# LOCAL LINEAR CONVERGENCE OF ISTA AND FISTA ON THE LASSO PROBLEM *

SHAOZHE TAO†, DANIEL BOLEY‡, AND SHUZHONG ZHANG§

**Abstract.** We use a model LASSO problem to analyze the convergence behavior of the ISTA and FISTA iterations, showing that both iterations satisfy local linear convergence rate bound when close enough to the solution. Using the observation that FISTA is an accelerated ISTA process, and a spectral analysis of the associated matrix operators, we show that FISTA's convergence rate can slow down as it proceeds, eventually becoming slower than ISTA. This observation leads to a proposed heuristic algorithm to take an ISTA step if it shows more progress compared to FISTA, as measured by the decrease in the objective function. We illustrate the results with some synthetic numerical examples.

**1. Introduction.** The $l_1$-norm regularized least squares model has received much attention recently due to its wide applications in the real problems including compressive sensing [6], statistics [8], sparse coding [11], geophysics [21] and so on. The model problem in question is:

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) = {}^1\!/_2\|A\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$ is a given matrix, $\mathbf{b}$ is a given vector and $\lambda$ is a positive scalar.

Though the idea of least squares with $l_1$ regularization is decades old [21], it was not until more recently that this idea was introduced to the computational mathematics community under the name of Least Absolute Selection and Shrinkage Operator (LASSO) [22] and Basis Pursuit Denoising (or Compressive Sensing) [6]. For example, in compressive sensing, we seek to recover a solution $\mathbf{x}$ to an underdetermined $m \times n$ system of linear equations $A\mathbf{x} = \mathbf{b}$ with $n \gg m$. This linear system either has no solution (if $A$ is rank-deficient) or has many solutions. A common approach is to find the solution with minimum $l_2$-norm. However, it is often desired to find a solution $\mathbf{x}$ with as few non-zero entries as possible, as a way to find the fewest columns of $A$ needed to obtain a good approximation of the target vector $\mathbf{b}$. It is by now well known that an efficient way to obtain a sparse solution $\mathbf{x}$ is to use $l_1$ regularization as in (1.1). Since LASSO has become one of the most popular names for this model, we will use term LASSO to denote the above model for the remainder of the paper.

Although the LASSO problem can be cast as a second order cone programming problem and solved by standard general algorithms like an interior point method [2], the computational complexity of such traditional methods is too high to handle large-scale data encountered in many real applications. Recently, several algorithms that take advantage of the special structure of the LASSO problem have been proposed including alternating direction method of multipliers (ADMM), coordinate descent method, iterative shrinkage thresholding algorithm (ISTA) and its accelerated version fast iterative shrinkage thresholding algorithm (FISTA). In [20], we compared the local convergence of these four methods on the LASSO problem, showing that they

*This research was partially supported by NSF grants (IIS-1319749, CMMI-1462408).

†Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, United States. Email: taoxx120@umn.edu.

‡Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, United States. Email: boley@cs.umn.edu.

§Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, United States. Email: zhangs@umn.edu.

all exhibit local linear convergence but at different rates. The present paper is devoted exclusively to a theoretical and experimental comparison of ISTA vs FISTA, with a simplified proof of linear convergence and a proposed heuristic algorithm as a natural consequence of the analysis.

ISTA is an example of a proximal gradient method, since it involves a gradient descent step with a penalty to limit the length of each step. Its computation involves only matrix-vector multiplications, which has great advantage over many other convex algorithms by avoiding a matrix factorization [19]. Recently, Beck and Teboulle [1] proposed an accelerated ISTA, named as FISTA, in which a specific relaxation parameter is chosen. The idea of acceleration was first developed by Nesterov in [14]. Both algorithms are designed for solving more general problems containing convex differentiable objectives combined with an $l_1$ regularization term in the following general form:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + g(\mathbf{x}),$$

where $f$ is a smooth convex function and $g$ is a continuous convex function but possibly nonsmooth. Clearly, the LASSO problem is a special case of the above formulation with $f(\mathbf{x}) = {}^1\!/_2\|A\mathbf{x}-\mathbf{b}\|^2$, $g(\mathbf{x}) = \|\mathbf{x}\|_1$. The gradient $\nabla f = A^{\mathrm{T}}A\mathbf{x} - A^{\mathrm{T}}\mathbf{b}$ is Lipschitz continuous with constant $L(f) = \rho(A^{\mathrm{T}}A) = \|A^{\mathrm{T}}A\|_2$, i.e., $\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\| \leq L(f)\|\mathbf{x}_1 - \mathbf{x}_2\|, \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$. It has been shown [1] that FISTA provides a convergence rate of $O(1/k^2)$ compared to the rate of $O(1/k)$ by ISTA, while maintaining practically the same per-iteration cost, where $k$ is the iteration number.

In contrast to the global convergence rate, there have been several previous studies of the local convergence of ISTA. Bredies and Lorenz in [4] provided the local linear convergence of the iterative soft-thresholding algorithms in infinite dimensional Hilbert spaces under certain conditions. Recently, Liang et al. [13] showed that a general forward-backward proximal splitting algorithm converges linearly if the function satisfies the so-called partly smooth properties. The local linear convergence of FISTA is not clear in the literature. A Lyapunov analysis on the local convergence FISTA has been established in [18] contemporaneously with the present paper.

In this work, we establish local bounds on the convergence behavior of ISTA and FISTA on the model LASSO problem. We compare the asymptotic convergence behavior of ISTA with that of FISTA. The latter can be considered an accelerated ISTA, but we show that as one approaches the solution FISTA can slow down and even become slower than ISTA. Extending the same techniques as in [3, 20], we show that linear convergence can be reached eventually, but not necessarily from the beginning. The method is based on a representation of ISTA and FISTA as a matrix recurrence and a spectrial analysis of that matrix recurrence. This yields a model that shows that the iterations pass through several phases or "regimes", each treated separately in terms of the spectral model. After passing through several regimes, some of which consist of taking constant steps, each iteration reaches a regime of linear convergence with a convergence rate bounded by the eigenvalues of the corresponding matrix operator. Unlike [4, 13], our analysis is conducted on the finite dimensional Euclidean space so that we can take advantage of the well-established matrix properties. This lets us study the local convergence of not only ISTA but also FISTA, which is not considered in [4, 13]. Apart from the local convergence results, our analysis in terms of regimes can model the behavior of entire iteration process and give a way to compare ISTA and FISTA with each other.

Throughout this paper, the matrix $p$-norm is the norm induced by the corresponding vector norm: $\|M\|_p = \max_{\|\mathbf{v}\|_p=1} \|M\mathbf{v}\|_p$, with $p = 1, 2$ or $\infty$. We use $\rho(M)$ to denote the spectral radius (largest absolute value of any eigenvalue of a matrix $M$). For any real symmetric matrix $M$, the matrix 2-norm is the same as the spectral radius: $\|M\|_2 = \rho(M)$, hence we use those interchangeably for symmetric matrices. We also use a so-called $G$-norm where $G$ is a non-singular matrix, defined to be $\|\mathbf{v}\|_G = \|G\mathbf{v}\|_\infty$ for any vector $\mathbf{v}$, and $\|M\|_G = \|GAG^{-1}\|_\infty$ for any matrix $M$.

The paper is organized as follows. Section 2 gives some preliminaries of the paper. We show how to represent ISTA as a matrix recurrence in Section 3.1 and give some of its spectral properties in Section 3.2, and do the same for FISTA in Sections 3.3 & 3.4. Section 4 gives details about four types of regimes that ISTA and FISTA will encounter in the iterations process based on our spectral analysis. Our first main result is given in Section 5, which shows the local linear convergence of ISTA and FISTA on the LASSO problem. In Section 6 we compare the behavior in each regime, showing that FISTA can be faster that ISTA through most of the regimes, but asymptotically can be slower as one approaches the optimal solution. A heuristic algorithm is provided based on this observation. Section 7 includes two numerical examples run by the standard ISTA and FISTA, to illustrate many of the predicted behaviors.

## 2. Preliminaries.

**2.1. Optimality Condition of the LASSO Problem.** The first order KKT optimality conditions for the LASSO problem (1.1) are

$$A^{\mathrm{T}}(\mathbf{b} - A\mathbf{x}) = \lambda\boldsymbol{\nu} \tag{2.1}$$

where at optimality each component of $\boldsymbol{\nu}$ satisfies

$$\begin{Bmatrix} \nu_i = \mathrm{Sign}\,(x_i) & \text{if } x_i \neq 0 \\ -1 \leq \nu_i \leq +1 & \text{if } x_i = 0 \end{Bmatrix} \qquad \text{for} \quad i = 1, 2, \cdots . \tag{2.2}$$

Here the "Sign" function is defined as

$$\mathrm{Sign}\,(x) = \begin{cases} +1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

**2.2. Uniqueness and Strict Complementarity.** There are various sufficient and necessary conditions for the uniqueness of the LASSO problem or its variants. For example, [17, 5, 9] show different sufficient conditions and [23] studies the necessary conditions for the LASSO problem. In fact, the problem (1.1) needs to have a unique solution in many situations. For example, in compressive sensing and signal recovery, having non-unique solutions results in unreliable recovery given the data. We refer readers to [23, 24] and references therein for a discussion of the uniqueness of the LASSO problem. Here we will occasionally assume uniqueness of the LASSO solution.

Let $\mathbf{x}^*$ be an optimal solution to (1.1) with corresponding residuals $\boldsymbol{\nu}^*$ (2.1). We say this solution has the property of *strict complementarity* if for every index $i$, the $i$-th components of $\mathbf{x}^*, \boldsymbol{\nu}^*$ satisfy either $x_i^* > 0$ and $\nu_i^* = \mathrm{Sign}\,(x_i) = \pm 1$ or else $x_i^* = 0$ and $|\nu_i^*| < 1$. The situation where $x_i^* = 0$ and $|\nu_i^*| = 1$ for the same index $i$ would violate the condition of strict complementarity. We remark that such violations occur only for finitely many values of $\lambda$ for a given $A$ and $\mathbf{b}$ [8].

**2.3. ISTA and FISTA iteration.** In this part, we review the basic iteration of ISTA and FISTA for solving the LASSO problem. To make clear the difference between ISTA and FISTA, we let $\widehat{\mathbf{x}}$ and $\widetilde{\mathbf{x}}$ denote the iterates of ISTA and FISTA respectively in the remainder of this paper. The basic step of ISTA for the LASSO problem can be reduced to [7, 1]

$$\begin{aligned}
\widehat{\mathbf{x}}^{[k+1]} &= \operatorname*{argmin}_{\widehat{\mathbf{x}}}\{g(\widehat{\mathbf{x}}) + {}^{L}\!/_{2}\|\widehat{\mathbf{x}} - (\widehat{\mathbf{x}}^{[k]} - {}^{1}\!/_{L}\nabla f(\widehat{\mathbf{x}}^{[k]}))\|^2\} \\
&= \operatorname*{argmin}_{\widehat{\mathbf{x}}}\{\lambda\|\widehat{\mathbf{x}}\|_1 + {}^{L}\!/_{2}\|\widehat{\mathbf{x}} - (\widehat{\mathbf{x}}^{[k]} - {}^{1}\!/_{L}(A^{\mathrm{T}}A\widehat{\mathbf{x}}^{[k]} - A^{\mathrm{T}}\mathbf{b})\|^2\} \\
&= \mathrm{Shr}_{\lambda/L}((I - {}^{1}\!/_{L}A^{\mathrm{T}}A)\widehat{\mathbf{x}}^{[k]} + {}^{1}\!/_{L}A^{\mathrm{T}}\mathbf{b}).
\end{aligned}$$

Here the shrinkage operator is defined in terms of a positive threshold $\xi$:

$$\mathrm{Shr}_{\xi}(s) = \begin{cases} s - \xi & \text{if } s \geq \xi \\ 0 & \text{if } -\xi < s < \xi \\ s + \xi & \text{if } s \leq -\xi, \end{cases}$$

where this is applied elementwise to vectors.

One advantage of ISTA is that the above step can be solved in closed form using only matrix-vector multiplications, leading to the following updates repeated until convergence. Here $L$ is a constant such that $L \geq \|A^{\mathrm{T}}A\|_2$, though for the ISTA analysis it could be as low as ${}^{1}\!/_{2}\|A^{\mathrm{T}}A\|_2$ (cf. §5.4). We show one pass through the algorithm, with $\widehat{\mathbf{x}}^{[k]}$ denoting the vector from previous pass and $\widehat{\mathbf{x}}^{[k+1]}$ denoting the new iterate.

Algorithm 1: One pass of ISTA

start with $\widehat{\mathbf{x}}^{[k]}$.
  Set $\widehat{\mathbf{x}}^{[k+1]} = \mathrm{Shr}_{\lambda/L}((I - {}^{1}\!/_{L}A^{\mathrm{T}}A)\widehat{\mathbf{x}}^{[k]} + {}^{1}\!/_{L}A^{\mathrm{T}}\mathbf{b})$.
Result is $\widehat{\mathbf{x}}^{[k+1]}$ for next pass.

FISTA differs from ISTA in that the shrinkage operator is employed not on the previous point $\widetilde{\mathbf{x}}^{[k-1]}$ but a different point $\mathbf{y}^{[k]}$, which uses a very specific linear combination of the previous two points $\{\widetilde{\mathbf{x}}^{[k-1]}, \widetilde{\mathbf{x}}^{[k-2]}\}$. The algorithm of FISTA for the LASSO problem is presented below, with the initial $\mathbf{y}^{[1]} = \widetilde{\mathbf{x}}^{[0]} \in \mathbb{R}^n$ and $t^{[0]} = t^{[1]} = 1$.

