SIAM J. IMAGING SCIENCES
Vol. 2, No. 1, pp. 183–202

A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*

Amir Beck† and Marc Teboulle‡

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, which can be viewed as an extension of the classical gradient algorithm, is attractive due to its simplicity and thus is adequate for solving large-scale problems even with dense matrix data. However, such methods are also known to converge quite slowly. In this paper we present a new 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 which is shown to be faster than ISTA by several orders of magnitude.

Key words. iterative shrinkage-thresholding algorithm, deconvolution, linear inverse problem, least squares and l_1 regularization problems, optimal gradient method, global rate of convergence, two-step iterative algorithms, image deblurring

AMS subject classifications. 90C25, 90C06, 65F22

DOI. 10.1137/080716542

1. Introduction.

Linear inverse problems arise in a wide range of applications such as astrophysics, signal and image processing, statistical inference, and optics, to name just a few. The interdisciplinary nature of inverse problems is evident through a vast literature which includes a large body of mathematical and algorithmic developments; see, for instance, the monograph [13] and the references therein.

A basic linear inverse problem leads us to study a discrete linear system of the form

$$\mathbf{A}\mathbf{x} = \mathbf{b} + \mathbf{w}, \tag{1.1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$ are known, $\mathbf{w}$ is an unknown noise (or perturbation) vector, and $\mathbf{x}$ is the "true" and unknown signal/image to be estimated. In image blurring problems, for example, $\mathbf{b} \in \mathbb{R}^m$ represents the blurred image, and $\mathbf{x} \in \mathbb{R}^n$ is the unknown true image, whose size is assumed to be the same as that of $\mathbf{b}$ (that is, $m = n$). Both $\mathbf{b}$ and $\mathbf{x}$ are formed by stacking the columns of their corresponding two-dimensional images. In these applications, the matrix $\mathbf{A}$ describes the blur operator, which in the case of spatially invariant blurs represents a two-dimensional convolution operator. The problem of estimating $\mathbf{x}$ from the observed blurred and noisy image $\mathbf{b}$ is called an *image deblurring* problem.

*Received by the editors February 25, 2008; accepted for publication (in revised form) October 23, 2008; published electronically March 4, 2009. This research was partially supported by the Israel Science Foundation, ISF grant 489-06.
http://www.siam.org/journals/siims/2-1/71654.html
†Department of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 32000, Israel (becka@ie.technion.ac.il.).
‡School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel (teboulle@post.tau.ac.il.).

1.1. Background. A classical approach to problem (1.1) is the least squares (LS) approach [4] in which the estimator is chosen to minimize the data error:

$$\text{(LS):}\quad \hat{\mathbf{x}}_{\text{LS}} = \operatorname*{argmin}_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2.$$

When $m = n$ (as is the case in some image processing applications) and $\mathbf{A}$ is nonsingular, the LS estimate is just the näive solution $\mathbf{A}^{-1}\mathbf{b}$. In many applications, such as image deblurring, it is often the case that $\mathbf{A}$ is ill-conditioned [22], 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. The basic idea of regularization is to replace the original ill-conditioned problem with a "nearby" well-conditioned problem whose solution approximates the required solution. One of the popular regularization techniques is Tikhonov regularization [33] in which a quadratic penalty is added to the objective function:

$$\text{(T):}\quad \hat{\mathbf{x}}_{\text{TIK}} = \operatorname*{argmin}_{\mathbf{x}} \{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \lambda\|\mathbf{L}\mathbf{x}\|^2\}. \tag{1.2}$$

The second term in the above minimization problem is a regularization term that controls the norm (or seminorm) of the solution. The regularization parameter $\lambda > 0$ provides a tradeoff between fidelity to the measurements and noise sensitivity. Common choices for $\mathbf{L}$ are the identity or a matrix approximating the first or second order derivative operator [19, 21, 17].

Another regularization method that has 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

$$\min_{\mathbf{x}} \{F(\mathbf{x}) \equiv \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \lambda\|\mathbf{x}\|_1\}, \tag{1.3}$$

where $\|\mathbf{x}\|_1$ stands for the sum of the absolute values of the components of $\mathbf{x}$; see, e.g., [15, 32, 10, 16]. More references on earlier works promoting the use of l_1 regularization, as well as its relevance to other research areas, can be found in the recent work [16]. In image deblurring applications, and in particular in wavelet-based restoration methods, $\mathbf{A}$ is often chosen as $\mathbf{A} = \mathbf{R}\mathbf{W}$, where $\mathbf{R}$ is the blurring matrix and $\mathbf{W}$ contains a wavelet basis (i.e., multiplying by $\mathbf{W}$ corresponds to performing inverse wavelet transform). The vector $\mathbf{x}$ contains the coefficients of the unknown image. The underlying philosophy in dealing with the l_1 norm regularization criterion is that most images have a sparse representation in the wavelet domain. The presence of the l_1 term is used to induce sparsity in the optimal solution of (1.3); see, e.g., [11, 8]. Another important advantage of the l_1-based regularization (1.3) over the l_2-based regularization (1.2) is that as opposed to the latter, l_1 regularization is less sensitive to outliers, which in image processing applications correspond to sharp edges.

The convex optimization problem (1.3) can be cast as a second order cone programming problem and thus could be solved via interior point methods [1]. However, in most 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 (1.3), where the dominant computational effort is a relatively cheap matrix-vector multiplication involving $\mathbf{A}$ and $\mathbf{A}^T$; see, for instance, the recent study [16], where problem (1.3) is reformulated as a box-constrained quadratic problem and solved by a gradient projection algorithm. One of the most popular methods for solving problem (1.3) is in the class of *iterative shrinkage-thresholding* algorithms (ISTA), where each iteration involves matrix-vector multiplication involving $\mathbf{A}$ and $\mathbf{A}^T$ followed by a shrinkage/soft-threshold step;[1] see, e.g., [7, 15, 10, 34, 18, 35]. Specifically, the general step of ISTA is

$$\mathbf{x}_{k+1} = \mathcal{T}_{\lambda t}\left(\mathbf{x}_k - 2t\mathbf{A}^T(\mathbf{A}\mathbf{x}_k - \mathbf{b})\right), \tag{1.4}$$

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

$$\mathcal{T}_\alpha(\mathbf{x})_i = (|x_i| - \alpha)_+ \operatorname{sgn}(x_i). \tag{1.5}$$

In the optimization literature, this algorithm can be traced back to the proximal forward-backward iterative scheme introduced in [6] and [30] within the general framework of splitting methods; see [14, Chapter 12] and the references therein for a very good introduction to this approach, including convergence results. Another interesting recent contribution including very general convergence results for the sequence $\mathbf{x}_k$ produced by proximal forward-backward algorithms under various conditions and settings relevant to linear inverse problems can be found in [9].

1.2. Contribution. The convergence analysis of ISTA has been well studied in the literature under various contexts and frameworks, including various modifications; see, e.g., [15, 10, 9] and the references therein, with a focus on establishing conditions under which the sequence $\{\mathbf{x}_k\}$ converges to a solution of (1.3). The advantage of ISTA is in its simplicity. However, ISTA has also been recognized as a slow method. The very recent manuscript [5] provides further rigorous grounds to that claim by proving that under some assumptions on the operator $\mathbf{A}$ the sequence $\{\mathbf{x}_k\}$ produced by ISTA shares an asymptotic rate of convergence that can be very slow and arbitrarily bad (for details, see in particular Theorem 3 and the conclusion in [5, section 6]).

