:*

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{L(f)\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2k}, \ \forall\, \mathbf{x}^* \in X_*. \qquad (8)$$

3. FISTA: A FAST ITERATIVE SHRINKAGE/THRESHOLDING ALGORITHM

We have just shown that ISTA has a worst-case complexity result of $O(1/k)$. In this section we present a simple fast iterative shrinkage/thresholding algorithm with an improved rate of $O(1/k^2)$.

We recall that when $g(\mathbf{x}) \equiv 0$ the general model (6) consists of minimizing a smooth convex function and ISTA reduces to the gradient method. In this smooth setting it was proven in [14] that there exists a gradient method with an $O(1/k^2)$ complexity result which is an "optimal" first-order method for smooth problems. The remarkable fact is that the method developed in [14] does not require more than one gradient evaluation at each iteration (namely, same as the gradient method), but just the computation of a smartly chosen linear combination of the two previous iterates.

In this section we extend the method of [14] to the general model (6) and we establish the improved complexity result. We begin by presenting the algorithm with a constant stepsize.

FISTA with constant stepsize
Input: $L = L(f)$ - A Lipschitz constant of ∇f.
Step 0. Take $\mathbf{y}_1 = \mathbf{x}_0 \in \mathbb{R}^n,\ t_1 = 1$.
Step k. ($k \geq 1$) Compute

$$\mathbf{x}_k = \text{prox}_{t_k}(g)\left(\mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k)\right), \quad (9)$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \quad (10)$$

$$\mathbf{y}_{k+1} = \mathbf{x}_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}_k - \mathbf{x}_{k-1}). \quad (11)$$

The main difference between the above algorithm and ISTA, is that that the iterative-shrinkage step (9) is not employed on the previous point $\mathbf{x}_{k-1}$, but rather at the point $\mathbf{y}_k$ which uses a very specific linear combination of the previous two points $\{\mathbf{x}_{k-1}, \mathbf{x}_k\}$. Obviously the main computational effort in both ISTA and FISTA remains the same. The requested additional computation for FISTA in the steps (10) and (11) is clearly marginal.

Since the Lipschitz constant is not always computable, we also provide a version of FISTA with a backtracking stepsize rule.

FISTA with backtracking
Step 0. Take $L_0 > 0$, some $\eta > 1$ and $\mathbf{x}_0 \in \mathbb{R}^n$. Set $\mathbf{y}_1 = \mathbf{x}_0,\ t_1 = 1$.
Step k. ($k \geq 1$) Find the smallest nonnegative integers i_k such that with $i = i_k$, $\bar{L} = \eta^{i_k} L_{k-1}$:

$$F(p_{\bar{L}}(\mathbf{y}_k)) \leq Q_{\bar{L}}(p_{\bar{L}}(\mathbf{y}_k), \mathbf{y}_k).$$

Set $L_k = \eta^{i_k} L_{k-1}$ and compute

$$\mathbf{x}_k = p_{L_k}(\mathbf{y}_k),$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2},$$

$$\mathbf{y}_{k+1} = \mathbf{x}_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}_k - \mathbf{x}_{k-1}).$$

The improved $O(1/k^2)$ convergence result for both variations is given in the following theorem.

Theorem 3.1 *Let $\{\mathbf{x}_k\}, \{\mathbf{y}_k\}$ be generated by FISTA. Then for any $k \geq 1$*

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{2\alpha L(f)\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{(k+1)^2}, \quad \forall \mathbf{x}^* \in X_*. \quad (12)$$

where $\alpha = 1$ for the constant stepsize setting and $\alpha = \eta$ for the backtracking stepsize setting.

4. NUMERICAL EXAMPLES

In this section we illustrate by some image deblurring problems the performance of the iterative shrinkage/thresholding algorithm FISTA compared to the basic iterative shrinkage/thresholding algorithm ISTA and to the recent TWIST algorithm of [11]. Since our simulations consider extremely ill-conditioned problems, the TWIST method is not guaranteed to converge and we thus use the montone version of TWIST termed MTWIST. The parameters for the MTWIST method were chosen as suggested in Section 6 of [11] for extremely ill-conditioned problems. All methods were used with a constant step size rule and applied on the l_1 regularization problem (2), that is $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ and $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$.

In the first example we look at the 256×256 cameraman test image. The image went through a Gaussian blur of size 9×9 and standard deviation 4 followed by a an additive zero-mean white Gaussian noise with standard deviation 10^{-3}.

For these experiments we assume reflexive (Neumann) boundary conditions. We then tested ISTA, FISTA and MTWIST for solving problem (2) where $\mathbf{b}$ represents the (vectorized) observed image and $\mathbf{A} = \mathbf{RW}$ where $\mathbf{R}$ is the matrix representing the blur operator and $\mathbf{W}$ is the inverse of a three stage Haar wavelet transform. The regularization parameter was chosen to be $\lambda =$ 2e-5 and the initial image was the blurred image. Iterations 100 and 200 are described in Figure 1. The function value at iteration k is denoted by F_k. The images produced by FISTA are of a better quality than those created by ISTA and MTWIST. It is also clear that MTWIST gives better results than ISTA. The function value of FISTA was consistently lower than the function value of ISTA and MTWIST. We also computed the function values produced after 1000 iterations for ISTA, MTWIST and FISTA which were respectively 2.45e-1, 2.31e-1 and 2.23e-1. Note that the function value of ISTA after 1000 iterations is still worse (that is, larger) than the function value of FISTA after 100 iterations and the function value of MTWIST after 1000 iterations is worse than the function value of FISTA after 200 iterations.