Algorithm 2: One pass of FISTA

start with $t^{[k]}$, $t^{[k-1]}$, $\widetilde{\mathbf{x}}^{[k-1]}$ and $\widetilde{\mathbf{x}}^{[k-2]}$.
1. Set $\mathbf{y}^{[k]} = \widetilde{\mathbf{x}}^{[k-1]} + \frac{t^{[k-1]}-1}{t^{[k]}}(\widetilde{\mathbf{x}}^{[k-1]} - \widetilde{\mathbf{x}}^{[k-2]})$.
2. Set $\widetilde{\mathbf{x}}^{[k]} = \mathrm{Shr}_{\lambda/L}((I - {}^{1}\!/_{L}A^{\mathrm{T}}A)\mathbf{y}^{[k]} + {}^{1}\!/_{L}A^{\mathrm{T}}\mathbf{b})$.
3. Set $t^{[k+1]} = \frac{1+\sqrt{1+4t^{[k]2}}}{2}$.

Result is $t^{[k+1]}$, $t^{[k]}$, $\widetilde{\mathbf{x}}^{[k]}$, $\widetilde{\mathbf{x}}^{[k-1]}$ for next pass.

## 3. Auxiliary Variables with Local Monotonic Behavior.

**3.1. ISTA as a Matrix Recurrence.** Instead of using variables $\widehat{\mathbf{x}}^{[k]}$, we use two auxiliary variables to carry the iteration. One variable, namely, $\widehat{\mathbf{w}}^{[k]}$ exhibits smooth behavior, with linear convergence locally around a fixed point, and the other variable $\widehat{\mathbf{d}}^{[k]}$ is simply a ternary vector based on the three cases of the shrinkage operator. We let, for all $k$, the common iterate be

$$\widehat{\mathbf{w}}^{[k]} = (I - {}^{1}\!/_{L}A^{\mathrm{T}}A)\widehat{\mathbf{x}}^{[k]} + {}^{1}\!/_{L}A^{\mathrm{T}}\mathbf{b} \tag{3.1}$$

and vector $\widehat{\mathbf{d}}^{[k]}$ is defined elementwise as

$$\widehat{d}_i^{[k]} = \mathrm{Sign}\left(\mathrm{Shr}_{\lambda/L}\, \widehat{w}_i^{[k]})\right) = \begin{cases} 1 & \text{if } \widehat{w}_i^{[k]} > \lambda/L \\ 0 & \text{if } -\lambda/L \leq \widehat{w}_i^{[k]} \leq \lambda/L \\ -1 & \text{if } \widehat{w}_i^{[k]} < -\lambda/L. \end{cases} \tag{3.2}$$

We also define the matrix $\widehat{D}^{[k]} = \mathrm{diag}(\widehat{\mathbf{d}}^{[k]})$. Since $\widehat{D}^{[k]}$ indicates the sign of the iterates, we call it flag matrix in the rest of the paper. By the updating rule in Alg. 1 and the above two equations, one can obtain the $\widehat{\mathbf{x}}^{[k]}$-update in terms of $\widehat{\mathbf{w}}^{[k]}$ and $\widehat{\mathbf{d}}^{[k]}$

$$\widehat{\mathbf{x}}^{[k+1]} = \mathrm{Shr}_{\lambda/L}(\widehat{\mathbf{w}}^{[k]}) = (\widehat{D}^{[k]})^2 \widehat{\mathbf{w}}^{[k]} - \lambda/L\, \widehat{\mathbf{d}}^{[k]}. \tag{3.3}$$

Using (3.1), (3.2) and (3.3), the update formula for $\widehat{\mathbf{w}}$ now can be expressed explicitly as follows:

$$\begin{aligned} \widehat{\mathbf{w}}^{[k+1]} &= R^{[k]}\widehat{\mathbf{w}}^{[k]} + \mathbf{h}^{[k]} \\ &= [(I - 1/L\, A^{\mathrm{T}}A)(\widehat{D}^{[k]})^2]\widehat{\mathbf{w}}^{[k]} - (I - 1/L\, A^{\mathrm{T}}A)\lambda/L\, \widehat{\mathbf{d}}^{[k]} + 1/L\, A^{\mathrm{T}}\mathbf{b} \end{aligned}$$

where we denote

$$\begin{array}{rcl} R^{[k]} & = & [(I - 1/L\, A^{\mathrm{T}}A)(\widehat{D}^{[k]})^2] \\ \mathbf{h}^{[k]} & = & -(I - 1/L\, A^{\mathrm{T}}A)\lambda/L\, \widehat{\mathbf{d}}^{[k]} + 1/L\, A^{\mathrm{T}}\mathbf{b} \end{array} \tag{3.4}$$

throughout this paper. Therefore, the ISTA in Alg. 1 with variable $\widehat{\mathbf{x}}$ can be modified to the following procedure using the new variables $\widehat{\mathbf{w}}$ and $\widehat{D}$.

Algorithm 3: One pass of modified ISTA

start with $\widehat{\mathbf{w}}^{[k]}$,$\widehat{D}^{[k]}$.

1. $\widehat{\mathbf{w}}^{[k+1]} = R^{[k]}\widehat{\mathbf{w}}^{[k]} + \mathbf{h}^{[k]}$ (with $R^{[k]}, \mathbf{h}^{[k]}$ defined by (3.4)).
2. $\widehat{D}^{[k+1]} = \mathrm{Diag}\left(\mathrm{Sign}\left(\mathrm{Shr}_{\lambda/L}(\widehat{\mathbf{w}}^{[k+1]})\right)\right)$.

Result is $\widehat{\mathbf{w}}^{[k+1]}$, $\widehat{D}^{[k+1]}$ for next pass.

Alg. 3 is mathematically equivalent to Alg. 1 and is designed only for the purpose of analysis, not intended for computation. We note that step 1 of Alg. 3 can be written as a homogeneous matrix recurrence in (3.5), which we will use to characterize ISTA's convergence.

$$\begin{aligned} \begin{pmatrix} \widehat{\mathbf{w}}^{[k+1]} \\ 1 \end{pmatrix} &= \mathbf{R}_{aug}^{[k]} \begin{pmatrix} \widehat{\mathbf{w}}^{[k]} \\ 1 \end{pmatrix} = \begin{pmatrix} R^{[k]} & \mathbf{h}^{[k]} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{w}}^{[k]} \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} (I - 1/L\, A^{\mathrm{T}}A)(\widehat{D}^{[k]})^2 & \mathbf{h}^{[k]} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{w}}^{[k]} \\ 1 \end{pmatrix} \end{aligned} \tag{3.5}$$

where we denote $\mathbf{R}_{\text{aug}}^{[k]}$ as $\begin{pmatrix} R^{[k]} & \mathbf{h}^{[k]} \\ 0 & 1 \end{pmatrix}$, the augmented matrix of $R^{[k]}$, in this paper.

It is known that the fixed point condition in Alg. 1 is equivalent to the KKT conditions (2.1) and (2.2). The following lemma shows the equivalence between the fixed point of the constructed matrix recurrence (3.5) and the KKT point of the LASSO problem.

Lemma 3.1. *Suppose* $\begin{pmatrix} \widehat{\mathbf{w}} \\ 1 \end{pmatrix}$ *is an eigenvector corresponding to eigenvalue 1 of* $\mathbf{R}_{\text{aug}}$*(omitting* $[k]$*) in* (3.5) *and* $\widehat{D}^{[k+1]} = \widehat{D}^{[k]} = \widehat{D} = \mathrm{Diag}\,(\widehat{\mathbf{d}})$ *with entries* $\widehat{d}_i = 1$ *if*

$\widehat{w}_i > {}^{\lambda}\!/_{L}$, $\widehat{d}_i = -1$ *if* $\widehat{w}_i < -{}^{\lambda}\!/_{L}$ *and* $\widehat{d}_i = 0$ *if* $-{}^{\lambda}\!/_{L} \le \widehat{w}_i \le {}^{\lambda}\!/_{L}$. *Then the variable defined by* $\widehat{\mathbf{x}} = Shr_{{}^{\lambda}\!/_{L}}(\widehat{\mathbf{w}})$ *satisfies the* 1*st order KKT conditions. Conversely, if* $\widehat{\mathbf{x}}$ *and* $\widehat{\boldsymbol{\nu}} = {}^{1}\!/_{\lambda} A^{\mathrm{T}}(\mathbf{b} - A\widehat{\mathbf{x}})$ *satisfy the KKT conditions, then* $\begin{pmatrix} \widehat{\mathbf{w}} \\ 1 \end{pmatrix}$, *with* $\widehat{\mathbf{w}} = \widehat{\mathbf{x}} + {}^{\lambda}\!/_{L}\widehat{\boldsymbol{\nu}}$, *is an eigenvector of* $\mathbf{R}_{\mathrm{aug}}$ *corresponding to eigenvalue 1, where* $\mathbf{R}_{\mathrm{aug}}$ *is defined in* (3.5) *and* $\widehat{D}^{[k+1]} = \widehat{D}^{[k]} = \widehat{D} = \mathrm{Diag}\,(\widehat{\mathbf{d}})$ *with entries* $\widehat{d}_i = 1$ *if* $\widehat{w}_i > {}^{\lambda}\!/_{L}$, $\widehat{d}_i = -1$ *if* $\widehat{w}_i < -{}^{\lambda}\!/_{L}$ *and* $\widehat{d}_i = 0$ *if* $-{}^{\lambda}\!/_{L} \le \widehat{w}_i \le {}^{\lambda}\!/_{L}$.

*Proof.* Define $\widehat{\boldsymbol{\nu}}^{[k]} = {}^{1}\!/_{\lambda} A^{\mathrm{T}}(\mathbf{b} - A\widehat{\mathbf{x}}^{[k]})$ for all $k$. Then from Algorithm 1, $\widehat{\mathbf{x}}^{[k+1]} = \mathrm{Shr}_{{}^{\lambda}\!/_{L}}(\widehat{\mathbf{w}}^{[k]}) = \mathrm{Shr}_{{}^{\lambda}\!/_{L}}(\widehat{\mathbf{x}}^{[k]} + {}^{\lambda}\!/_{L}\widehat{\boldsymbol{\nu}}^{[k]})$. $\widehat{\mathbf{w}}^{[k]}$ is a fixed point if and only if $\begin{pmatrix} \widehat{\mathbf{w}}^{[k]} \\ 1 \end{pmatrix}$ is a eigenvector corresponding to eigenvalue 1. But it is a fixed point if and only if $\widehat{\mathbf{x}}^{[k]} = \mathrm{Shr}_{{}^{\lambda}\!/_{L}}(\widehat{\mathbf{x}}^{[k]} + {}^{\lambda}\!/_{L}\widehat{\boldsymbol{\nu}}^{[k]})$, which holds if and only if $\widehat{\mathbf{x}}^{[k]}, \widehat{\boldsymbol{\nu}}^{[k]}$ satisfy conditions (2.1) (2.2). This last statement follows directly from the fact that $x = \mathrm{Shr}_{\xi}(x+y)$ if and only if $y = \mathrm{Thr}_{\xi}(x+y)$, where $\mathrm{Thr}_{\xi}$ is the threshold function satisfying $\mathrm{Shr}_{\xi} + \mathrm{Thr}_{\xi} = \mathrm{Id}$ (the identity function). □

**3.2. Spectral Properties of $\mathbf{R}_{\mathrm{aug}}^{[k]}$.** We give here some spectral properties of $\mathbf{R}_{\mathrm{aug}}^{[k]}$ that will play key roles in our convergence analysis. We first recall some theory relating the spectral radius to the matrix norm.

THEOREM 3.2. *Let* $\rho(M)$ *denote the spectral radius of an arbitrary square real matrix* $M$. *Then we have the following:*

(a). $\rho(M) \le \|M\|_2$.

(b). *If* $\|M\|_2 = \rho(M)$ *then for any eigenvalue* $\lambda$ *such that* $|\lambda| = \rho(M)$, *the algebraic and geometric multiplicities of* $\lambda$ *are the same (all Jordan blocks for* $\lambda$ *is* $1 \times 1$*). Such a matrix is said to be a member of Class M [12, 16].*

(c). *If a* $\lambda$ *such that* $|\lambda| = \rho(M)$ *has a Jordan block of dimension larger than* 1 *(the geometric multiplicity is strictly less than the algebraic multiplicity), then for any* $\epsilon > 0$ *there exists a matrix norm* $\|\cdot\|_G$ *such that* $\rho(M) < \|M\|_G \equiv \|GMG^{-1}\|_\infty \le \rho(M) + \epsilon$.

We refer readers to [3, 16, 12, 10] for the proof of the above theorem. We remark that (a) and (b) are true for any operator norm but here we need it only for the matrix 2-norm.

LEMMA 3.3. *Regarding ISTA, there are three properties of* $R^{[k]}$*:*

(a). $\|R^{[k]}\|_2 = \|(I - {}^{1}\!/_{L} A^{\mathrm{T}}A)(\widehat{D}^{[k]})^2\|_2 \le 1$.

(b). *All eigenvalues must lie in the interval* $[0, 1]$.

(c). *If there exists one or more eigenvalues equal to* 1, *then eigenvalue* 1 *must have a complete set of eigenvectors.*

*Proof.* We here omit the pass number ${}^{[k]}$ for simplicity.

(a). $\|R\|_2 = \|(I - {}^{1}\!/_{L} A^{\mathrm{T}}A)\widehat{D}^2\|_2 \le \|(I - {}^{1}\!/_{L} A^{\mathrm{T}}A)\|_2 \|\widehat{D}^2\|_2 \le 1$.

(b). All eigenvalues of $R$ are the same as those of $R' = \widehat{D}(I - {}^{1}\!/_{L} A^{\mathrm{T}}A)\widehat{D}$. Noticing $L \ge \|A^{\mathrm{T}}A\|_2$, we obtain $\|R'\|_2 \le \|\widehat{D}^2\|_2 \|(I - {}^{1}\!/_{L} A^{\mathrm{T}}A)\|_2 \le 1$. In addition, $R'$ is symmetric and a positive semidefinite matrix. Hence all eigenvalues of $R'$ must lie in the interval $[0, 1]$. Hence so should those of $R$.