In this paper, we focus on the *nonasymptotic* global rate of convergence and efficiency of methods like ISTA measured through function values. Our development and analysis will consider the more general nonsmooth convex optimization model

$$\min_{\mathbf{x}}\{F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}, \tag{1.6}$$

where f, g are convex functions, with g possibly nonsmooth (see section 2.2 for a precise description). Basically, the general step of ISTA is of the form

$$\mathbf{x}_{k+1} = \mathcal{T}_{\lambda t}(G(\mathbf{x}_k)),$$

where $G(\cdot)$ stands for a gradient step of the fit-to-data LS term in (1.3) and ISTA is an extension of the classical gradient method (see section 2 for details). Therefore, ISTA belongs

[1]Other names in the signal processing literature include, for example, threshold Landweber method, iterative denoising, and deconvolution algorithms.

to the class of first order methods, that is, optimization methods that are based on function values and gradient evaluations. It is well known that for large-scale problems first order methods are often the only practical option, but as alluded to above it has been observed that the sequence $\{\mathbf{x}_k\}$ converges quite slowly to a solution. In fact, as a first result we further confirm this property by proving that ISTA behaves like

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \simeq O(1/k),$$

namely, shares a *sublinear global rate* of convergence.

The important question then is whether we can devise a faster method than the iterative shrinkage-thresholding scheme described above, in the sense that the computational effort of the new method will keep the simplicity of ISTA, while its global rate of convergence will be significantly better, both theoretically and practically. This is the main contribution of this paper which answers this question affirmatively. To achieve this goal, we consider a method which is similar to ISTA and of the form

$$\mathbf{x}_{k+1} = \mathcal{T}_{\lambda t}(G(\mathbf{y}_k)),$$

where the new point $\mathbf{y}_k$ will be smartly chosen and easy to compute; see section 4. This idea builds on an algorithm which is not so well known and which was introduced and developed by Nesterov in 1983 [27] for minimizing a *smooth* convex function, and proven to be an "optimal" first order (gradient) method in the sense of complexity analysis [26].

Here, the problem under consideration is convex but *nonsmooth*, due to the l_1 term. Despite the presence of a nonsmooth regularizer in the objective function, we prove that we can construct a faster algorithm than ISTA, called FISTA, that keeps its simplicity but shares the improved rate $O(1/k^2)$ of the optimal gradient method devised earlier in [27] for minimizing smooth convex problems. Our theoretical analysis is general and can handle an objective function with any convex nonsmooth regularizers (beyond l_1) and any smooth convex function (instead of the LS term), and constraints can also be handled.

1.3. Some recent algorithms accelerating ISTA. Very recently other researchers have been working on alternative algorithms that could speed up the performance of ISTA. Like FISTA proposed in this paper, these methods also rely on computing the next iterate based not only on the previous one, but on two or more previously computed iterates. One such line of research was very recently considered in [3], where the authors proposed an interesting two-step ISTA (TWIST) which, under some assumptions on the problem's data and appropriately chosen parameters defining the algorithm, is proven to converge to a minimizer of an objective function of the form

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \varphi(\mathbf{x}), \tag{1.7}$$

where φ is a convex nonsmooth regularizer. The effectiveness of TWIST as a faster method than ISTA was demonstrated experimentally on various linear inverse problems [3].

Another line of analysis toward an acceleration of ISTA for the same class of problems (1.7) was recently considered in [12] by using sequential subspace optimization techniques and relying on generating the next iterate by minimizing a function over an affine subspace spanned by two or more previous iterates and the current gradient. The speedup gained by this approach has been shown through numerical experiments for denoising application problems. For both of these recent methods [3, 12], global nonasymptotic rate of convergence has not been established.

After this paper was submitted for publication we recently became aware[2] of a very recent unpublished manuscript by Nesterov [28], who has independently investigated a multistep version of an accelerated gradient-like method that also solves the general problem model (1.6) and, like FISTA, is proven to converge in function values as $O(1/k^2)$, where k is the iteration counter. While both algorithms theoretically achieve the same global rate of convergence, the two schemes are remarkably different both conceptually and computationally. In particular, the main differences between FISTA and the new method proposed in [28] are that (a) on the building blocks of the algorithms, the latter uses an accumulated history of the past iterates to build recursively a sequence of estimate functions $\psi_k(\cdot)$ that approximates $F(\cdot)$, while FISTA uses just the usual projection-like step, evaluated at an auxiliary point very specially constructed in terms of the two previous iterates and an explicit dynamically updated stepsize; (b) the new Nesterov's method requires two projection-like operations per iteration, as opposed to one single projection-like operation needed in FISTA. As a consequence of the key differences between the building blocks and iterations of FISTA versus the new method of [28], the theoretical analysis and proof techniques developed here to establish the global rate convergence rate result are completely different from that given in [28].

1.4. Outline of the paper. In section 2, we recall some basic results pertinent to gradient-based methods and provide the building blocks necessary to the analysis of ISTA and, more importantly, of FISTA. Section 3 proves the aforementioned slow rate of convergence for ISTA, and in section 4 we present the details of the new algorithm FISTA and prove the promised faster rate of convergence. In section 5 we present some preliminary numerical results for image deblurring problems, which demonstrate that FISTA can be even faster than the proven theoretical rate and can outperform ISTA by several orders of magnitude, thus showing the potential promise of FISTA. To gain further insights into the potential of FISTA we have also compared it with the recent algorithm TWIST of [3]. These preliminary numerical results show evidence that FISTA can also be faster than TWIST by several orders of magnitude.

Notation. The inner product of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is denoted by $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T\mathbf{y}$. For a matrix $\mathbf{A}$, the maximum eigenvalue is denoted by $\lambda_{\max}(\mathbf{A})$. For a vector $\mathbf{x}$, $\|\mathbf{x}\|$ denotes the Euclidean norm of $\mathbf{x}$. The spectral norm of a matrix $\mathbf{A}$ is denoted by $\|\mathbf{A}\|$.

2. The building blocks of the analysis.

In this section we first recall some basic facts on gradient-based methods. We then formulate our problem and establish in Lemma 2.3 a result which will play a central role in the global convergence rate of analysis of the algorithms under study.

2.1. Gradient methods and ISTA.

Consider the unconstrained minimization problem of a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$:

$$\text{(U)} \qquad \min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}.$$

[2]We are also grateful to a referee for pointing out to us Nesterov's manuscript [28].

One of the simplest methods for solving (U) is the gradient algorithm which generates a sequence $\{\mathbf{x}_k\}$ via

$$\mathbf{x}_0 \in \mathbb{R}^n, \quad \mathbf{x}_k = \mathbf{x}_{k-1} - t_k \nabla f(\mathbf{x}_{k-1}), \tag{2.1}$$

where $t_k > 0$ is a suitable stepsize. It is very well known (see, e.g., [31, 2]) that the gradient iteration (2.1) can be viewed as a proximal regularization [24] of the linearized function f at $\mathbf{x}_{k-1}$, and written equivalently as

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x}} \left\{ f(\mathbf{x}_{k-1}) + \langle \mathbf{x} - \mathbf{x}_{k-1}, \nabla f(\mathbf{x}_{k-1}) \rangle + \frac{1}{2t_k} \|\mathbf{x} - \mathbf{x}_{k-1}\|^2 \right\}.$$

Adopting this same basic gradient idea to the nonsmooth l_1 regularized problem

$$\min\{f(\mathbf{x}) + \lambda \|\mathbf{x}\|_1 : \mathbf{x} \in \mathbb{R}^n\} \tag{2.2}$$

leads to the iterative scheme

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x}} \left\{ f(\mathbf{x}_{k-1}) + \langle \mathbf{x} - \mathbf{x}_{k-1}, \nabla f(\mathbf{x}_{k-1}) \rangle + \frac{1}{2t_k} \|\mathbf{x} - \mathbf{x}_{k-1}\|^2 + \lambda \|\mathbf{x}\|_1 \right\}.$$

After ignoring constant terms, this can be rewritten as

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x}} \left\{ \frac{1}{2t_k} \|\mathbf{x} - (\mathbf{x}_{k-1} - t_k \nabla f(\mathbf{x}_{k-1}))\|^2 + \lambda \|\mathbf{x}\|_1 \right\}, \tag{2.3}$$

which is a special case of the scheme introduced in [30, model (BF), p. 384] for solving (2.2). Since the l_1 norm is separable, the computation of $\mathbf{x}_k$ reduces to solving a one-dimensional minimization problem for each of its components, which by simple calculus produces

$$\mathbf{x}_k = \mathcal{T}_{\lambda t_k}(\mathbf{x}_{k-1} - t_k \nabla f(\mathbf{x}_{k-1})),$$

where $\mathcal{T}_\alpha : \mathbb{R}^n \to \mathbb{R}^n$ is the shrinkage operator given in (1.5).