We also considered an example in which the optimal solution is known. For that sake we considered a 64×64 version of the previous test image which undergoes the same blur operator as the previous example. No noise was added and we solved the least squares problem, that is $\lambda = 0$. The optimal solution of this problem is zero. The function values of the three methods for 10000 iterations are described in Figure 2. The results produced by FISTA are better than those produced by ISTA and MTWIST by several orders of magnitude and clearly demonstrate the effective performance of FISTA. One can see that after 10000 iterations FISTA reaches accuracy of approximately 10^{-7} while ISTA and MTWIST reach accuracies of 10^{-3} and 10^{-4} respectively. Finally, we observe that the values obtained by ISTA and MTWIST at iteration 10000 was already obtained by FISTA at iterations 275 and 468 respectively. These preliminary computational results indicate that FISTA is a simple and promising iterative scheme, which can be even faster than the proven predicted theoretical rate.

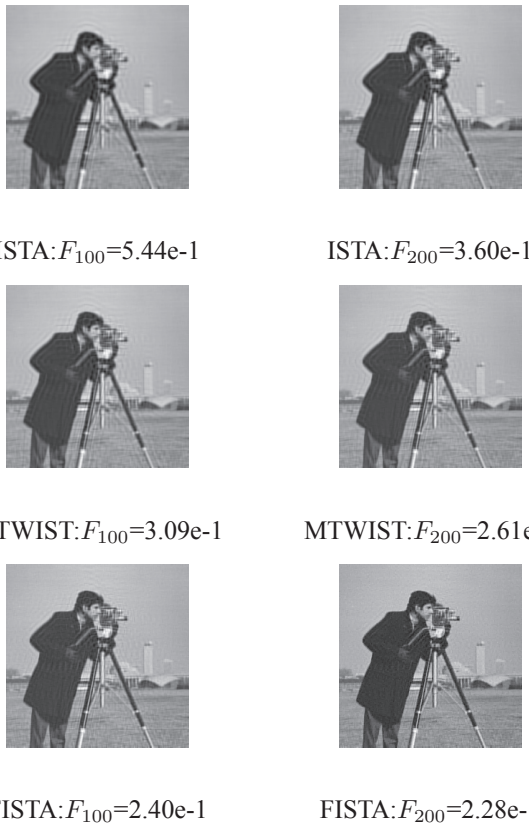

Fig. 1. Iterations of ISTA, MTWIST and FISTA methods for deblurring of the cameraman

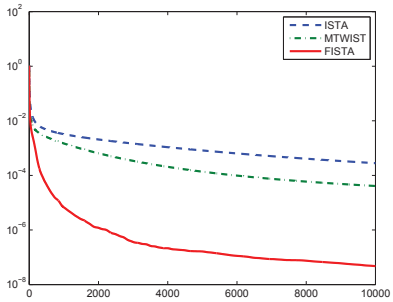

Fig. 2. Comparison of function values errors $F(\mathbf{x}_k) - F(\mathbf{x}^*)$ of ISTA, MTWIST and FISTA

5. REFERENCES

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

[2] J. L. Starck, D. L. Donoho, and E. J. Candès, "Astronomical image representation by the curevelet transform," *Astronomy and Astrophisics*, vol. 398, no. 2, pp. 785–800, February 2003.

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

[4] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," To appear in IEEE J. Selected Topics in Signal Processing, 2007.

[5] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1200–1224, 1995.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61 (electronic), 1998.

[7] A. Chambolle, R. A. DeVore, N. Y. Lee, and B. J. Lucier, "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Processing*, vol. 7, pp. 319–335, 1998.

[8] C. Vonesch and M. Unser, "Fast iterative thresholding algorithm for wavelet-regularized deconvolution," in *Proceedings of the SPIE Conference on Mathematical Imaging: Wavelet XII*, San Diego, CA, USA, 2007, vol. 6701, pp. 1–5.

[9] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling and Simulation*, vol. 4, pp. 1168–1200, 2005.

[10] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," Accepted for Publication in SIAM J. on Imaging Sciences.

[11] J. Bioucas-Dias and M. Figueiredo, "A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. on Image Processing*, vol. 16, pp. 2992–3004, 2007.

[12] M. Elad, B. Matalon, and M. Zibulevsky, "Subspace optimization methods for linear least squares with non-quadratic regularization," *Applied and Computational Harmonic Analysis*, vol. 23, pp. 346–367, 2007.

[13] Y. E. Nesterov, "Gradient methods for minimizing composite objective function," 2007, CORE Report. Available at http://www.ecore.beDPs/dp1191313936.pdf.

[14] Y. E. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," *Dokl. Akad. Nauk SSSR*, vol. 269, no. 3, pp. 543–547, 1983.