(c). Because $\rho(R) = \|R\|_2 = 1$, this statement follows directly from Theorem 3.2. □

**3.3. FISTA as a Matrix Recurrence.** Similar to ISTA, we use auxiliary variables $\widetilde{\mathbf{w}}^{[k]}, \widetilde{D}^{[k]}$ to replace variable $\widetilde{\mathbf{x}}^{[k]}$ for carrying the FISTA iterations. We set

$$\widetilde{\mathbf{w}}^{[k]} = (I - {}^{1}\!/_{L} A^{\mathrm{T}} A)\mathbf{y}^{[k]} + {}^{1}\!/_{L} A^{\mathrm{T}}\mathbf{b}. \tag{3.6}$$

Hence,

$$\widetilde{\mathbf{x}}^{[k+1]} = \mathrm{Shr}_{\lambda/_L}(\widetilde{\mathbf{w}}^{[k]}) = (\widetilde{D}^{[k]})^2 \widetilde{\mathbf{w}}^{[k]} - {}^{\lambda}\!/_{L} \widetilde{\mathbf{d}}^{[k]} \tag{3.7}$$

where for all $k$, the flag matrix $\widetilde{D}^{[k]} = \mathrm{diag}(\widetilde{\mathbf{d}}^{[k]})$, and the vector $\widetilde{\mathbf{d}}^{[k]}$ is defined elementwise as

$$\widetilde{d}_i^{[k]} = \mathrm{Sign}\,(\mathrm{Shr}_{\lambda/_L} \widetilde{w}_i^{[k]})) = \begin{cases} 1 & \text{if } \widetilde{w}_i^{[k]} > {}^{\lambda}\!/_{L} \\ 0 & \text{if } -{}^{\lambda}\!/_{L} \le \widetilde{w}_i^{[k]} \le {}^{\lambda}\!/_{L} \\ -1 & \text{if } \widetilde{w}_i^{[k]} < -{}^{\lambda}\!/_{L}. \end{cases} \tag{3.8}$$

Using (3.6), (3.7) and the updating formula in Alg. 2, we arrive at

$$\begin{array}{rcl} \widetilde{\mathbf{w}}^{[k+1]} & = & (I - {}^{1}\!/_{L} A^{\mathrm{T}} A)\left[\widetilde{\mathbf{x}}^{[k]} + \frac{t^{[k]}-1}{t^{[k+1]}}(\widetilde{\mathbf{x}}^{[k]} - \widetilde{\mathbf{x}}^{[k-1]})\right] + {}^{1}\!/_{L} A^{\mathrm{T}}\mathbf{b} \\ & = & (I - {}^{1}\!/_{L} A^{\mathrm{T}} A)\left[(\frac{t^{[k]}-1}{t^{[k+1]}} + 1)((\widetilde{D}^{[k]})^2 \widetilde{\mathbf{w}}^{[k]} - {}^{\lambda}\!/_{L} \widetilde{\mathbf{d}}^{[k]})\right] \\ & & -(I - {}^{1}\!/_{L} A^{\mathrm{T}} A)\left[\frac{t^{[k]}-1}{t^{[k+1]}}((\widetilde{D}^{[k-1]})^2 \widetilde{\mathbf{w}}^{[k-1]} - {}^{\lambda}\!/_{L} \widetilde{\mathbf{d}}^{[k-1]})\right] + {}^{1}\!/_{L} A^{\mathrm{T}}\mathbf{b} \\ & = & (1 + \frac{t^{[k]}-1}{t^{[k+1]}})\left[(I - {}^{1}\!/_{L} A^{\mathrm{T}} A)(\widetilde{D}^{[k]})^2\right] \widetilde{\mathbf{w}}^{[k]} \\ & & -\frac{t^{[k]}-1}{t^{[k+1]}}\left[(I - {}^{1}\!/_{L} A^{\mathrm{T}} A)(\widetilde{D}^{[k-1]})^2\right] \widetilde{\mathbf{w}}^{[k-1]} \\ & & +(I - {}^{1}\!/_{L} A^{\mathrm{T}} A)\left[-(1 + \frac{t^{[k]}-1}{t^{[k+1]}}){}^{\lambda}\!/_{L} \widetilde{\mathbf{d}}^{[k]} + \frac{t^{[k]}-1}{t^{[k+1]}}{}^{\lambda}\!/_{L} \widetilde{\mathbf{d}}^{[k-1]}\right] + {}^{1}\!/_{L} A^{\mathrm{T}}\mathbf{b} \\ & = & (1 + \tau^{[k]})\widetilde{R}^{[k]} \widetilde{\mathbf{w}}^{[k]} - \tau^{[k]} \widetilde{R}^{[k-1]} \widetilde{\mathbf{w}}^{[k-1]} + \widetilde{\mathbf{h}} \\ & = & P^{[k]} \widetilde{\mathbf{w}}^{[k]} + Q^{[k-1]} \widetilde{\mathbf{w}}^{[k-1]} + \widetilde{\mathbf{h}}^{[k]} \end{array} \tag{3.9}$$

where we denote

$$\left\{ \begin{array}{rcl} \tau^{[k]} & = & \frac{t^{[k]}-1}{t^{[k+1]}} \\ P^{[k]} & = & (1 + \tau^{[k]})\widetilde{R}^{[k]} \\ Q^{[k]} & = & -\tau^{[k+1]} \widetilde{R}^{[k]} \\ \widetilde{R}^{[k]} & = & (I - {}^{1}\!/_{L} A^{\mathrm{T}} A)(\widetilde{D}^{[k]})^2 \\ \widetilde{\mathbf{h}}^{[k]} & = & (I - {}^{1}\!/_{L} A^{\mathrm{T}} A)\left[-(1 + \tau^{[k]}){}^{\lambda}\!/_{L} \widetilde{\mathbf{d}}^{[k]} + \tau^{[k]}\,{}^{\lambda}\!/_{L} \widetilde{\mathbf{d}}^{[k-1]}\right] + {}^{1}\!/_{L} A^{\mathrm{T}}\mathbf{b} \end{array} \right. \tag{3.10}$$

in the rest of this paper. Note that $R^{[k]}$ in (3.4) refers to the mapping at the $k$-th iteration of ISTA while $\widetilde{R}^{[k]}$ in (3.10) refers to the mapping that would occur if one took one step of ISTA starting at the $k$-th iterate of FISTA. Note also that if $\widetilde{\mathbf{d}}^{[k]} = \widetilde{\mathbf{d}}^{[k-1]}$, then $\widetilde{\mathbf{h}}^{[k]} = (I - {}^{1}\!/_{L} A^{\mathrm{T}} A)\left[{}^{\lambda}\!/_{L} \widetilde{\mathbf{d}}^{[k]}\right] + {}^{1}\!/_{L} A^{\mathrm{T}}\mathbf{b}$ does not vary with $\tau$. For the purposes of analysis, the modified FISTA iteration then can be equivalently expressed as in Alg. 4.

Algorithm 4: One pass of modified FISTA

start with $\widetilde{\mathbf{w}}^{[k-1]}$, $\widetilde{\mathbf{w}}^{[k]}$, $t^{[k]}$, $\widetilde{D}^{[k-1]}$ and $\widetilde{D}^{[k]}$.

1. $\widetilde{\mathbf{w}}^{[k+1]} = P^{[k]}\widetilde{\mathbf{w}}^{[k]} + Q^{[k-1]}\widetilde{\mathbf{w}}^{[k-1]} + \widetilde{\mathbf{h}}^{[k]}$ (with $P^{[k]}, Q^{[k-1]}, \widetilde{\mathbf{h}}^{[k]}$ defined by (3.10)).
2. $t^{[k+1]} = \frac{1+\sqrt{1+4t^{[k]2}}}{2}$ so that $\tau^{[k]} = \frac{t^{[k]}-1}{t^{[k+1]}}$.
3. $\widetilde{D}^{[k+1]} = \mathrm{Diag}\left(\mathrm{Sign}\left(\mathrm{Shr}_{\lambda/L}(\widetilde{\mathbf{w}}^{[k+1]})\right)\right)$.

Result is $\widetilde{\mathbf{w}}^{[k]}$, $\widetilde{\mathbf{w}}^{[k+1]}$, $t^{[k+1]}$, $\widetilde{D}^{[k]}$ and $\widetilde{D}^{[k+1]}$ for next pass.

Step 1 of above procedure can also be formulated as a homogeneous matrix recurrence analogous to (3.5) for ISTA, but with a larger (approximately double) dimension:

$$
\begin{aligned}
\begin{pmatrix} \widetilde{\mathbf{w}}^{[k+1]} \\ \widetilde{\mathbf{w}}^{[k]} \\ 1 \end{pmatrix} &= \mathbf{N}_{aug}^{[k]} \begin{pmatrix} \widetilde{\mathbf{w}}^{[k]} \\ \widetilde{\mathbf{w}}^{[k-1]} \\ 1 \end{pmatrix} = \begin{pmatrix} N^{[k]} & \widetilde{\mathbf{h}}_{aug}^{[k]} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{w}}^{[k]} \\ \widetilde{\mathbf{w}}^{[k-1]} \\ 1 \end{pmatrix} \\
&= \begin{pmatrix} P^{[k]} & Q^{[k-1]} & \widetilde{\mathbf{h}}^{[k]} \\ I & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{w}}^{[k]} \\ \widetilde{\mathbf{w}}^{[k-1]} \\ 1 \end{pmatrix}.
\end{aligned} \tag{3.11}
$$

We denote $N^{[k]} = \begin{pmatrix} P^{[k]} & Q^{[k-1]} \\ I & 0 \end{pmatrix}$ and $\widetilde{\mathbf{h}}_{aug}^{[k]} = \begin{pmatrix} \widetilde{\mathbf{h}}^{[k]} \\ 0 \end{pmatrix}$ such that $\mathbf{N}_{aug}^{[k]} = \begin{pmatrix} N^{[k]} & \widetilde{\mathbf{h}}_{aug}^{[k]} \\ 0 & 1 \end{pmatrix}$ in the remainder of this paper.

Analogous to Lemma 3.1 for ISTA, the following lemma shows the equivalence between the fixed point of the constructed matrix recurrence (3.11) and the KKT point of the LASSO problem.

Lemma 3.4. *Suppose* $\begin{pmatrix} \widetilde{\mathbf{w}}_1 \\ \widetilde{\mathbf{w}}_2 \\ 1 \end{pmatrix}$ *is an eigenvector corresponding to eigenvalue 1 of* $\mathbf{N}_{\mathrm{aug}}$ *(omitting* $[k]$*) in* (3.11)*, then* $\widetilde{\mathbf{w}}_1 = \widetilde{\mathbf{w}}_2 := \widetilde{\mathbf{w}}$*. Suppose further that* $\widetilde{D}^{[k+1]} = \widetilde{D}^{[k]} = \widetilde{D} = \mathrm{Diag}\,(\widetilde{\mathbf{d}})$ *with entries* $\widetilde{d}_i = 1$ *if* $\widetilde{w}_i > \lambda/L$, $\widetilde{d}_i = -1$ *if* $\widetilde{w}_i < -\lambda/L$ *and* $\widetilde{d}_i = 0$ *if* $-\lambda/L \le \widetilde{w}_i \le \lambda/L$*. Then the variables defined by* $\widetilde{\mathbf{x}} = Shr_{\lambda/L}(\widetilde{\mathbf{w}})$ *satisfies the* $1$*st order KKT conditions* (2.1) *and* (2.2)*. Conversely, if* $\widetilde{\mathbf{x}}$ *and* $\widetilde{\boldsymbol{\nu}} = 1/\lambda A^{\mathrm{T}}(\mathbf{b} - A\widetilde{\mathbf{x}})$ *satisfy the KKT conditions, then* $\begin{pmatrix} \widetilde{\mathbf{w}} \\ \widetilde{\mathbf{w}} \\ 1 \end{pmatrix}$*, with* $\widetilde{\mathbf{w}} = \widetilde{\mathbf{x}} + \lambda/L\widetilde{\boldsymbol{\nu}}$*, is an eigenvector of* $\mathbf{N}_{\mathrm{aug}}$ *corresponding to eigenvalue 1, where* $\mathbf{N}_{\mathrm{aug}}$ *is defined in* (3.11) *and* $\widetilde{D}^{[k+1]} = \widetilde{D}^{[k]} = \widetilde{D} = \mathrm{Diag}\,(\widetilde{\mathbf{d}})$ *with entries* $\widetilde{d}_i = 1$ *if* $\widetilde{w}_i > \lambda/L$, $\widetilde{d}_i = -1$ *if* $\widetilde{w}_i < -\lambda/L$ *and* $\widetilde{d}_i = 0$ *if* $-\lambda/L \le \widetilde{w}_i \le \lambda/L$.

*Proof.* It is easy to show that a vector is a fixed point for FISTA if and only if it is a fixed point for ISTA, so this follows from Lemma 3.1. □

To prepare for further discussion, we make three remarks.

(a). $\tau^{[k]} \longrightarrow 1$ from below as $k \longrightarrow \infty$.

(b). $R^{[k]} = \widetilde{R}^{[k]}$ if the flag matrix of ISTA and FISTA are the same, i.e. $\widehat{D}^{[k]} = \widetilde{D}^{[k]}$ and $\mathbf{h}^{[k]} = \widetilde{\mathbf{h}}^{[k]}$ if $\widehat{D}^{[k]} = \widetilde{D}^{[k]} = \widetilde{D}^{[k-1]}$. This observation relates the $\mathbf{R}_{aug}^{[k]}$ to $\mathbf{N}_{aug}^{[k]}$. It is the foundation upon which we establish the properties to compare ISTA and FISTA in Section 6.