Thus, with $f(\mathbf{x}) := \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$, the popular ISTA is recovered as a natural extension of a gradient-based method. As already mentioned in the introduction, for solving the l_1 problem (1.3), ISTA has been developed and analyzed independently through various techniques by many researchers. A typical condition ensuring convergence of $\mathbf{x}_k$ to a minimizer $\mathbf{x}^*$ of (1.3) is to require that $t_k \in (0, 1/\|\mathbf{A}^T\mathbf{A}\|)$. For example, this follows as a special case of a more general result which can be found in [14, Theorem 12.4.6] (see also Chapter 12 of [14] and its references for further details, modifications, and extensions).

2.2. The general model. As recalled above, the basic idea of the iterative shrinkage algorithm is to build at each iteration a regularization of the linearized differentiable function part in the objective. For the purpose of our analysis, we consider the following general formulation which naturally extends the problem formulation (1.3):

$$\text{(P)} \qquad \min\{F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}. \tag{2.4}$$

The following assumptions are made throughout the paper:

- $g : \mathbb{R}^n \to \mathbb{R}$ is a continuous convex function which is possibly *nonsmooth*.
- $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth convex function of the type $C^{1,1}$, i.e., continuously differentiable with Lipschitz continuous gradient $L(f)$:
$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L(f)\|\mathbf{x} - \mathbf{y}\| \quad \text{for every } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$
where $\|\cdot\|$ denotes the standard Euclidean norm and $L(f) > 0$ is the Lipschitz constant of ∇f.
- Problem (P) is solvable, i.e., $X_* := \operatorname{argmin} F \neq \emptyset$, and for $\mathbf{x}^* \in X_*$ we set $F_* := F(\mathbf{x}^*)$.

Example 2.1. When $g(\mathbf{x}) \equiv 0$, (P) is the general unconstrained smooth convex minimization problem.

Example 2.2. The l_1 regularization problem (1.3) is obviously a special instance of problem (P) by substituting $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$, $g(\mathbf{x}) = \|\mathbf{x}\|_1$. The (smallest) Lipschitz constant of the gradient ∇f is $L(f) = 2\lambda_{\max}(\mathbf{A}^T\mathbf{A})$.

2.3. The basic approximation model. In accordance with the basic results recalled in section 2.1, we adopt the following approximation model. For any $L > 0$, consider the following quadratic approximation of $F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})$ at a given point $\mathbf{y}$:

$$Q_L(\mathbf{x}, \mathbf{y}) := f(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y})\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 + g(\mathbf{x}), \tag{2.5}$$

which admits a unique minimizer

$$p_L(\mathbf{y}) := \operatorname{argmin}\{Q_L(\mathbf{x}, \mathbf{y}) : \ \mathbf{x} \in \mathbb{R}^n\}. \tag{2.6}$$

Simple algebra shows that (ignoring constant terms in $\mathbf{y}$)

$$p_L(\mathbf{y}) = \operatorname*{argmin}_{\mathbf{x}} \left\{ g(\mathbf{x}) + \frac{L}{2}\left\| \mathbf{x} - \left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right)\right\|^2 \right\}.$$

Clearly, the basic step of ISTA for problem (P) thus reduces to

$$\mathbf{x}_k = p_L(\mathbf{x}_{k-1}),$$

where L plays the role of a stepsize. Even though in our analysis we consider a general nonsmooth convex regularizer $g(\mathbf{x})$ in place of the l_1 norm, we will still refer to this more general method as ISTA.

2.4. The two pillars. Before proceeding with the analysis of ISTA we establish a key result (see Lemma 2.3 below) that will be crucial for the analysis of not only ISTA but also the new faster method introduced in section 4. For that purpose we first need to recall the first pillar, which is the following well-known and fundamental property for a smooth function in the class $C^{1,1}$; see, e.g., [29, 2].

Lemma 2.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function with Lipschitz continuous gradient and Lipschitz constant $L(f)$. Then, for any $L \geq L(f)$,*

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y})\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 \text{ for every } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{2.7}$$

We also need the following simple result which characterize the optimality of $p_L(\cdot)$.

Lemma 2.2. *For any $\mathbf{y} \in \mathbb{R}^n$, one has $\mathbf{z} = p_L(\mathbf{y})$ if and only if there exists $\gamma(\mathbf{y}) \in \partial g(\mathbf{z})$, the subdifferential of $g(\cdot)$, such that*

$$\nabla f(\mathbf{y}) + L(\mathbf{z} - \mathbf{y}) + \gamma(\mathbf{y}) = \mathbf{0}. \tag{2.8}$$

Proof. The proof is immediate from optimality conditions for the strongly convex problem (2.6). ∎

We are now ready to state and prove the promised key result.

Lemma 2.3. *Let $\mathbf{y} \in \mathbb{R}^n$ and $L > 0$ be such that*

$$F(p_L(\mathbf{y})) \le Q(p_L(\mathbf{y}), \mathbf{y}). \tag{2.9}$$

Then for any $\mathbf{x} \in \mathbb{R}^n$,

$$F(\mathbf{x}) - F(p_L(\mathbf{y})) \ge \frac{L}{2}\|p_L(\mathbf{y}) - \mathbf{y}\|^2 + L\langle \mathbf{y} - \mathbf{x}, p_L(\mathbf{y}) - \mathbf{y}\rangle.$$

Proof. From (2.9), we have

$$F(\mathbf{x}) - F(p_L(\mathbf{y})) \ge F(\mathbf{x}) - Q(p_L(\mathbf{y}), \mathbf{y}). \tag{2.10}$$

Now, since f, g are convex we have

$$\begin{aligned} f(\mathbf{x}) &\ge f(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y})\rangle, \\ g(\mathbf{x}) &\ge g(p_L(\mathbf{y})) + \langle \mathbf{x} - p_L(\mathbf{y}), \gamma(\mathbf{y})\rangle, \end{aligned}$$

where $\gamma(\mathbf{y})$ is defined in the premise of Lemma 2.2. Summing the above inequalities yields

$$F(\mathbf{x}) \ge f(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y})\rangle + g(p_L(\mathbf{y})) + \langle \mathbf{x} - p_L(\mathbf{y}), \gamma(\mathbf{y})\rangle. \tag{2.11}$$

On the other hand, by the definition of $p_L(\mathbf{y})$ one has

$$Q(p_L(\mathbf{y}), \mathbf{y}) = f(\mathbf{y}) + \langle p_L(\mathbf{y}) - \mathbf{y}, \nabla f(\mathbf{y})\rangle + \frac{L}{2}\|p_L(\mathbf{y}) - \mathbf{y}\|^2 + g(p_L(\mathbf{y})). \tag{2.12}$$

Therefore, using (2.11) and (2.12) in (2.10) it follows that

$$\begin{aligned} F(\mathbf{x}) - F(p_L(\mathbf{y})) &\ge -\frac{L}{2}\|p_L(\mathbf{y}) - \mathbf{y}\|^2 + \langle \mathbf{x} - p_L(\mathbf{y}), \nabla f(\mathbf{y}) + \gamma(\mathbf{y})\rangle \\ &= -\frac{L}{2}\|p_L(\mathbf{y}) - \mathbf{y}\|^2 + L\langle \mathbf{x} - p_L(\mathbf{y}), \mathbf{y} - p_L(\mathbf{y})\rangle, \\ &= \frac{L}{2}\|p_L(\mathbf{y}) - \mathbf{y}\|^2 + L\langle \mathbf{y} - \mathbf{x}, p_L(\mathbf{y}) - \mathbf{y}\rangle, \end{aligned}$$

where in the first equality above we used (2.8). ∎

Note that from Lemma 2.1, it follows that if $L \ge L(f)$, then the condition (2.9) is always satisfied for $p_L(\mathbf{y})$.

Remark 2.1. As a final remark in this section, we point out that all the above results and the forthcoming results in this paper also hold in any real Hilbert space setting. Moreover, all the results can be adapted for problem (P) with convex constraints. In that case, if C is a nonempty closed convex subset of $\mathbb{R}^n$, the computation of $p_L(\cdot)$ might require intensive computation, unless C is very simple (e.g., the nonnegative orthant). For simplicity of exposition, all the results are developed in the unconstrained and finite-dimensional setting.

3. The global convergence rate of ISTA. The convergence analysis of ISTA has been well studied for the l_1 regularization problem (1.3) and the more general problem (P), with a focus on conditions ensuring convergence of the sequence $\{\mathbf{x}_k\}$ to a minimizer. In this section we focus on the *nonasymptotic* global rate of convergence/efficiency of such methods, measured by function values.

We begin by stating the basic iteration of ISTA for solving problem (P) defined in (2.4).

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

$$\mathbf{x}_k = p_L(\mathbf{x}_{k-1}). \tag{3.1}$$

If $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ and $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ $(\lambda > 0)$, then algorithm (3.1) reduces to the basic iterative shrinkage method (1.4) with $t = \frac{1}{L(f)}$. Clearly, such a general algorithm will be useful when $p_L(\cdot)$ can be computed analytically or via a low cost scheme. This situation occurs particularly when $g(\cdot)$ is separable, since in that case the computation of p_L reduces to solving a one-dimensional minimization problem, e.g., with $g(\cdot)$ being the pth power of the l_p norm of $\mathbf{x}$, with $p \geq 1$. For such computation and other separable regularizers, see, for instance, the general formulas derived in [25, 7, 9].

A possible drawback of this basic scheme is that the Lipschitz constant $L(f)$ is not always known or computable. For instance, the Lipschitz constant in the l_1 regularization problem (1.3) depends on the maximum eigenvalue of $\mathbf{A}^T\mathbf{A}$ (see Example 2.2). For large-scale problems, this quantity is not always easily computable. We therefore also analyze ISTA with a backtracking stepsize rule.

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

$$F(p_{\bar{L}}(\mathbf{x}_{k-1})) \leq Q_{\bar{L}}(p_{\bar{L}}(\mathbf{x}_{k-1}), \mathbf{x}_{k-1}). \tag{3.2}$$

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

$$\mathbf{x}_k = p_{L_k}(\mathbf{x}_{k-1}). \tag{3.3}$$

Remark 3.1. Note that the sequence of function values $\{F(\mathbf{x}_k)\}$ produced by ISTA is nonincreasing. Indeed, for every $k \geq 1$,

$$F(\mathbf{x}_k) \leq Q_{L_k}(\mathbf{x}_k, \mathbf{x}_{k-1}) \leq Q_{L_k}(\mathbf{x}_{k-1}, \mathbf{x}_{k-1}) = F(\mathbf{x}_{k-1}),$$

where L_k is either chosen by the backtracking rule or $L_k \equiv L(f)$ is a given Lipschitz constant of ∇f.

Remark 3.2. Since inequality (3.2) is satisfied for $\bar{L} \geq L(f)$, where $L(f)$ is the Lipschitz constant of ∇f, it follows that for ISTA with backtracking one has $L_k \leq \eta L(f)$ for every $k \geq 1$. Overall,

$$\beta L(f) \leq L_k \leq \alpha L(f), \tag{3.4}$$

where $\alpha = \beta = 1$ for the constant stepsize setting and $\alpha = \eta$, $\beta = \frac{L_0}{L(f)}$ for the backtracking case.

Recall that ISTA reduces to the gradient method when $g(\mathbf{x}) \equiv \mathbf{0}$. 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 $O(1/k)$, which is also called a "sublinear" rate of convergence. That is, $F(\mathbf{x}_k) - F_* \leq C/k$ for some positive constant C; see, e.g., [23]. We show below that ISTA shares the same rate of convergence.

Theorem 3.1. *Let $\{\mathbf{x}_k\}$ be the sequence generated by either (3.1) or (3.3). Then for any $k \geq 1$*

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{\alpha L(f) \|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2k} \quad \forall \mathbf{x}^* \in X_*, \tag{3.5}$$

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

Proof. Invoking Lemma 2.3 with $\mathbf{x} = \mathbf{x}^*$, $\mathbf{y} = \mathbf{x}_n$, and $L = L_{n+1}$, we obtain

$$\begin{aligned} \frac{2}{L_{n+1}} (F(\mathbf{x}^*) - F(\mathbf{x}_{n+1})) &\geq \|\mathbf{x}_{n+1} - \mathbf{x}_n\|^2 + 2\langle \mathbf{x}_n - \mathbf{x}^*, \mathbf{x}_{n+1} - \mathbf{x}_n \rangle \\ &= \|\mathbf{x}^* - \mathbf{x}_{n+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}_n\|^2, \end{aligned}$$

which combined with (3.4) and the fact that $F(\mathbf{x}^*) - F(\mathbf{x}_{n+1}) \leq 0$ yields

$$\frac{2}{\alpha L(f)} (F(\mathbf{x}^*) - F(\mathbf{x}_{n+1})) \geq \|\mathbf{x}^* - \mathbf{x}_{n+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}_n\|^2. \tag{3.6}$$

Summing this inequality over $n = 0, \ldots, k-1$ gives

$$\frac{2}{\alpha L(f)} \left(k F(\mathbf{x}^*) - \sum_{n=0}^{k-1} F(\mathbf{x}_{n+1}) \right) \geq \|\mathbf{x}^* - \mathbf{x}_k\|^2 - \|\mathbf{x}^* - \mathbf{x}_0\|^2. \tag{3.7}$$

Invoking Lemma 2.3 one more time with $\mathbf{x} = \mathbf{y} = \mathbf{x}_n$ and $L = L_{n+1}$ yields

$$\frac{2}{L_{n+1}} (F(\mathbf{x}_n) - F(\mathbf{x}_{n+1})) \geq \|\mathbf{x}_n - \mathbf{x}_{n+1}\|^2.$$

Since $L_{n+1} \geq \beta L(f)$ (see (3.4) and $F(\mathbf{x}_n) - F(\mathbf{x}_{n+1}) \geq 0$), it follows that

$$\frac{2}{\beta L(f)} (F(\mathbf{x}_n) - F(\mathbf{x}_{n+1})) \geq \|\mathbf{x}_n - \mathbf{x}_{n+1}\|^2.$$

Multiplying the last inequality by n and summing over $n = 0, \ldots, k-1$, we obtain

$$\frac{2}{\beta L(f)} \sum_{n=0}^{k-1} (n F(\mathbf{x}_n) - (n+1) F(\mathbf{x}_{n+1}) + F(\mathbf{x}_{n+1})) \geq \sum_{n=0}^{k-1} n \|\mathbf{x}_n - \mathbf{x}_{n+1}\|^2,$$

which simplifies to

$$\frac{2}{\beta L(f)}\left(-kF(\mathbf{x}_k)+\sum_{n=0}^{k-1}F(\mathbf{x}_{n+1})\right)\geq\sum_{n=0}^{k-1}n\|\mathbf{x}_n-\mathbf{x}_{n+1}\|^2. \tag{3.8}$$

Adding (3.7) and (3.8) times β/α, we get

$$\frac{2k}{\alpha L(f)}\left(F(\mathbf{x}^*)-F(\mathbf{x}_k)\right)\geq\|\mathbf{x}^*-\mathbf{x}_k\|^2+\frac{\beta}{\alpha}\sum_{n=0}^{k-1}n\|\mathbf{x}_n-\mathbf{x}_{n+1}\|^2-\|\mathbf{x}^*-\mathbf{x}_0\|^2,$$

and hence it follows that

$$F(\mathbf{x}_k)-F(\mathbf{x}^*)\leq\frac{\alpha L(f)\|\mathbf{x}-\mathbf{x}_0\|^2}{2k}. \qquad \blacksquare$$

The above result can be interpreted as follows. The number of iterations of ISTA required to obtain an ε-optimal solution, that is, an $\tilde{\mathbf{x}}$ such that $F(\tilde{\mathbf{x}})-F_*\leq\varepsilon$, is at most $\lceil C/\varepsilon\rceil$, where $C=\frac{\alpha L(f)\|\mathbf{x}_0-\mathbf{x}^*\|^2}{2}$.