(c). One main difference between operators of ISTA and FISTA (i.e. $\mathbf{R}_{\text{aug}}^{[k]}$ and $\mathbf{N}_{\text{aug}}^{[k]}$) is that $\mathbf{R}_{\text{aug}}^{[k]}$ is fixed when the flag matrix is fixed while $\mathbf{N}_{\text{aug}}^{[k]}$ changes at each step $k$ even if the flag matrix is fixed. In other words, for all $k$, $\mathbf{R}_{\text{aug}}^{[k]} = \mathbf{R}_{\text{aug}}^{[k+1]}$ if $\widehat{D}^{[k]} = \widehat{D}^{[k+1]}$. But $\mathbf{N}_{\text{aug}}^{[k]} \neq \mathbf{N}_{\text{aug}}^{[k+1]}$ even if $\widetilde{D}^{[k]} = \widetilde{D}^{[k+1]}$. The reason is that $\mathbf{N}_{\text{aug}}^{[k]}$ depends on the changing stepsize $\tau^{[k]}$. Nevertheless, one can still use the same similar argument for $\mathbf{N}_{\text{aug}}^{[k]}$ as for $\mathbf{R}_{\text{aug}}^{[k]}$ by additional lemmas, as we will show in Section 5.

**3.4. Spectral Properties of $\mathbf{N}_{\text{aug}}^{[k]}$.** We give the spectral properties of the FISTA matrix operator that we will use in our convergence analysis. Lemma 3.5 give the eigenstructure of $N^{[k]}$ and its relation to that of the ISTA operator $\widetilde{R}^{[k]}$.

LEMMA 3.5. *Suppose $\widetilde{D}^{[k]} = \widetilde{D}^{[k-1]}$ and hence $\widetilde{R}^{[k]} = \widetilde{R}^{[k-1]}$, then (omitting index $^{[k]}$) we have the following results:*

(a). *For any given eigenvalue $\gamma$ of $N$, it must have a corresponding eigenvector with the form of $\begin{pmatrix}\gamma\widetilde{\mathbf{w}}\\ \widetilde{\mathbf{w}}\end{pmatrix}$. And for that given $\gamma$, there exists an eigenvalue of $\widetilde{R}$, $\beta = \frac{\gamma^2}{[(1+\tau)\gamma-\tau]}$ with corresponding eigenvector $\widetilde{\mathbf{w}}$. Conversely, let $\beta$ be any eigenvalue of $\widetilde{R}$ with corresponding eigenvector $\widetilde{\mathbf{w}}$. Then for given $\beta$, there exists a pair of eigenvalues of $N$, $\gamma_1$ and $\gamma_2$, which are the solutions of $\gamma^2 - \gamma(1+\tau)\beta + \tau\beta = 0$. Furthermore, $\begin{pmatrix}\gamma_1\widetilde{\mathbf{w}}\\ \widetilde{\mathbf{w}}\end{pmatrix}$ and $\begin{pmatrix}\gamma_2\widetilde{\mathbf{w}}\\ \widetilde{\mathbf{w}}\end{pmatrix}$ are two corresponding eigenvectors of $N$.*

(b). *For $0 < \tau \leq 1$, the eigenvalues of $N$ defined in (3.11) lie in the closed disk in the complex plane with center $1/2$ and radius $1/2$, denoted as $\mathcal{D}(1/2, 1/2)$. In particular, if $N$ has any eigenvalue with absolute value $\rho(N) = 1$, then that eigenvalue must be exactly 1.*

(c). *$N$ has an eigenvalue equal to 1 if and only if $\widetilde{R}$ has an eigenvalue equal to 1.*

(d). *Assuming $\tau < 1$, then if $N$ has an eigenvalue equal to 1, this eigenvalue must have a complete set of eigenvectors.*

*Proof.* (a). Given any eigenvalue $\gamma$ of $N$, by definition (just after (3.11))

$$N \cdot \begin{pmatrix}\widetilde{\mathbf{w}}_1\\ \widetilde{\mathbf{w}}_2\end{pmatrix} = \begin{pmatrix}P & Q\\ I & 0\end{pmatrix}\begin{pmatrix}\widetilde{\mathbf{w}}_1\\ \widetilde{\mathbf{w}}_2\end{pmatrix} = \begin{pmatrix}P\widetilde{\mathbf{w}}_1 + Q\widetilde{\mathbf{w}}_2\\ \widetilde{\mathbf{w}}_1\end{pmatrix} = \gamma\begin{pmatrix}\widetilde{\mathbf{w}}_1\\ \widetilde{\mathbf{w}}_2\end{pmatrix}$$

and thus $\widetilde{\mathbf{w}}_1 = \gamma\widetilde{\mathbf{w}}_2$ is observed from the second row. Then

$$N \cdot \begin{pmatrix}\gamma\widetilde{\mathbf{w}}_2\\ \widetilde{\mathbf{w}}_2\end{pmatrix} = \begin{pmatrix}\gamma P\widetilde{\mathbf{w}}_2 + Q\widetilde{\mathbf{w}}_2\\ \gamma\widetilde{\mathbf{w}}_2\end{pmatrix} = \begin{pmatrix}(1+\tau)\gamma\widetilde{R}\widetilde{\mathbf{w}}_2 - \tau\widetilde{R}\widetilde{\mathbf{w}}_2\\ \gamma\widetilde{\mathbf{w}}_2\end{pmatrix} = \gamma\begin{pmatrix}\gamma\widetilde{\mathbf{w}}_2\\ \widetilde{\mathbf{w}}_2\end{pmatrix}.$$

Therefore,

$$\widetilde{R}\widetilde{\mathbf{w}}_2 = \frac{\gamma^2}{[(1+\tau)\gamma-\tau]}\widetilde{\mathbf{w}}_2 = \beta\widetilde{\mathbf{w}}_2 \iff \gamma^2 - (1+\tau)\gamma\beta + \tau\beta = 0. \tag{3.12}$$

(b). We first study the spectrum of matrix $N - 1/2 I$, then the spectrum of $N$ should be a shift by $1/2$. Let $\alpha = \gamma - 1/2$ be the eigenvalue of $N - 1/2 I$ associated with eigenvector $\begin{pmatrix}\widetilde{\mathbf{w}}_1\\ \widetilde{\mathbf{w}}_2\end{pmatrix}$, then according to (3.12), $\alpha$ and $\beta$ satisfy

$$\alpha^2 + (1 - \beta - \tau\beta)\alpha + 1/2\tau\beta - 1/2\beta + 1/4 = 0.$$

Note that $\tau \in (0,1]$ and $\beta \in [0,1]$ by definition of $\widetilde{R}$. We consider two situations for the above quadratic equation. First, suppose $\alpha_1$ and $\alpha_2$ are two conjugate complex roots. Then $\alpha_1 = \bar{\alpha_2}$, $\alpha_1 + \alpha_2 = \tau\beta + \beta - 1$ and $\alpha_1\alpha_2 = {}^1\!/_2\tau\beta - {}^1\!/_2\beta + {}^1\!/_4$ such that

$$|\alpha|^2 = |\alpha_1\bar{\alpha_1}| = |\alpha_1\alpha_2| = \left|{}^1\!/_2\tau\beta - {}^1\!/_2\beta + {}^1\!/_4\right| \le \frac{1}{4} \tag{3.13}$$

which gives $|\alpha| \le {}^1\!/_2$. The second situation is that both roots are real numbers and they are

$$\alpha_1 = \frac{1+\tau}{2}\beta + \frac{\sqrt{\beta}\sqrt{(1+\tau)^2\beta - 4\tau}}{2} - {}^1\!/_2 \quad \alpha_2 = \frac{1+\tau}{2}\beta - \frac{\sqrt{\beta}\sqrt{(1+\tau)^2\beta - 4\tau}}{2} - {}^1\!/_2.$$

To get the largest possible value of $\alpha$, we only look at $\alpha_1$ because $\alpha_1 \ge \alpha_2$ for any fixed $\beta$. Since $\alpha_1$ is an increasing function of $\beta$ and $\beta \in [0,1]$, the largest real value of $\alpha$ should be ${}^1\!/_2$ when $\beta = 1$. On the other hand, to get the smallest negative real value of $\alpha$, we only need to look at $\alpha_2$. One can write $\alpha_2 = \frac{1+\tau}{2}(\beta - \sqrt{\beta^2 - \frac{4\tau}{(1+\tau)^2}}) - {}^1\!/_2$ to see that $\alpha_2 \ge -{}^1\!/_2$. So we conclude that if $\alpha$ is real, then $-{}^1\!/_2 \le \alpha \le {}^1\!/_2$.

Under both situations, one can conclude that $\alpha$ must satisfy $|\alpha| \le {}^1\!/_2$, lying in a disk centered at 0 with radius ${}^1\!/_2$, i.e. $\mathcal{D}(0, {}^1\!/_2)$. So the eigenvalues of $N$ must lie in the disk $\mathcal{D}({}^1\!/_2, {}^1\!/_2)$ by the shift. Consequently, the only possible eigenvalue on the unit circle is 1.

(c). $\gamma_1, \gamma_2$ are the two roots of the quadratic polynomial, i.e. $\gamma^2 - (1+\tau)\gamma\beta + \tau\beta = (\gamma - \gamma_1)(\gamma - \gamma_2) = 0$. For given $\beta$, they must satisfy

$$\gamma_1\gamma_2 = \tau\beta \quad \text{and} \quad \gamma_1 + \gamma_2 = (1+\tau)\beta = \beta + \gamma_1\gamma_2.$$

If $N$ has an eigenvalue $\gamma_1 = 1$, then $\gamma_2 = (1+\tau)\beta - 1 = \beta + \gamma_1\gamma_2 - \gamma_1 = \beta + \gamma_2 - 1$, hence $\beta = 1$ must be true and $\widetilde{R}$ has an eigenvalue equal to 1. Conversely, if $\widetilde{R}$ has an eigenvalue $\beta = 1$, the quadratic polynomial (3.12) will reduce to $\gamma^2 - (1+\tau)\gamma + \tau = 0$, which gives $\gamma_1 = 1$ and $\gamma_2 = \tau$. Then $N$ has an eigenvalue equal to 1.

(d). Notice in (3.12) that each eigenvalue $\beta$ of $\widetilde{R}$ maps to two eigenvalues of $N$, $\gamma_1$ and $\gamma_2$, and associated eigenvector $\widetilde{\mathbf{w}}_2$ of $\widetilde{R}$ maps to two eigenvectors of $N$, $\begin{pmatrix}\gamma_1\widetilde{\mathbf{w}}_2 \\ \widetilde{\mathbf{w}}_2\end{pmatrix}$ and $\begin{pmatrix}\gamma_2\widetilde{\mathbf{w}}_2 \\ \widetilde{\mathbf{w}}_2\end{pmatrix}$. As shown in (c), $N$ has an eigenvalue equal to 1 if and only if $\widetilde{R}$ has an eigenvalue equal to 1. Since $R$ and $\widetilde{R}$ have the similar eigenstructure, eigenvalue 1 of $\widetilde{R}$ must have a complete set of eigenvectors. So the only possible situation that eigenvalue 1 of $N$ does not have a complete set of eigenvectors is that both $\gamma_1$ and $\gamma_2$ equal to 1. However, this is impossible because we have shown in (c) that $\beta = 1$ gives $\gamma_1 = 1$ and $\gamma_2 = \tau$ which is close to 1 but not equal. As a result, if $N$ has an eigenvalue 1, and then its algebraic and geometric multiplicities coincide. $\square$

**4. Regimes.** Since the ISTA and FISTA updating steps have been converted into a variation of an eigenproblem in previous sections, we can study the convergence in terms of the spectral properties of the operator $\mathbf{R}_{aug}^{[k]}$ in (3.5) and $\mathbf{N}_{aug}^{[k]}$ in (3.11). Hence in this section, we show how these properties reflected in the possible convergence "regimes" that ISTA and FISTA can encounter.

**4.1. Spectral Properties.** The eigenvalues of the augmented matrices $\mathbf{R}_{aug}^{[k]}$ and $\mathbf{N}_{aug}^{[k]}$ consist of those of $R^{[k]}$, $N^{[k]}$, respectively, plus an extra eigenvalue equal to

1. If $R^{[k]}$ (or $N^{[k]}$) already has an eigenvalue equal to 1, then the extra eigenvalue 1 may or may not add a corresponding eigenvector. The next lemma gives limits on the properties of the eigenalue 1 for any augmented matrix of the general form of $\mathbf{R}^{[k]}_{aug}$, $\mathbf{N}^{[k]}_{aug}$.

LEMMA 4.1. *Let* $\mathbf{M}_{\text{aug}} = \begin{pmatrix} M & \mathbf{p} \\ 0 & 1 \end{pmatrix}$ *be any block upper triangular matrix with a* $1 \times 1$ *lower right block. The matrix* $\mathbf{M}_{\text{aug}}$ *has an eigenvalue* $\alpha_1 = 1$ *and suppose its corresponding eigenvector has a non-zero last element. We scale that eigenvector to take the form* $\begin{pmatrix} \mathbf{w} \\ 1 \end{pmatrix} = \mathbf{M}_{\text{aug}} \begin{pmatrix} \mathbf{w} \\ 1 \end{pmatrix}$. *If the upper left block* $M$ *either has no eigenvalue equal to* 1 *or the eigenvalue* 1 *of* $M$ *has a complete set of eigenvectors, then* $\alpha_1 = 1$ *has no non-trivial Jordan block. Furthermore, if the given eigenvector* $\begin{pmatrix} \mathbf{w} \\ 1 \end{pmatrix}$ *is unique, then* $M$ *has no eigenvalue equal to 1.*

We refer readers to [3] for the proof of Lemma 4.1. Now we summarize spectral properties of our specific operators $\mathbf{R}^{[k]}_{aug}$ and $\mathbf{N}^{[k]}_{aug}$ in terms of their possible Jordan canonical forms as given in the following lemmas. Essentially these lemmas say that all their eigenvalues must have absolute value strictly less than 1, except for the eigenvalue equal to 1. And the eigenvalue 1's geometric multiplicity either equal to or one less than its algebraic multiplicity.

LEMMA 4.2. *Assuming* $\widehat{D}^{[k+1]} = \widehat{D}^{[k]}$, *then* $\mathbf{R}^{[k]}_{\text{aug}}$ *in* (3.5) *is fixed and has a spectral decomposition* $\mathbf{R}^{[k]}_{\text{aug}} = \mathbf{P}_R \mathbf{J}^{[k]}_R \mathbf{P}_R^{-1}$, *where* $\mathbf{J}^{[k]}_R$ *is a block diagonal matrix*

$$\mathbf{J}^{[k]}_R = \text{Diag}\left\{ \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, I^{[k]}_R, \widehat{\mathbf{J}}^{[k]}_R \right\} \tag{4.1}$$

*where any of these blocks might be missing. Here* $I^{[k]}_R$ *is an identity matrix and* $\widehat{\mathbf{J}}^{[k]}_R$ *is a matrix with spectral radius strictly less than* 1.

*Proof.* For $\mathbf{R}^{[k]}_{\text{aug}}$, the upper left block of (3.5) (i.e. $R^{[k]}$) satisfies Lemma 3.3 and hence contributes blocks of the form $I^{[k]}_R$, $\widehat{\mathbf{J}}^{[k]}_R$. No eigenvalue with absolute value 1 can have a nondiagonal Jordan block, so the blocks corresponding to those eigenvalues must be diagonal. Embedding that upper left block $R^{[k]}$ into the entire matrix yields a matrix $\mathbf{R}^{[k]}_{\text{aug}}$, with the exact same set of eigenvalues with the same algebraic and geometric multiplicities, except for eigenvalue 1.

If the upper left block of $\mathbf{R}^{[k]}_{\text{aug}}$ has no eigenvalue equal to 1, then $\mathbf{R}^{[k]}_{\text{aug}}$ has a simple eigenvalue 1. In general for eigenvalue 1, the algebraic multiplicity goes up by one and the geometric multiplicity can either stay the same or increase by 1. In other words, $\mathbf{R}^{[k]}_{\text{aug}}$ either satisfies the conditions of Lemma 4.1, or the algebraic and geometric multiplicities of eigenvalue 1 for $\mathbf{R}^{[k]}_{\text{aug}}$ differ by 1, meaning we have a single $2 \times 2$ Jordan block $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. $\square$

LEMMA 4.3. *Assuming* $\widetilde{D}^{[k+1]} = \widetilde{D}^{[k]}$, *since* $\mathbf{N}^{[k]}_{\text{aug}}$ *in* (3.11) *is different at each step, for each step* $k$, *there exists a* $\mathbf{P}^{[k]}_N$ *such that* $\mathbf{N}^{[k]}_{\text{aug}}$ *has a spectral decomposition* $\mathbf{N}^{[k]}_{\text{aug}} = \mathbf{P}^{[k]}_N \mathbf{J}^{[k]}_N (\mathbf{P}^{[k]}_N)^{-1}$, *where* $\mathbf{J}^{[k]}_N$ *is the block diagonal matrix:*

$$\mathbf{J}^{[k]}_N = \text{Diag}\left\{ \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, I^{[k]}_N, \widetilde{\mathbf{J}}^{[k]}_N \right\} \tag{4.2}$$

*where any of these blocks might be missing. Here $I_N^{[k]}$ is an identity matrix, $\widetilde{\mathbf{J}}_N^{[k]}$ is a matrix with spectral radius strictly less than* 1.

*Proof.* The proof is similar to $\mathbf{R}_{aug}^{[k]}$. We only note here that the upper left block of (3.11) (i.e. $N^{[k]}$) satisfies Lemma 3.5 and hence contributes blocks of the form $I_N^{[k]}$, $\widetilde{\mathbf{J}}_N^{[k]}$. $\square$

**4.2. Four regimes.** Lemmas 4.2 & 4.3 give rise to the four possible "regimes" associated with the ISTA and FISTA iterations, depending on the flag matrix and the eigenvalues of $\mathbf{R}_{aug}^{[k]}$, $\mathbf{N}_{aug}^{[k]}$. We treat separately the case where the flag matrix remains the same at each iteration, in which there are three possible regimes, and treat all the transitional cases together in their own fourth regime. For simplicity of the notation, we let $D$ denote the flag matrix instead of $\widehat{D}$ and $\widetilde{D}$ unless specified.

When the flag matrix does not change, i.e. $D^{[k+1]} = D^{[k]}$, the ISTA operator $\mathbf{R}_{aug}^{[k]}$ remains invariant over those passes while the FISTA operator $\mathbf{N}_{aug}^{[k]}$ is slightly different at each iteration due to the changing parameter $\tau^{[k]}$. In both cases, the convergence behavior of the algorithm should depend on the eigenvalue of its corresponding operator, which can be categorized into three situations. In the following, we specifically describe these three possible regimes distinguished by the eigenstructure of the operators $\mathbf{R}_{aug}^{[k]}$, $\mathbf{N}_{aug}^{[k]}$. One of the these three regimes must occur when the flag matrix is unchanged from one step to the next: $D^{[k+1]} = D^{[k]}$.

[A]. The spectral radius of $R^{[k]}$ (or $N^{[k]}$) is strictly less than 1. The block $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is absent from (4.1) (or (4.2)), and the block $I_R^{[k]}$ (or $I_N^{[k]}$) is $1\times 1$. In the case where the optimal solution exists and is unique, the result is linear convergence to the solution when close enough to that solution, as we will show in Theorem 5.3 & 5.5. For $\mathbf{R}_{aug}^{[k]}$, as long as the flags remain the same, the recurrence (3.5) hence will converge linearly to a unique fixed point at a rate determined by the next largest eigenvalue in absolute value (largest eigenvalue of the block $\widehat{\mathbf{J}}_R^{[k]}$), according to the theory of the power method. If the flags $\widehat{D}^{[k]}$ are consistent with the eigenvector satisfying (3.2), then the eigenvector must satisfy the KKT conditions (2.1) and (2.2) because of Lemma 3.1.

For $\mathbf{N}_{aug}^{[k]}$, though it changes slightly at each step, we will show in the case of a unique solution that the left and right eigenvectors for eigenvalue 1 do not depend on $\tau$ (Lemma 5.4), and the remaining eigenvalues remain smaller and bounded away from 1. The result is that we observe linear convergence to a unique fixed point with a slightly changing convergence rate. If the flags $\widetilde{D}^{[k]}$ are consistent with the eigenvector satisfying (3.8), then that eigenvector must satisfy the KKT conditions because of Lemma 3.4.

[B]. $R^{[k]}$ (or $N^{[k]}$) has an eigenvalue equal to 1 which results in a $2\times 2$ Jordan block $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ for $\mathbf{R}_{aug}^{[k]}$ (or $\mathbf{N}_{aug}^{[k]}$). Therefore, the iteration process tends to a constant step.

For $\mathbf{R}_{aug}^{[k]}$, the theory of power method implies that the vector iterates will converge to the invariant subspace corresponding to the largest eigenvalue 1. The presence of $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ means that there is a Jordan chain: two non-zero vectors $\mathbf{q}, \mathbf{r}$ such that $(\mathbf{R}_{aug} - I)\mathbf{q} = \mathbf{r}$, $(\mathbf{R}_{aug} - I)\mathbf{r} = 0$. Any vector which includes a component of the form $\alpha\mathbf{q} + \beta\mathbf{r}$ will be transformed into $\mathbf{R}_{aug}(\alpha\mathbf{q} + \beta\mathbf{r}) = \alpha\mathbf{q} + (\alpha+\beta)\mathbf{r}$, i.e. each pass would add a constant vector $\alpha\mathbf{r}$, plus fading lower order terms from the other lesser eigenvalues. As long as the flags do not change, this will result in constant steps: the difference between consecutive iterates, $\begin{pmatrix}\widehat{\mathbf{w}}^{[k+1]}\\ 1\end{pmatrix} - \begin{pmatrix}\widehat{\mathbf{w}}^{[k]}\\ 1\end{pmatrix}$, will converge to a constant vector, asymptotically as the effects of the smaller eigenvalues fade. That constant vector is an eigenvector for eigenvalue 1. The ISTA iteration will not converge until a flag change in $\widehat{\mathbf{w}}^{[k]}$ forces a change in the flags $\widehat{D}^{[k]}$. If we satisfy the conditions for global convergence of ISTA, then such a flag change is guaranteed to occur.

The same analysis applies to $\mathbf{N}_{aug}^{[k]}$. The difference between two consecutive iterates, $\begin{pmatrix}\widetilde{\mathbf{w}}^{[k+1]}\\ 1\end{pmatrix} - \begin{pmatrix}\widetilde{\mathbf{w}}^{[k]}\\ 1\end{pmatrix}$, will asyptotically converge to a constant vector. The FISTA iteration will not converge until a flag change in $\widetilde{\mathbf{w}}^{[k]}$ forces a change in the flags $\widetilde{D}^{[k]}$. Such a flag change is guaranteed to occur due to the global convergence of FISTA under the assumption of unique solution. In Section 6.1.1, we will show that FISTA can jump out of such regime very fast, which is the main reason why it is faster than ISTA. See Section 7 for more discussions on its numerical behavior.

[C]. $R^{[k]}$ (or $N^{[k]}$) has an eigenvalue equal to 1, but the block $\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$ is absent. For $\mathbf{R}_{aug}^{[k]}$ (or $\mathbf{N}_{aug}^{[k]}$), the convergence rate of this regime will still depend on $\rho(\widehat{\mathbf{J}}_R^{[k]})$ and $\rho(\widetilde{\mathbf{J}}_N^{[k]})$. If we assume the solution to be unique, the eigenvalue 1 of $\mathbf{R}_{aug}^{[k]}$ (or $\mathbf{N}_{aug}^{[k]}$) must be simple by Lemma 4.1. So the iteration will eventually jump out of this type of regime.

When the flag matrix does change, it means the set of active constraints at the current pass in the process has changed, and the current pass is a transition to a different operator with a different eigenstructure.

[D]. The operator $R^{[k+1]}$ (or $N^{[k+1]}$) will be different from $R^{[k]}$ (or $N^{[k]}$) due to different flag matrix.

**5. Unique Solution: Linear Convergence.** In the case that (1.1) has a unique solution with strict complementarity (§2.2), we can guarantee that eventually the flag matrix will not change. Once the flag matrix stays fixed, the ISTA (or FISTA) iteration behaves just like the power method (or similar to power method) for the matrix eigenvalue problem. In this case, the spectral theory developed here gives a guarantee of linear convergence with the rate equal to the second largest eigenvalue of the matrix operator. In this section, we denote the fixed point of the respective matrix recursions (3.5) and (3.11) as

$$\widehat{\mathbf{w}}_{aug}^* = \begin{pmatrix}\widehat{\mathbf{w}}^*\\ 1\end{pmatrix} \text{ and } \widetilde{\mathbf{w}}_{aug}^* = \begin{pmatrix}\widetilde{\mathbf{w}}^*\\ \widetilde{\mathbf{w}}^*\\ 1\end{pmatrix}, \tag{5.1}$$

where in the final regime, $\widehat{\mathbf{w}}^* = \widetilde{\mathbf{w}}^* = \mathbf{w}^* = \mathbf{x}^* + {}^{\lambda}\!/_{L}\boldsymbol{\nu}^*$, with $\mathbf{x}^*, \boldsymbol{\nu}^*$ the unique solution satisfying (2.1) (2.2).

**5.1. Global Convergence Theory.** We first invoke the global convergence property of ISTA and FISTA.

THEOREM 5.1. *If problem* (1.1) *is solvable, let* $F^* = \min_{\mathbf{x}} F(\mathbf{x})$*, where* $F(\mathbf{x}) = {}^{1}\!/_{2}\|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1$*. Let* $\widehat{\mathbf{x}}^{[0]}$ *be any starting point in* $\mathbb{R}^n$ *and* $\widehat{\mathbf{x}}^{[k]}$ *be the sequence generated by ISTA. Then for any* $k \geq 1$*,* $F(\widehat{\mathbf{x}}^{[k]}) - F^*$ *decreases to zero at a rate bounded by* $O({}^{1}\!/_{k})$*. On the other hand, if we let* $\mathbf{y}^{[1]} = \widetilde{\mathbf{x}}^{[0]}$ *be any starting point in*

$\mathbb{R}^n$, $t^{[0]} = t^{[1]} = 1$ *and* $\{\widetilde{\mathbf{x}}^{[k]}\}, \{\mathbf{y}^{[k]}\}, \{t^{[k]}\}$ *be the sequence generated by FISTA, then for any* $k \geq 1$, $F(\widetilde{\mathbf{x}}^{[k]}) - F^*$ *decreases to zero at a rate bounded by* $O({}^1\!/_{k^2})$.

This is a restatement of the convergence theorem in [1]. It says little on the local behavior of the algorithm. Under the assumption of unique solution with strict complementarity, we can prove the specific result that ISTA and FISTA iterations must eventually reach and remain in "linear convergence" regime [A] so that the optimal flag matrix is identified.

LEMMA 5.2. *Suppose the LASSO problem* (1.1) *has a unique solution and this solution has strict complementarity (§2.2). Then eventually the ISTA (FISTA) iteration reaches and remains in the final regime where optimal flag matrix is identified.*

*Proof.* For ISTA, by Lemma 3.1, the strict complementarity is equivalent to $\widehat{w}_i^* \neq \pm{}^\lambda\!/_L$. Note that $\widehat{w}_i^{[k]}$ (where $k$ is the pass number) could only be in one of three cases: $\widehat{w}_i^{[k]} > {}^\lambda\!/_L$, $\widehat{w}_i^{[k]} < -{}^\lambda\!/_L$ or $-{}^\lambda\!/_L \leq \widehat{w}_i^{[k]} \leq {}^\lambda\!/_L$. We can let $C_1 = \min\{|\widehat{w}_i^* - {}^\lambda\!/_L| - \epsilon_1, |\widehat{w}_i^* + {}^\lambda\!/_L| - \epsilon_1\} > 0$ for a positive constant $\epsilon_1$ sufficiently small to make $C_1 > 0$. We define the following ball around eigenvector $\widehat{\mathbf{w}}_{aug}^*$:

$$\mathcal{B}_1 = \{\widehat{\mathbf{w}}_{aug} : \|\widehat{\mathbf{w}}_{aug} - \widehat{\mathbf{w}}_{aug}^*\|_\infty \leq C_1\}. \tag{5.2}$$

By Theorem 5.1 and uniqueness of the solution, the iterates $\widehat{\mathbf{w}}_{\text{aug}}^{[k]}$ converge to the $\widehat{\mathbf{w}}_{aug}^*$. Therefore, there exists a pass $K_1$ such that $\widehat{\mathbf{w}}_{aug}^{[k]}$ lies in $\mathcal{B}_1$ (i.e. $\|\widehat{\mathbf{w}}_{\text{aug}}^{[k]} - \widehat{\mathbf{w}}_{\text{aug}}^*\|_\infty \leq C_1$) for all $k > K_1$. This means that $\widehat{w}_i^{[k]}$ will remain in one of three cases: $\widehat{w}_i^{[k]} > {}^\lambda\!/_L$, $\widehat{w}_i^{[k]} < -{}^\lambda\!/_L$ or $-{}^\lambda\!/_L \leq \widehat{w}_i^{[k]} \leq {}^\lambda\!/_L$ and will never change to another case for all $k > K_1$. This, combined with the definition of flag matrix $\widehat{D}^{[k]} = \text{Diag}\,(\text{Sign}\,(\text{Shr}_{{}^\lambda\!/_L}(\widehat{\mathbf{w}}^{[k]})))$, implies that the flag matrix remain unchanged for all $k > K$.