In the next section we will show that we can devise a different method, which is as simple as ISTA, but with a significantly improved complexity result.

4. FISTA: A fast iterative shrinkage-thresholding algorithm.

In the previous section we showed that ISTA has a worst-case complexity result of $O(1/k)$. In this section we will introduce a new ISTA with an improved complexity result of $O(1/k^2)$.

We recall that when $g(\mathbf{x})\equiv 0$, the general model (2.4) consists of minimizing a smooth convex function and ISTA reduced to the gradient method. In this smooth setting it was proven in [27] that there exists a gradient method with an $O(1/k^2)$ complexity result which is an "optimal" first order method for smooth problems, in the sense of Nemirovsky and Yudin [26]. The remarkable fact is that the method developed in [27] does not require more than one gradient evaluation at each iteration (namely, same as the gradient method) but just an additional point that is smartly chosen and easy to compute.

In this section we extend the method of [27] to the general model (2.4) and we establish the improved complexity result. Our analysis also provides a simple proof for the special smooth case (i.e., with $g(\mathbf{x})\equiv 0$) as well.

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

$$\mathbf{x}_k=p_L(\mathbf{y}_k), \tag{4.1}$$

$$t_{k+1}=\frac{1+\sqrt{1+4t_k^2}}{2}, \tag{4.2}$$

$$\mathbf{y}_{k+1}=\mathbf{x}_k+\left(\frac{t_k-1}{t_{k+1}}\right)(\mathbf{x}_k-\mathbf{x}_{k-1}). \tag{4.3}$$

The main difference between the above algorithm and ISTA is that the iterative shrinkage operator $p_L(\cdot)$ 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 $\{\mathbf{x}_{k-1}, \mathbf{x}_{k-2}\}$. Obviously the main computational effort in both ISTA and FISTA remains the same, namely, in the operator p_L. The requested additional computation for FISTA in the steps (4.2) and (4.3) is clearly marginal. The specific formula for (4.2) emerges from the recursive relation that will be established below in Lemma 4.1.

For the same reasons already explained in section 3, we will also analyze FISTA with a backtracking stepsize rule, which we now state explicitly.

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 $\bar{L} = \eta^{i_k} L_{k-1}$

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

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

$$\begin{aligned} \mathbf{x}_k &= p_{L_k}(\mathbf{y}_k), \\ t_{k+1} &= \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \\ \mathbf{y}_{k+1} &= \mathbf{x}_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}_k - \mathbf{x}_{k-1}). \end{aligned}$$

Note that the upper and lower bounds on L_k given in Remark 3.2 still hold true for FISTA, namely,

$$\beta L(f) \leq L_k \leq \alpha L(f).$$

The next result provides the key recursive relation for the sequence $\{F(\mathbf{x}_k) - F(\mathbf{x}^*)\}$ that will imply the better complexity rate $O(1/k^2)$. As we shall see, Lemma 2.3 of section 2 plays a central role in the proofs.

Lemma 4.1. *The sequences $\{\mathbf{x}_k, \mathbf{y}_k\}$ generated via FISTA with either a constant or backtracking stepsize rule satisfy for every $k \geq 1$*

$$\frac{2}{L_k} t_k^2 v_k - \frac{2}{L_{k+1}} t_{k+1}^2 v_{k+1} \geq \|\mathbf{u}_{k+1}\|^2 - \|\mathbf{u}_k\|^2,$$

where $v_k := F(\mathbf{x}_k) - F(\mathbf{x}^*)$, $\mathbf{u}_k := t_k \mathbf{x}_k - (t_k - 1)\mathbf{x}_{k-1} - \mathbf{x}^*$.

Proof. First we apply Lemma 2.3 at the points $(\mathbf{x} := \mathbf{x}_k, \mathbf{y} := \mathbf{y}_{k+1})$ with $L = L_{k+1}$, and likewise at the points $(\mathbf{x} := \mathbf{x}^*, \mathbf{y} := \mathbf{y}_{k+1})$, to get

$$\begin{aligned} 2L_{k+1}^{-1}(v_k - v_{k+1}) &\geq \|\mathbf{x}_{k+1} - \mathbf{y}_{k+1}\|^2 + 2\langle \mathbf{x}_{k+1} - \mathbf{y}_{k+1}, \mathbf{y}_{k+1} - \mathbf{x}_k\rangle, \\ -2L_{k+1}^{-1} v_{k+1} &\geq \|\mathbf{x}_{k+1} - \mathbf{y}_{k+1}\|^2 + 2\langle \mathbf{x}_{k+1} - \mathbf{y}_{k+1}, \mathbf{y}_{k+1} - \mathbf{x}^*\rangle, \end{aligned}$$

where we used the fact that $\mathbf{x}_{k+1} = p_{L_{k+1}}(\mathbf{y}_{k+1})$. To get a relation between v_k and v_{k+1}, we multiply the first inequality above by $(t_{k+1} - 1)$ and add it to the second inequality:

$$\frac{2}{L_{k+1}}((t_{k+1}-1)v_k - t_{k+1}v_{k+1}) \geq t_{k+1}\|\mathbf{x}_{k+1}-\mathbf{y}_{k+1}\|^2 + 2\langle \mathbf{x}_{k+1}-\mathbf{y}_{k+1}, t_{k+1}\mathbf{y}_{k+1}-(t_{k+1}-1)\mathbf{x}_k-\mathbf{x}^*\rangle.$$

Multiplying the last inequality by t_{k+1} and using the relation $t_k^2 = t_{k+1}^2 - t_{k+1}$ which holds thanks to (4.2), we obtain

$$\frac{2}{L_{k+1}}(t_k^2 v_k - t_{k+1}^2 v_{k+1}) \geq \|t_{k+1}(\mathbf{x}_{k+1}-\mathbf{y}_{k+1})\|^2 + 2t_{k+1}\langle \mathbf{x}_{k+1}-\mathbf{y}_{k+1}, t_{k+1}\mathbf{y}_{k+1}-(t_{k+1}-1)\mathbf{x}_k-\mathbf{x}^*\rangle.$$

Applying the usual Pythagoras relation

$$\|\mathbf{b}-\mathbf{a}\|^2 + 2\langle \mathbf{b}-\mathbf{a}, \mathbf{a}-\mathbf{c}\rangle = \|\mathbf{b}-\mathbf{c}\|^2 - \|\mathbf{a}-\mathbf{c}\|^2$$

to the right-hand side of the last inequality with

$$\mathbf{a} := t_{k+1}\mathbf{y}_{k+1}, \quad \mathbf{b} := t_{k+1}\mathbf{x}_{k+1}, \quad \mathbf{c} := (t_{k+1}-1)\mathbf{x}_k + \mathbf{x}^*,$$

we thus get

$$\frac{2}{L_{k+1}}(t_k^2 v_k - t_{k+1}^2 v_{k+1}) \geq \|t_{k+1}\mathbf{x}_{k+1} - (t_{k+1}-1)\mathbf{x}_k - \mathbf{x}^*\|^2 - \|t_{k+1}\mathbf{y}_{k+1} - (t_{k+1}-1)\mathbf{x}_k - \mathbf{x}^*\|^2.$$

Therefore, with $\mathbf{y}_{k+1}$ (cf. (4.3)) and $\mathbf{u}_k$ defined by

$$t_{k+1}\mathbf{y}_{k+1} = t_{k+1}\mathbf{x}_k + (t_k-1)(\mathbf{x}_k - \mathbf{x}_{k-1}) \quad \text{and} \quad \mathbf{u}_k = t_k\mathbf{x}_k - (t_k-1)\mathbf{x}_{k-1} - \mathbf{x}^*,$$

it follows that

$$\frac{2}{L_{k+1}}(t_k^2 v_k - t_{k+1}^2 v_{k+1}) \geq \|\mathbf{u}_{k+1}\|^2 - \|\mathbf{u}_k\|^2,$$

which combined with the inequality $L_{k+1} \geq L_k$ yields

$$\frac{2}{L_k}t_k^2 v_k - \frac{2}{L_{k+1}}t_{k+1}^2 v_{k+1} \geq \|\mathbf{u}_{k+1}\|^2 - \|\mathbf{u}_k\|^2. \qquad \blacksquare$$

We also need the following trivial facts.

Lemma 4.2. *Let $\{a_k, b_k\}$ be positive sequences of reals satisfying*

$$a_k - a_{k+1} \geq b_{k+1} - b_k \quad \forall k \geq 1, \text{ with } a_1 + b_1 \leq c, \ c > 0.$$

Then, $a_k \leq c$ for every $k \geq 1$.

Lemma 4.3. *The positive sequence $\{t_k\}$ generated in FISTA via (4.2) with $t_1 = 1$ satisfies $t_k \geq (k+1)/2$ for all $k \geq 1$.*

We are now ready to prove the promised improved complexity result for FISTA.

Theorem 4.4. *Let $\{\mathbf{x}_k\}, \{\mathbf{y}_k\}$ be generated by FISTA. Then for any $k \geq 1$*

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{2\alpha L(f)\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{(k+1)^2} \quad \forall \mathbf{x}^* \in X_*, \tag{4.4}$$

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

Proof. Let us define the quantities

$$a_k := \frac{2}{L_k} t_k^2 v_k, \quad b_k := \|\mathbf{u}_k\|^2, \quad c := \|\mathbf{y}_1 - \mathbf{x}^*\|^2 = \|\mathbf{x}_0 - \mathbf{x}^*\|^2,$$

and recall (cf. Lemma 4.1) that $v_k := F(\mathbf{x}_k) - F(\mathbf{x}^*)$. Then, by Lemma 4.1 we have for every $k \geq 1$

$$a_k - a_{k+1} \geq b_{k+1} - b_k,$$

and hence assuming that $a_1 + b_1 \leq c$ holds true, invoking Lemma 4.2, we obtain that

$$\frac{2}{L_k} t_k^2 v_k \leq \|\mathbf{x}_0 - \mathbf{x}^*\|^2,$$

which combined with $t_k \geq (k+1)/2$ (by Lemma 4.3) yields

$$v_k \leq \frac{2L_k \|\mathbf{x}_0 - \mathbf{x}^*\|^2}{(k+1)^2}.$$

Utilizing the upper bound on L_k given in (3.4), the desired result (4.4) follows. Thus, all that remains is to prove the validity of the relation $a_1 + b_1 \leq c$. Since $t_1 = 1$, and using the definition of $\mathbf{u}_k$ given in Lemma 4.1, we have here

$$a_1 = \frac{2}{L_1} t_1 v_1 = \frac{2}{L_1} v_1, \quad b_1 = \|\mathbf{u}_1\|^2 = \|\mathbf{x}_1 - \mathbf{x}^*\|.$$

Applying Lemma 2.3 to the points $\mathbf{x} := \mathbf{x}^*$, $\mathbf{y} := \mathbf{y}_1$ with $L = L_1$, we get

$$F(\mathbf{x}^*) - F(p(\mathbf{y}_1)) \geq \frac{L_1}{2} \|p(\mathbf{y}_1) - \mathbf{y}_1\|^2 + L_1 \langle \mathbf{y}_1 - \mathbf{x}^*, p(\mathbf{y}_1) - \mathbf{y}_1 \rangle. \tag{4.5}$$

Thus,

$$\begin{aligned}
F(\mathbf{x}^*) - F(\mathbf{x}_1) &= F(\mathbf{x}^*) - F(p(\mathbf{y}_1)) \\
&\overset{(4.5)}{\geq} \frac{L_1}{2} \|p(\mathbf{y}_1) - \mathbf{y}_1\|^2 + L_1 \langle \mathbf{y}_1 - \mathbf{x}^*, p(\mathbf{y}_1) - \mathbf{y}_1 \rangle \\
&= \frac{L_1}{2} \|\mathbf{x}_1 - \mathbf{y}_1\|^2 + L_1 \langle \mathbf{y}_1 - \mathbf{x}^*, \mathbf{x}_1 - \mathbf{y}_1 \rangle \\
&= \frac{L_1}{2} \{\|\mathbf{x}_1 - \mathbf{x}^*\|^2 - \|\mathbf{y}_1 - \mathbf{x}^*\|^2\}.
\end{aligned}$$

Consequently,

$$\frac{2}{L_1} v_1 \leq \|\mathbf{y}_1 - \mathbf{x}^*\|^2 - \|\mathbf{x}_1 - \mathbf{x}^*\|^2;$$

that is, $a_1 + b_1 \leq c$ holds true. ■

The number of iterations of FISTA required to obtain an ε-optimal solution, that is, an $\tilde{\mathbf{x}}$ such that $F(\tilde{\mathbf{x}}) - F_* \leq \varepsilon$, is at most $\lceil C/\sqrt{\varepsilon} - 1 \rceil$, where $C = \sqrt{2\alpha L(f) \|\mathbf{x}_0 - \mathbf{x}^*\|^2}$, and which clearly improves ISTA. In the next section we demonstrate the practical value of this theoretical global convergence rate estimate derived for FISTA on the l_1 wavelet-based regularization problem (1.3).

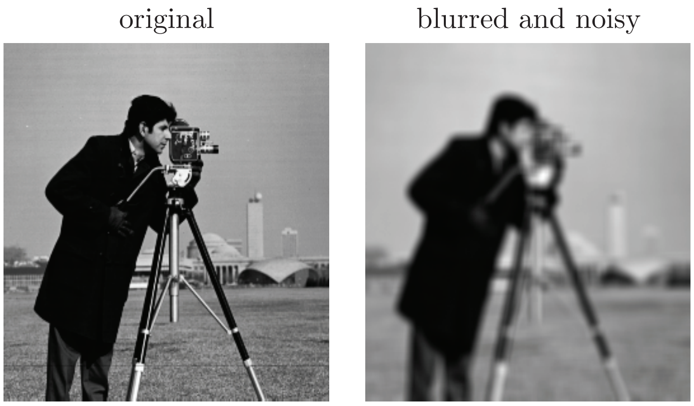

Figure 1. *Deblurring of the cameraman.*

5. Numerical examples.

In this section we illustrate the performance of FISTA compared to the basic ISTA and to the recent TWIST algorithm of [3]. Since our simulations consider extremely ill-conditioned problems (the smallest eigenvalue of $\mathbf{A}^T\mathbf{A}$ is nearly zero, and the maximum eigenvalue is 1), the TWIST method is not guaranteed to converge, and we thus use the monotone version of TWIST termed MTWIST. The parameters for the MTWIST method were chosen as suggested in [3, section 6] for extremely ill-conditioned problems. All methods were used with a constant stepsize rule and applied to the l_1 regularization problem (1.3), that is, $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ and $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$.