Similarly, for FISTA, by Lemma 3.4, the strict complementarity is equivalent to $\widetilde{w}_i^* \neq \pm{}^\lambda\!/_L$. We can let $C_2 = \min\{|\widetilde{w}_i^* - {}^\lambda\!/_L| - \epsilon_2, |\widetilde{w}_i^* + {}^\lambda\!/_L| - \epsilon_2\} > 0$ for a positive constant $\epsilon_2$ sufficiently small to make $C_2 > 0$. We then define the following ball around the optimal eigenvector $\widetilde{\mathbf{w}}_{aug}^*$:

$$\mathcal{B}_2 = \{\widehat{\mathbf{w}}_{aug} : \|\widetilde{\mathbf{w}}_{aug} - \widetilde{\mathbf{w}}_{aug}^*\|_\infty \leq C_2\}. \tag{5.3}$$

By Theorem 5.1 and uniqueness of the solution, the iterates $\widetilde{\mathbf{w}}_{aug}^{[k]}$ converge to $\widetilde{\mathbf{w}}_{aug}^*$. Therefore, there exists a pass $K_2$ such that $\widetilde{\mathbf{w}}_{aug}^{[k]}$ lies in $\mathcal{B}_2$ for all $k > K_2$. Combined with the definition of flag matrix $\widetilde{D}^{[k]} = \text{Diag}\,(\text{Sign}\,(\text{Shr}_{{}^\lambda\!/_L}(\widetilde{\mathbf{w}}^{[k]})))$, this implies that the flag matrix remain unchanged for all $k > K_2$. $\square$

**5.2. Local Linear Convergence of ISTA.** Once the optimal flag matrix is identified at step $K_1$, the iteration matrices $R^{[k]}$ and $\mathbf{R}_{\text{aug}}^{[k]}$ remain fixed for all $k > K_1$. We denote them as $R^*$ and $\mathbf{R}_{\text{aug}}^*$ in this section.

THEOREM 5.3. *Suppose the LASSO problem* (1.1) *has a unique solution and this solution satisfies the strict complementarity. Then eventually the ISTA iteration reaches a stage where it converges linearly to that unique solution.*

*Proof.* As shown in the proof of Lemma 5.2, there exists a pass number $K_1$ such that $\widehat{\mathbf{w}}_{\text{aug}}^{[k]}$ lies in $\mathcal{B}_1$ and $D^{[k]} = D^{[k+1]}$, for all $k > K_1$. Hence for all $k > K_1$, $R^{[k]} = R^*$ and $\mathbf{h}^{[k]} = \mathbf{h}^*$ remain invariant. By Lemma 3.1, the unique solution, if it exists, must correspond to a unique eigenvector of eigenvalue 1 for the matrix $\mathbf{R}_{\text{aug}}^*$. Additionally, by Lemma 4.1, the matrix $R^*$ has no eigenvalue equal to 1, and by Lemma 3.3, all the eigenvalues must be strictly less than 1 in the absolute value. Hence, starting at the $K_1$-th pass, the ISTA iteration reduces to the power method on the constant matrix $\mathbf{R}^*_{aug}$ associated with the optimal flag matrix.

Let the error vector at the $k$-th pass of the power method be

$$\widehat{\mathbf{e}}^{[k]}_{aug} = \widehat{\mathbf{w}}^{[k]}_{aug} - \widehat{\mathbf{w}}^*_{aug} = \begin{pmatrix} \widehat{\mathbf{w}}^{[k]} \\ 1 \end{pmatrix} - \begin{pmatrix} \widehat{\mathbf{w}}^* \\ 1 \end{pmatrix} = \begin{pmatrix} \widehat{\mathbf{e}}^{[k]} \\ 0 \end{pmatrix}.$$

Then the power iteration on $\widehat{\mathbf{w}}^{[k]}_{\text{aug}}$ yields

$$\begin{aligned} \widehat{\mathbf{w}}^{[k+1]}_{\text{aug}} &= \widehat{\mathbf{w}}^*_{aug} + \widehat{\mathbf{e}}^{[k+1]}_{\text{aug}} = \begin{pmatrix} R^*\widehat{\mathbf{w}}^{[k]} + \mathbf{h}^* \\ 1 \end{pmatrix} = \begin{pmatrix} R^*(\widehat{\mathbf{w}}^* + \widehat{\mathbf{e}}^{[k]}) + \mathbf{h}^* \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} \widehat{\mathbf{w}}^* + R^*\widehat{\mathbf{e}}^{[k]} \\ 1 \end{pmatrix} = \widehat{\mathbf{w}}^*_{aug} + \mathbf{R}^*_{aug}\widehat{\mathbf{e}}^{[k]}_{\text{aug}} \end{aligned}$$

with $\widehat{\mathbf{e}}^{[k+1]} = R^* \cdot \widehat{\mathbf{e}}^{[k]}$. According to Theorem 3.2, for any $\epsilon > 0$, there exists a matrix norm $\|\cdot\|_{\widehat{G}}$ such that $\rho(R^*) < \|R^*\|_{\widehat{G}} \leq \rho(R^*) + \epsilon < 1$. Let $\widehat{G}_{aug} = \begin{pmatrix} \widehat{G} & 0 \\ 0 & 1 \end{pmatrix}$, then

$$\|\widehat{\mathbf{e}}^{[k+1]}_{\text{aug}}\|_{\widehat{G}_{aug}} = \|\widehat{\mathbf{e}}^{[k+1]}\|_{\widehat{G}} \leq \|R^*\|_{\widehat{G}}\|\widehat{\mathbf{e}}^{[k+1]}\|_{\widehat{G}} = \|R^*\|_{\widehat{G}}\|\widehat{\mathbf{e}}^{[k]}_{\text{aug}}\|_{\widehat{G}_{aug}}$$

Hence starting from $K_1$-th pass,

$$\|\widehat{\mathbf{e}}^{[k]}_{\text{aug}}\|_{\widehat{G}_{aug}} \leq O(\|R^*\|^{k-K_1}_{\widehat{G}}) < O((\rho(R^*) + \epsilon)^{k-K_1}) \longrightarrow 0$$

as $k \longrightarrow \infty$. Therefore, $\|\widehat{\mathbf{w}}^{[k]}_{\text{aug}} - \widehat{\mathbf{w}}^*_{aug}\|_{\widehat{G}_{aug}}$ converges at a linear rate bounded by $\rho(R^*) + \epsilon < 1$ for any $\epsilon > 0$. $\square$

Theorem 5.3 indicates that when the iterate of ISTA is close enough to the optimal solution, then it converges linearly. We remark here that [4, 13] also present the same asymptotic local linear convergence result for ISTA. [4] requires the condition called strict sparsity pattern, which is identical to our strict complementarity condition while [13] requires partial smoothness property in a more general setting.

Different from [4, 13], our analysis only focuses on finite Euclidean space so that we can take advantage of well established matrix properties. Apart from the local linear convergence results, our analysis can characterizes the convergence behavior in terms of regimes. We show how ISTA stagnates before convergence with examples in Section 7. Moreover, the spectral analysis we established can be applied to the FISTA's local linear convergence, which was not studied in [4, 13]. In Section 6, we shall make comparison between ISTA and FISTA from the perspective of our spectral analysis.

**5.3. Local Linear Convergence of FISTA.** Once the optimal flag matrix is identified, for all $k > K_2$, $\widetilde{R}^{[k]}$, $\widetilde{\mathbf{h}}^{[k]}$ and $\widetilde{\mathbf{h}}^{[k]}_{\text{aug}}$ remain fixed. We denote them as $\widetilde{R}^*$, $\widetilde{\mathbf{h}}^*$ and $\widetilde{\mathbf{h}}^*_{\text{aug}}$ in this section. Though $\mathbf{N}^{[k]}_{\text{aug}}$ still changes at each step $k$, one can decompose it as follows.

LEMMA 5.4. *Assume that the LASSO problem* (1.1) *has a unique solution and this solution satisfies the strict complementarity. For all $k > K_2$, with $K_2$ defined in Lemma 5.2, we denote matrix $\mathbf{N}^{[\infty]}_{\text{aug}}$ as $\mathbf{N}^{[k]}_{\text{aug}}$ in which $\tau = 1$. By (3.11) one can write*

$\mathbf{N}_{aug}^{[k]} = \mathbf{N}_{aug}^{[\infty]} + (1-\tau^{[k]})\boldsymbol{\Delta}\mathbf{N}_{aug}$, *where*

$$\mathbf{N}_{aug}^{[\infty]} = \begin{pmatrix} N^{[\infty]} & \widetilde{\mathbf{h}}^*_{aug} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2\widetilde{R}^* & -\widetilde{R}^* & \widetilde{\mathbf{h}}^* \\ I & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
$$\text{and} \quad \boldsymbol{\Delta}\mathbf{N}_{aug} = \begin{pmatrix} \Delta N & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -\widetilde{R}^* & \widetilde{R}^* & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{5.4}$$

*Consequently,* $N^{[k]} = N^{[\infty]} + (1-\tau^{[k]})\Delta N$. *Also,* $\mathbf{N}_{\rm aug}^{[\infty]}$ *must also have a simple eigenvalue equal to* 1 *and the spectral radius of* $N^{[\infty]}$ *is strictly less than* 1. *In addition, the left and right eigenvectors of* $\mathbf{N}_{\rm aug}^{[k]}$ *corresponding to eigenvalue 1 are the same as that of* $\mathbf{N}_{\rm aug}^{[\infty]}$.

*Proof.* By Lemma 3.5(b), eigenvalue of $\mathbf{N}_{\rm aug}^{[\infty]}$ must lie in the disk $\mathcal{D}(1/2, 1/2)$. Having shown that $\mathbf{N}_{\rm aug}^{[k]}$ has only a simple eigenvalue equal to 1, by Lemma 3.5(c), $\widetilde{R}^*$ should have no eigenvalue equal to 1. This indicates matrix $\begin{pmatrix} 2\widetilde{R}^* & -\widetilde{R}^* \\ I & 0 \end{pmatrix}$ has no eigenvalue equal to 1. And hence $\mathbf{N}_{\rm aug}^{[\infty]}$ must also have a simple eigenvalue equal to 1. Consequently, the absolute value of any eigenvalue of $N^{[\infty]}$ is less than 1. A simple calculation shows that the left and right eigenvectors of $\mathbf{N}_{aug}^{[k]}$ are exactly $(0,\dots,0,1)$ and $\widetilde{\mathbf{w}}^*_{aug}$ ((5.1)), the same as that of $\mathbf{N}_{\rm aug}^{[\infty]}$. $\square$

Now we present one of the main results of our paper, the local linear convergence of FISTA, as below.