In all the simulations we have performed, we observed that FISTA significantly outperformed ISTA with respect to the number of iterations required to achieve a given accuracy. Similar conclusions can be made when compared with MTWIST. Below, we describe representative examples and results from these simulations.

5.1. Example 1: The cameraman test image.

All pixels of the original images described in the examples were first scaled into the range between 0 and 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 (applied by the MATLAB functions imfilter and fspecial) followed by an additive zero-mean white Gaussian noise with standard deviation 10^{-3}. The original and observed images are given in Figure 1.

For these experiments we assume reflexive (Neumann) boundary conditions [22]. We then tested ISTA, FISTA, and MTWIST for solving problem (1.3), where $\mathbf{b}$ represents the (vectorized) observed image, and $\mathbf{A} = \mathbf{R}\mathbf{W}$, 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. The Lipschitz constant was computable in this example (and those in what follows) since the eigenvalues of the matrix $\mathbf{A}^T\mathbf{A}$ can be easily calculated using the two-dimensional cosine transform [22]. Iterations 100 and 200 are described in Figure 2. 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

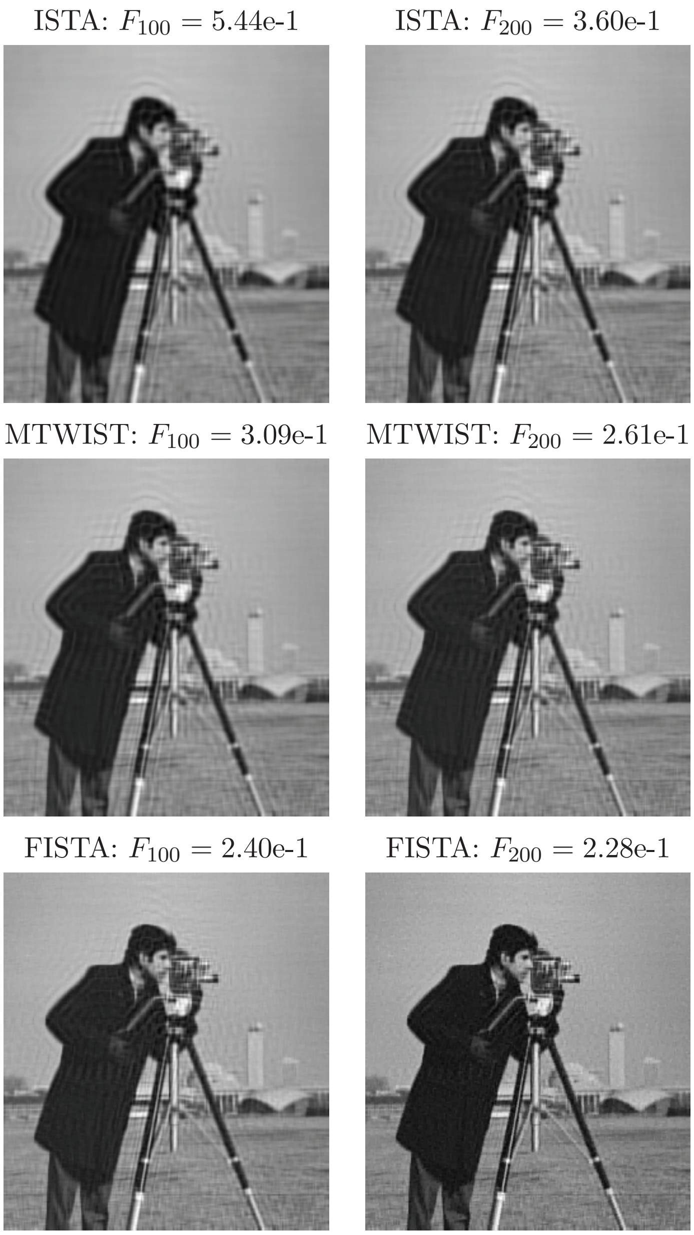

Figure 2. *Iterations of ISTA, MTWIST, and FISTA methods for deblurring of the cameraman.*

ISTA. The function value of FISTA was consistently lower than the function values 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.

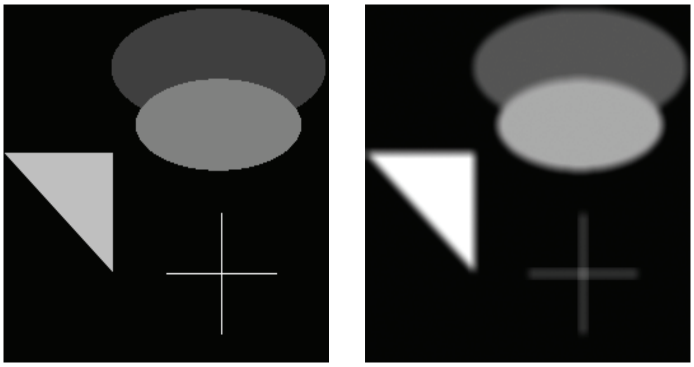

Figure 3. *Deblurring of the simple test image.*

5.2. Example 2: A simple test image. In this example we will further show the benefit of FISTA. The 256×256 simple test image was extracted from the function blur from the regularization toolbox [20]. The image then undergoes the same blurring and noise-adding procedure described in the previous example. The original and observed images are given in Figure 3.

The algorithms were tested with regularization parameter λ=1e-4 and with the same wavelet transform. The results of iterations 100 and 200 are described in Figure 4. Clearly, FISTA provides clearer images and improved function values. Moreover, the function value 0.321 obtained at iteration number 100 of FISTA is better than the function values of both ISTA and MTWIST methods at iteration number 200 (0.427 and 0.341, respectively). Moreover, MTWIST needed 416 iterations to reach the value that FISTA obtained after 100 iterations (0.321) and required 1102 iterations to reach the value 0.309 produced by FISTA after 200 iterations. In addition we ran the algorithm for tens of thousands of iterations and noticed that ISTA seems to get stuck at a function value of 0.323 and MTWIST gets stuck at a function value of 0.318.

From the previous example it seems that practically FISTA is able to reach accuracies that are beyond the capabilities of ISTA and MTWIST. To test this hypothesis 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 LS 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 5. 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 an 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 were 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. Its potential for analyzing and designing faster algorithms in other application areas and with