THEOREM 5.5. *Suppose the LASSO problem* (1.1) *has a unique solution and this solution has strict complementarity. Then eventually the FISTA iteration reaches a stage where it converges linearly to that unique solution.*

*Proof.* As shown in the proof of Lemma 5.2, there exists a pass number $K_2$ such that $\widetilde{\mathbf{w}}_{aug}^{[k]}$ lies in $\mathcal{B}_2$ for all $k > K_2$. Since the optimal flag matrix is identified, by Lemma 3.4, the unique solution, if it exists, must correspond to a unique eigenvector of eigenvalue 1 for the matrix $\mathbf{N}_{\rm aug}^{[k]}$. Starting from $K_2$-th pass,

$$\begin{aligned} \widetilde{\mathbf{w}}_{aug}^{[k+1]} &= \mathbf{N}_{\rm aug}^{[k]}\mathbf{N}_{\rm aug}^{[k-1]}\cdots\mathbf{N}_{\rm aug}^{[K_2]}\widetilde{\mathbf{w}}_{aug}^{[K_2]} \\ &= \begin{pmatrix} N^{[k]} & \widetilde{\mathbf{h}}^*_{aug} \\ 0 & 1 \end{pmatrix}\begin{pmatrix} N^{[k-1]} & \widetilde{\mathbf{h}}^*_{aug} \\ 0 & 1 \end{pmatrix}\cdots\begin{pmatrix} N^{[K_2]} & \widetilde{\mathbf{h}}^*_{aug} \\ 0 & 1 \end{pmatrix}\widetilde{\mathbf{w}}_{aug}^{[K_2]} \\ &= \begin{pmatrix} N^{[k]}N^{[k-1]}\cdots N^{[K_2]} & \boxed{*} \\ 0 & 1 \end{pmatrix}\widetilde{\mathbf{w}}_{aug}^{[K_2]}, \end{aligned} \tag{5.5}$$

where $\boxed{*}$ denotes entries that do not need to be specified in detail. For each $N^{[k]}$, we can write $N^{[k]} = N^{[\infty]} + (1-\tau^{[k]})\Delta N$. Due to the fixed flag matrix after $K_2$-th pass, $N^{[\infty]}$ and $\Delta N$ remain fixed. By Theorem 3.2, $\forall\epsilon > 0$ with $\epsilon < 1 - \rho(N^{[\infty]})$, $\exists\ \|\cdot\|_{\widetilde{G}}$ such that $\|N^{[\infty]}\|_{\widetilde{G}} < \rho(N^{[\infty]}) + 1/2\epsilon < 1 - 1/2\epsilon$. Since $\tau^{[k]} \to 1$, there must exist a pass $K_3(> K_2)$ such that $(1-\tau^{[k]})\cdot\|\Delta N\|_{\widetilde{G}} < 1/2\epsilon$ for all $k > K_3$. Therefore, starting at $K_3$-th pass, one has $\|N^{[k]}\|_{\widetilde{G}} \le \|N^{\infty}\|_{\widetilde{G}} + (1-\tau^{[k]})\|\Delta N\|_{\widetilde{G}} < \rho(N^{[\infty]}) + \epsilon < 1$. As in the proof of Thm 5.3, let the error vector at the $k$-th pass for FISTA be

$$\widetilde{\mathbf{e}}_{aug}^{[k]} = \widetilde{\mathbf{w}}_{aug}^{[k]} - \widetilde{\mathbf{w}}^*_{aug} = \begin{pmatrix} \widetilde{\mathbf{e}}^{[k]} \\ \widetilde{\mathbf{e}}^{[k-1]} \\ 0 \end{pmatrix} \text{ where } \widetilde{\mathbf{e}}^{[k]} = \begin{pmatrix} \widetilde{\mathbf{w}}^{[k]} \\ \widetilde{\mathbf{w}}^{[k-1]} \\ 1 \end{pmatrix} - \begin{pmatrix} \widetilde{\mathbf{w}}^* \\ \widetilde{\mathbf{w}}^* \\ 1 \end{pmatrix}.$$

Then

$$\widetilde{\mathbf{w}}_{aug}^{[k+1]} = \widetilde{\mathbf{w}}_{aug}^{*} + \widetilde{\mathbf{e}}_{aug}^{[k+1]} = \mathbf{N}_{aug}^{[k]}(\widetilde{\mathbf{w}}_{aug}^{*} + \widetilde{\mathbf{e}}_{aug}^{[k]}) = \widetilde{\mathbf{w}}_{aug}^{*} + \begin{pmatrix} N^{[k]} & 0 \\ 0 & 1 \end{pmatrix} \widetilde{\mathbf{e}}_{aug}^{[k]} \tag{5.6}$$

with $\widetilde{\mathbf{e}}^{[k+1]} = N^{[k]}\widetilde{\mathbf{e}}^{[k]}$. Let $\widetilde{G}_{aug} = \begin{pmatrix} \widetilde{G} & 0 \\ 0 & 1 \end{pmatrix}$, then

$$\|\widetilde{\mathbf{e}}_{\text{aug}}^{[k+1]}\|_{\widetilde{G}_{aug}} = \|\widetilde{\mathbf{e}}^{[k+1]}\|_{\widetilde{G}} \le \|N^{[k]}\|_{\widetilde{G}}\|\widetilde{\mathbf{e}}^{[k]}\|_{\widetilde{G}} = \|N^{[k]}\|_{\widetilde{G}}\|\widetilde{\mathbf{e}}_{\text{aug}}^{[k]}\|_{\widetilde{G}_{aug}}$$

Hence starting from $K_3$-th pass,

$$\|\widetilde{\mathbf{e}}_{\text{aug}}^{[k]}\|_{\widetilde{G}_{aug}} \le O(\|N^{[k-1]}\|_{\widetilde{G}}\|N^{[k-2]}\|_{\widetilde{G}} \cdots \|N^{[K_3]}\|_{\widetilde{G}}) \le O((\rho(N^{[\infty]}) + \epsilon)^{k-K_3}) \longrightarrow 0 \tag{5.7}$$

as $k \longrightarrow \infty$. Therefore, $\|\widetilde{\mathbf{w}}_{\text{aug}}^{[k]} - \widetilde{\mathbf{w}}_{\text{aug}}^{*}\|_{\widetilde{G}_{aug}}$ converges at a linear rate bounded by $\rho(N^{[\infty]}) + \epsilon < 1$ for any $\epsilon > 0$. From the proof, one can also observe that the linear rate at step $k$ depends on $N^{[k]}$. ☐

**5.4. Other Choices of $L$.** The previous analysis is based on $L \ge \|A^T A\|_2$. However, this assumption is not necessary to get our main results for ISTA. [4] proves that the ISTA iterates can converge to the optimal point as long as $L > 1/2\|A^T A\|_2$. It can be verified that our analysis allows the same $L$ choice for ISTA.

THEOREM 5.6. *Lemmas 3.1, 3.3, 4.2 and Theorem 5.3 hold for $L > 1/2\|A^T A\|_2$.*

However, from our analysis, we can show that there is no guarantee that FISTA would converge if $L < \|A^T A\|_2$. In Lemma 3.5(b), (3.13) cannot hold if $L < \|A^T A\|_2$, indicating that some of the eigenvalues may be outside the disk $\mathcal{D}(1/2, 1/2)$. Hence convergence is not guaranteed.

## 6. Acceleration.

**6.1. Comparison between ISTA and FISTA.** It is known that FISTA exhibits a global convergence rate of $O(1/k^2)$, which accelerates ISTA's $O(1/k)$ convergence rate. Compared to this worst case convergence result, we analyze how FISTA and ISTA behave through all iterations on the perspective of spectral analysis we establish in this paper. First, we characterize one important property based on three possible regimes.

LEMMA 6.1. *Suppose $R$ and $N$ have the same the flag matrix, ISTA and FISTA have the following relations:*

(a). *If FISTA is in regime [A] or [C], then so is ISTA, and vice versa.*

(b). *If FISTA is in regime [B], then so is ISTA, and vice versa.*

*Proof.* We note that if FISTA and ISTA start at the same iterate, we have $\widehat{D} = \widetilde{D}$, hence $\widetilde{R}$ defined in (3.9) is exactly operator $R$ defined in (3.5).

(a). If FISTA is in regime [A] or [C], then $N$ either has no eigenvalue equal to 1 or has a complete set of eigenvectors associated with eigenvalue 1. In other words, the augmented matrix $\mathbf{N}_{\text{aug}}$ must have a complete set of eigenvectors for eigenvalue

1. Let $\begin{pmatrix}\mathbf{w}_1\\ \mathbf{w}_2\\ 1\end{pmatrix}$ be the eigenvector for eigenvalue 1, then

$$
\begin{aligned}
(N-I)\begin{pmatrix}\mathbf{w}_1\\ \mathbf{w}_2\\ 1\end{pmatrix} &= \begin{pmatrix}(1+\tau)R & -\tau R & \mathbf{h}\\ I & -I & 0\\ 0 & 0 & 0\end{pmatrix}\begin{pmatrix}\mathbf{w}_1\\ \mathbf{w}_2\\ 1\end{pmatrix} = 0\\
\iff\quad & \mathbf{w}_1 = \mathbf{w}_2 \quad \text{(by second row)}\\
\iff\quad & R\mathbf{w}_1 - \mathbf{w}_1 + \mathbf{h} = (R-I)\mathbf{w}_1 + \mathbf{h} = 0\\
\iff\quad & \begin{pmatrix}R & \mathbf{h}\\ 0 & 1\end{pmatrix}\begin{pmatrix}\mathbf{w}_1\\ 1\end{pmatrix} = \begin{pmatrix}\mathbf{w}_1\\ 1\end{pmatrix}.
\end{aligned}
\tag{6.1}
$$

Therefore, $\begin{pmatrix}\mathbf{w}_1\\ 1\end{pmatrix}$ becomes the eigenvector for eigenvalue 1 of $\mathbf{R}_{aug}$. $R$ must either have no eigenvalue equal to 1 (in regime [A]) or have a complete set of eigenvectors associated with eigenvalue 1 (in regime [C]). The opposite direction follows by similar argument.

(b). Since one of the regimes [A], [B], [C] must occur, this statement can be considered as the contraposition of (a). □

This lemma suggests that both ISTA and FISTA are in the same regime as long as both operators have the same flag matrix. It motivates one to compare in each regime between FISTA and ISTA when starting from the same starting point (which results in the same flag matrix). By assuming the same starting point and a fixed flag matrix, we have $\widehat{D}^{[k]} = \widetilde{D}^{[k]} = \widehat{D}^{[k+1]} = \widetilde{D}^{[k+1]}$ and thus $\widetilde{R} = R$, $\mathbf{h} = \widetilde{\mathbf{h}}$. We will use these notations interchangeably and omit $^{[k]}$ for anything but iterates in the following analysis. It turns out that FISTA is faster in regime [B], but not always faster in regimes [A] and [C] depending on the parameter $\tau^{[k]}$.

**6.1.1. In Regime [B].** In regime [B], as mentioned in Section 4.2, there exist Jordan chains such that the difference between consecutive iterates will converge to a constant step. Let $\begin{pmatrix}\widehat{\mathbf{w}}^{[k+1]}\\ 1\end{pmatrix}$, $\begin{pmatrix}\widehat{\mathbf{w}}^{[k]}\\ 1\end{pmatrix}$ and $\begin{pmatrix}\widetilde{\mathbf{w}}^{[k+1]}\\ \widetilde{\mathbf{w}}^{[k]}\\ 1\end{pmatrix}$, $\begin{pmatrix}\widetilde{\mathbf{w}}^{[k]}\\ \widetilde{\mathbf{w}}^{[k-1]}\\ 1\end{pmatrix}$ be two consecutive iterates for ISTA and FISTA, respectively. In the following lemmas, we will show that the constant step for FISTA is larger than ISTA when starting at the same point, which yields a speedup.

Lemma 6.2. *The constant step vector for ISTA is* $\begin{pmatrix}\mathbf{v}\\ 0\end{pmatrix}$, *where* $\mathbf{v} = R\mathbf{v}$ *is an eigenvector of* $R$.

*Proof.* For ISTA, there must be a Jordan block $\mathbf{J}_R^1$ for the augmented matrix $\mathbf{R}_{aug}$. Then there exists a Jordan chain such that

$$
\begin{pmatrix}R & \mathbf{h}\\ 0 & 1\end{pmatrix}\begin{pmatrix}\widehat{\mathbf{w}}\\ 1\end{pmatrix} = \begin{pmatrix}\widehat{\mathbf{w}}+\mathbf{v}\\ 1\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}R & \mathbf{h}\\ 0 & 1\end{pmatrix}\begin{pmatrix}\mathbf{v}\\ 0\end{pmatrix} = \begin{pmatrix}\mathbf{v}\\ 0\end{pmatrix}.
\tag{6.2}
$$

In other words, each pass of ISTA will add a constant vector $\begin{pmatrix}\mathbf{v}\\ 0\end{pmatrix}$ in regime [B]. □

LEMMA 6.3. *The constant step vector for FISTA has the form* $\begin{pmatrix} c\mathbf{v} \\ c\mathbf{v} \\ 0 \end{pmatrix}$, *where* $\mathbf{v}$ *is the same* $\mathbf{v}$ *in Lemma 6.2, c is a scalar to be determined.*

*Proof.* Assume the constant vector is $\begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ 0 \end{pmatrix}$. Then the basic iteration of FISTA is

$$\begin{pmatrix} \widetilde{\mathbf{w}}^{[k+1]} \\ \widetilde{\mathbf{w}}^{[k]} \\ 1 \end{pmatrix} = N \begin{pmatrix} \widetilde{\mathbf{w}}^{[k]} \\ \widetilde{\mathbf{w}}^{[k-1]} \\ 1 \end{pmatrix} = \begin{pmatrix} (1+\tau)R & -\tau R & \mathbf{h} \\ I & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{w}}^{[k]} \\ \widetilde{\mathbf{w}}^{[k-1]} \\ 1 \end{pmatrix} = \begin{pmatrix} \widetilde{\mathbf{w}}^{[k]} \\ \widetilde{\mathbf{w}}^{[k-1]} \\ 1 \end{pmatrix} + \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ 0 \end{pmatrix}.$$

Due to the presence of Jordan block $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, there exists a Jordan chain

$$\text{(a)}(N-I) \begin{pmatrix} \widetilde{\mathbf{w}}^{[k]} \\ \widetilde{\mathbf{w}}^{[k-1]} \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ 0 \end{pmatrix} \quad \text{and} \quad \text{(b)}N \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ 0 \end{pmatrix}. \tag{6.3}$$

In (6.3a), the second row implies $\mathbf{v}_1 = \mathbf{v}_2$. Then, the first row implies $R\mathbf{v}_1 = \mathbf{v}_1$. Since both $\mathbf{v}_1$ and $\mathbf{v}$ are eigenvectors for eigenvalue 1 of $R$, we can write $\mathbf{v}_1 = c\mathbf{v}$ where $c$ is a scalar to be determined. Hence the constant step should be $\begin{pmatrix} c\mathbf{v} \\ c\mathbf{v} \\ 0 \end{pmatrix}$. ☐

LEMMA 6.4. *Suppose ISTA and FISTA start from the same point in the same regime [B], i.e.* $\widehat{\mathbf{w}}^{[k]} = \widetilde{\mathbf{w}}^{[k]}$, *then c in Lemma 6.3 equals* $\frac{1}{1-\tau}$, *where* $\tau$ *is a scalar close to* 1. *The constant step vector for FISTA is* $\frac{1}{1-\tau}\begin{pmatrix} \mathbf{v} \\ \mathbf{v} \\ 0 \end{pmatrix}$, *which is larger than* $\begin{pmatrix} \mathbf{v} \\ \mathbf{v} \\ 0 \end{pmatrix}$, *the ISTA constant step.*

*Proof.* By Lemma 6.3, the equation (6.3) expands to

$$\begin{aligned} (N-I) \begin{pmatrix} \widetilde{\mathbf{w}}^{[k]} \\ \widetilde{\mathbf{w}}^{[k-1]} \\ 1 \end{pmatrix} &= \left[ \begin{pmatrix} (1+\tau)R & -\tau R & \mathbf{h} \\ I & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} - I \right] \begin{pmatrix} \widetilde{\mathbf{w}}^{[k]} \\ \widetilde{\mathbf{w}}^{[k-1]} \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} ((1+\tau)R - I)\widetilde{\mathbf{w}}^{[k]} - \tau R \widetilde{\mathbf{w}}^{[k-1]} + \mathbf{h} \\ \widetilde{\mathbf{w}}^{[k]} - \widetilde{\mathbf{w}}^{[k-1]} \\ 0 \end{pmatrix} \end{aligned} \tag{6.4}$$

which is supposed to be equal to $\begin{pmatrix} c\mathbf{v} \\ c\mathbf{v} \\ 0 \end{pmatrix}$. From the second row, $\widetilde{\mathbf{w}}^{[k]} - \widetilde{\mathbf{w}}^{[k-1]} = c\mathbf{v}$ or $\widetilde{\mathbf{w}}^{[k-1]} = \widetilde{\mathbf{w}}^{[k]} - c\mathbf{v}$. Hence, the first row should be $c\mathbf{v} = ((1+\tau)R - I)\widetilde{\mathbf{w}}^{[k]} - \tau R \widetilde{\mathbf{w}}^{[k-1]} + \mathbf{h} = ((1+\tau)R - I)\widetilde{\mathbf{w}}^{[k]} - \tau R(\widetilde{\mathbf{w}}^{[k]} - c\mathbf{v}) + \mathbf{h} = (R-I)\widetilde{\mathbf{w}}^{[k]} + \mathbf{h} + c\tau\mathbf{v}$. The last equality follows from $R\mathbf{v} = \mathbf{v}$.