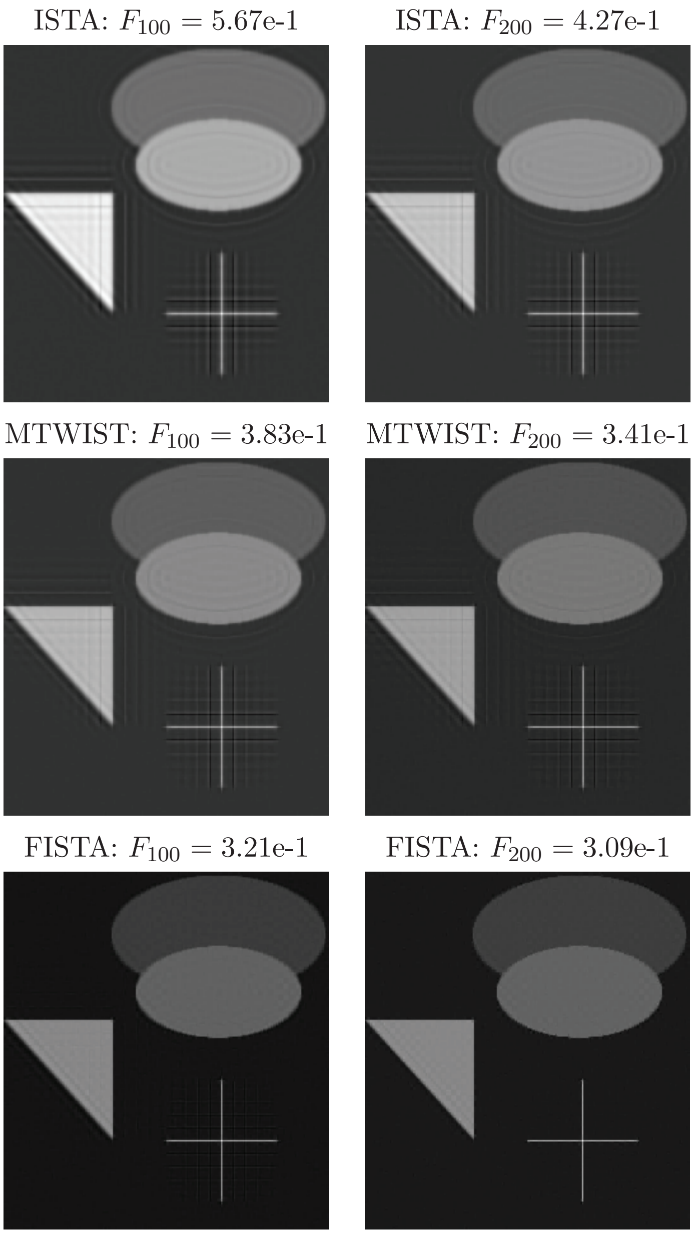

Figure 4. *Outputs of ISTA, MTWIST and FISTA for the simple test image.*

other types of regularizers, as well as a more thorough computational study, are topics of future research.

Acknowledgment. The authors are indebted to the two anonymous referees for their useful suggestions and for having drawn the authors' attention to additional relevant references.

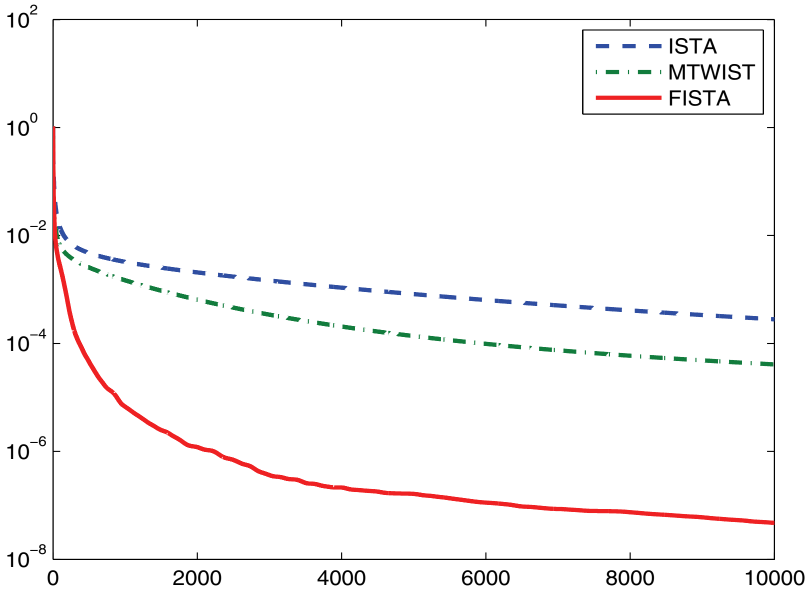

Figure 5. *Comparison of function value errors $F(\mathbf{x}_k) - F(\mathbf{x}^*)$ of ISTA, MTWIST, and FISTA.*

REFERENCES

[1] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS/SIAM Ser. Optim., SIAM, Philadelphia, 2001.

[2] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[3] J. Bioucas-Dias and M. Figueiredo, *A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration*, IEEE Trans. Image Process., 16 (2007), pp. 2992–3004.

[4] A. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.

[5] K. Bredies and D. Lorenz, *Iterative Soft-Thresholding Converges Linearly*, Technical report, 2008; available online at http://arxiv.org/abs/0709.1598v3.

[6] R. J. Bruck, *On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space*, J. Math. Anal. Appl., 61 (1977), pp. 159–164.

[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 Process., 7 (1998), pp. 319–335.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61.

[9] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.

[10] I. Daubechies, M. Defrise, and C. D. Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math., 57 (2004), pp. 1413–1457.

[11] D. L. Donoho and I. M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, J. Amer. Statist. Assoc., 90 (1995), pp. 1200–1224.

[12] M. Elad, B. Matalon, and M. Zibulevsky, *Subspace optimization methods for linear least squares with non-quadratic regularization*, Appl. Comput. Harmon. Anal., 23 (2007), pp. 346–367.

[13] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl. 375, Kluwer Academic Publishers Group, Dordrecht, The Netherlands, 1996.

[14] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vol. II, Springer Ser. Oper. Res., Springer-Verlag, New York, 2003.

[15] M. A. T. Figueiredo and R. D. Nowak, *An EM algorithm for wavelet-based image restoration*, IEEE Trans. Image Process., 12 (2003), pp. 906–916.

[16] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, *Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems*, IEEE J. Sel. Top. Signal Process., 1 (2007), pp. 586–597.

[17] G. H. Golub, P. C. Hansen, and D. P. O'Leary, *Tikhonov regularization and total least squares*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 185–194.

[18] E. Hale, W. Yin, and Y. Zhang, *A Fixed-Point Continuation Method for l_1-Regularized Minimization with Applications to Compressed Sensing*, CAAM Technical report TR07-07, Rice University, Houston, TX, 2007.

[19] P. C. Hansen and D. P. O'Leary, *The use of the L-curve in the regularization of discrete ill-posed problems*, SIAM J. Sci. Comput., 14 (1993), pp. 1487–1503.

[20] P. C. Hansen, *Regularization tools: A MATLAB package for analysis and solution of discrete ill-posed problems*, Numer. Algorithms, 6 (1994), pp. 1–35.

[21] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM, Philadelphia, 1997.

[22] P. C. Hansen, J. G. Nagy, and D. P. O'Leary, *Deblurring Images: Matrices, Spectra, and Filtering*, Fundam. Algorithms 3, SIAM, Philadelphia, 2006.

[23] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, Comput. Math. Math. Phys., 6 (1966), pp. 787–823.

[24] B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Française Informat. Recherche Opérationnelle, 4 (1970), pp. 154–158.

[25] P. Moulin and J. Liu, *Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors*, IEEE Trans. Inform. Theory, 45 (1999), pp. 909–919.

[26] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons, New York, 1983.

[27] Y. E. Nesterov, *A method for solving the convex programming problem with convergence rate $O(1/k^2)$*, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543–547 (in Russian).

[28] Y. E. Nesterov, *Gradient Methods for Minimizing Composite Objective Function*, CORE report, 2007; available at http://www.ecore.be/DPs/dp_1191313936.pdf.

[29] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Classics Appl. Math. 30, SIAM, Philadelphia, 2000.

[30] G. B. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, J. Math. Anal. Appl., 72 (1979), pp. 383–390.

[31] B. T. Polyak, *Introduction to Optimization*, Translations Series in Mathematics and Engineering, Optimization Software, Publications Division, New York, 1987.

[32] J. L. Starck, D. L. Donoho, and E. J. Candès, *Astronomical image representation by the curevelet transform*, Astron. Astrophys., 398 (2003), pp. 785–800.

[33] A. N. Tikhonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*, V. H. Winston, Washington, DC, 1977.

[34] C. Vonesch and M. Unser, *Fast iterative thresholding algorithm for wavelet-regularized deconvolution*, in Proceedings of the SPIE Optics and Photonics 2007 Conference on Mathematical Methods: Wavelet XII, Vol. 6701, San Diego, CA, 2007, pp. 1–5.

[35] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, *Sparse reconstruction by separable approximation*, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), 2008, pp. 3373–3376.