If FISTA and ISTA start from the same point $\widehat{\mathbf{w}}^{[k]} = \widetilde{\mathbf{w}}^{[k]}$, then $c\mathbf{v} = (R - I)\widetilde{\mathbf{w}}^{[k]} + \mathbf{h} + c\tau\mathbf{v} = (R-I)\widehat{\mathbf{w}}^{[k]} + \mathbf{h} + c\tau\mathbf{v} = \mathbf{v} + c\tau\mathbf{v}$, leading to $c(1-\tau) = 1$. Hence $c = \frac{1}{1-\tau}$. ☐

Lemma 6.4 indicates that if FISTA and ISTA start from the same starting point in one specific regime [B], then it will cost FISTA fewer iterations to leave this regime with larger constant step. Hence FISTA represents an acceleration compared to ISTA in regime [B].

**6.1.2. In Regimes [A] and [C].** On the other hand, in regimes [A] and [C], the convergence rate of the two algorithms are related to the spectral radius of $R$ and $N$. Particularly, the rate of FISTA depends on $\tau$ and the iteration number, since $\tau$ is a determined sequence based on iteration numbers. Let $\beta$, $\gamma$ denote an eigenvalue of $R$, $N$, respectively, and $\beta_{\max}$, $\gamma_{\max}$ denote the corresponding eigenvalues of largest absolute value. As stated in Section 4.2, we must have $1 > \beta_{\max}, \gamma_{\max} \geq 0$ in regimes [A] or [C]. In addition, by Lemma 3.5, $\beta$ and $\gamma$ satisfy the relation $\gamma^2 - \gamma(1+\tau)\beta + \tau\beta = 0$. Let $\gamma_1$ and $\gamma_2$ be two roots of $\gamma$. We conclude our result in the following proposition.

Proposition 6.5. *Suppose ISTA and FISTA start from the same point in a certain regime [A] or [C] and $D^{[k]} = D^{[k+1]}$, FISTA is faster than ISTA if $0 < \tau < \beta_{\max} < 1$ but slower if $0 < \beta_{\max} < \tau < 1$. Noting that $\beta_{\max}$ is a fixed value for one specific regime, if it is well separated from $1$, with the $\tau$ growing to $1$ such that $\beta_{\max} < \tau$, ISTA will be faster than FISTA toward the end.*

*Proof.* The roots $\gamma$ of (3.12) are real if $\frac{4\tau}{(1+\tau)^2} < \beta$ and are complex if $\beta < \frac{4\tau}{(1+\tau)^2}$. Noting that $\tau \leq \frac{4\tau}{(1+\tau)^2} \leq 1$ with equality only if $\tau = 1$, we consider two cases.

(i). If $\frac{4\tau}{(1+\tau)^2} < \beta$, then $\tau < \beta$. Without loss of generality, $\gamma_1 = \max\{\gamma_1, \gamma_2\} = \frac{(1+\tau)\beta}{2} + \sqrt{\frac{(1+\tau)^2\beta^2 - 4\tau\beta}{4}} < \frac{(1+\tau)\beta}{2} + \sqrt{\frac{(1+\tau)^2\beta^2 - 4\beta^2}{4}} < \beta$. The first inequality is due to $\beta > \tau$ and the second one is due to $\tau < 1$.

(ii). If $\beta < \frac{4\tau}{(1+\tau)^2}$, then $\gamma_1$ and $\gamma_2$ are a conjugate complex pair such that $|\gamma_1|^2 = \gamma_1\gamma_2 = \tau\beta$. If $\tau < \beta < \frac{4\tau}{(1+\tau)^2}$, then $|\gamma_1| = \sqrt{\tau\beta} < \beta$. If $\beta < \tau < \frac{4\tau}{(1+\tau)^2}$, then $|\gamma_1| = \sqrt{\tau\beta} > \beta$.

If FISTA were to continue long enough, eventually $\tau^{[k]}$ will become larger than $\beta_{\max} < 1$ at some step $K_4$, at which point the asymptotic convergence rate for FISTA will be slower than that for ISTA. $\square$

Proposition 6.5 concludes that if the starting points are the same in regimes [A] or [C], then ISTA will first be slower but then be faster as the iteration progresses.

**6.2. A Heuristic Algorithm.** The above analysis would indicate that one should try to take advantage of the generally faster $O(1/k^2)$ rate of convergence of FISTA, but switch to ISTA when it becomes faster during the final regime. We test this idea using a Hybrid F/ISTA method in which we run FISTA until reaching the final linear regime and then switch to ISTA. The result is illustrated in the examples. However, there is no way in practice to know when one reaches the final regime without knowing the optimal solution. So we also propose a simple heuristic algorithm in which both ISTA and FISTA iterates are computed at every iteration, choosing whichever iterate shows greater progress measured in terms of the decrease in the objective function (1.1). This heuristic algorithm is given as Alg. 5, with initialization $\mathbf{y}^{[1]} = \widetilde{\mathbf{x}}^{[0]} \in \mathbb{R}^n$ and $t^{[0]} = t^{[1]} = 1$.

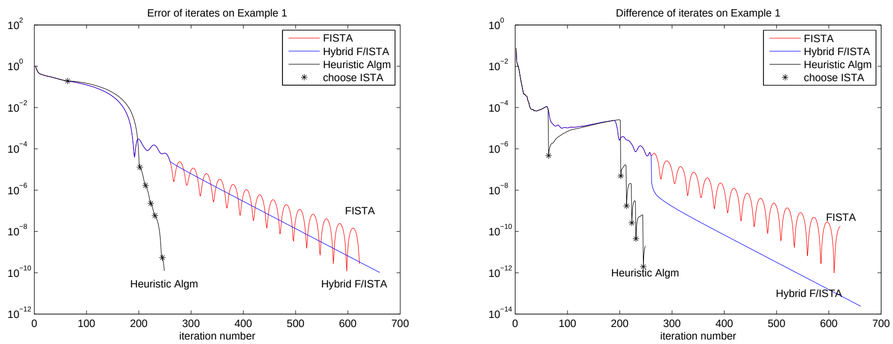


FIG. 1. *Left: Error of iterates of FISTA, Hybrid F/ISTA and Heuristic Algorithm. Right: Difference of iterates of FISTA, Hybrid F/ISTA and Heuristic algorithms. The star * on the Heuristic Alg. curve marks the iterations where the ISTA iterate was selected.*

---

Algorithm 5: One pass of the heuristic algorithm

---

start with $t^{[k]}$, $t^{[k-1]}$, $\mathbf{x}^{[k-1]}$ and $\mathbf{x}^{[k-2]}$.

1. Set $\mathbf{y}^{[k]} = \mathbf{x}^{[k-1]} + \frac{t^{[k-1]}-1}{t^{[k]}}(\mathbf{x}^{[k-1]} - \mathbf{x}^{[k-2]})$.
2. Set $\widetilde{\mathbf{x}}^{[k]} = \mathrm{Shr}_{\lambda/L}((I - {}^1\!/_L A^{\mathrm{T}}A)\mathbf{y}^{[k]} + {}^1\!/_L A^{\mathrm{T}}\mathbf{b})$.
   and $\widehat{\mathbf{x}}^{[k]} = \mathrm{Shr}_{\lambda/L}((I - {}^1\!/_L A^{\mathrm{T}}A)\mathbf{x}^{[k-1]} + {}^1\!/_L A^{\mathrm{T}}\mathbf{b})$.
3. If $F(\widehat{\mathbf{x}}^{[k]}) > F(\widetilde{\mathbf{x}}^{[k]})$, then $\mathbf{x}^{[k]} = \widetilde{\mathbf{x}}^{[k]}$. Else, $\mathbf{x}^{[k]} = \widehat{\mathbf{x}}^{[k]}$
4. Set $t^{[k+1]} = \frac{1+\sqrt{1+4t^{[k]2}}}{2}$.

Result is $t^{[k+1]}$, $t^{[k]}$, $\widetilde{\mathbf{x}}^{[k]}$, $\widetilde{\mathbf{x}}^{[k-1]}$ for next pass.

---

Though this updating rule is simple and lacks the theoretical $O(1/k^2)$ global convergence rate, the experimental results show that it can converge very fast, sometimes faster than we expect. The computational cost of one iteration of the heuristic algorithm is the same as for FISTA plus one extra shrinkage operation and two evaluations of the objective function. However, in our following examples, we observe that the proposed algorithm can sometimes be so much faster than ISTA or FISTA alone that the overall cost can be much less. An alternative acceleration heuristic (with similar behavior) was discussed in [15].

**7. Examples. Example** 1. We illustrate the eigen-analysis of the behavior of ISTA and FISTA on a uniform randomly generated LASSO problem. Specifically, in problem (1.1), $A$ and $\mathbf{b}$ are generated independently by a uniform distribution over $[-1, 1]$, $A$ being $20 \times 40$, $\lambda = 0.1$. Since $A$ is drawn by a continuous distribution, as noted in Lemma 4 of [23], problem (1.1) must have a unique solution. Figures 1 & 2 show the ISTA and FISTA convergence behavior. The figures show the error of $\mathbf{x}$, $\|\mathbf{x}^{[k]} - \mathbf{x}^*\|$ (A: top curve) and the difference between two consecutive iterates of $\mathbf{x}$: $\|\mathbf{x}^{[k]} - \mathbf{x}^{[k-1]}\|$ (B: bottom curve).

Figure 2 (left) shows the behavior of ISTA. ISTA takes 5324 iterations to converge and the flag matrix $D$ changes 25 times in total. During the first 174 iterations, the iteration passes through a few transitional phases and the flag matrix $D$ changes 20 times. After that, from iteration 175 to 483, it stays in regime [B] with an invariant $D$. Then from iteration 484 to 530, from 531 to 756, from 767 to 4722 and from 4723 to 4972, it passes through four different regimes [B]. Within each regime [B], the flag matrix $D$ is invariant. According to our analysis in Section 4.2, there exists a Jordan

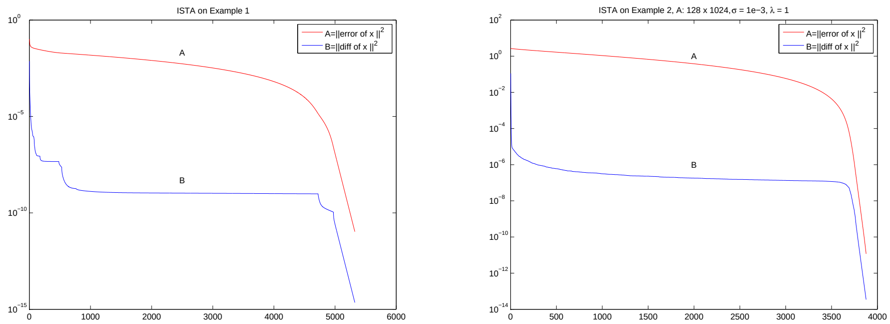


FIG. 2. *ISTA on Example 1 (left) and Example 2 (right): Curves* $\mathbf{A}$*:* $\|\mathbf{x}^{[k]} - \mathbf{x}^*\|^2$*.* $\mathbf{B}$*:* $\|\mathbf{x}^{[k]} - \mathbf{x}^{[k-1]}\|^2$*.*

chain in each of these regimes [B], indicating that we are indeed in a "constant step" regime. In other words, the difference between two consecutive iterates $\|\mathbf{x}^{[k]} - \mathbf{x}^{[k-1]}\|$ quickly converges to $\mathbf{R}_{aug}$'s eigenvector for eigenvalue 1 in each of these regimes [B]. Taking iterations from 767 to 4972 for example, one could notice curve $\mathbf{B}$ in Figure 2 (left) that $\|\mathbf{x}^{[k]} - \mathbf{x}^{[k-1]}\|$ is a constant from iteration 767 to 4722. Finally, at iteration 4973, it reaches and stays in the final regime [A], converging linearly in 351 steps. Indeed, the iterates are close enough to the final optimum so that the flags never change again. The linear convergence rate depends on the spectral radius of $R$, i.e. upper left part of $\mathbf{R}_{aug}$, which is $\rho(R) = 0.9817$, separated from the $\mathbf{R}_{aug}$'s largest eigenvalue 1, consistent with Theorem 5.3.

Figure 1 (FISTA) shows the behavior of FISTA. FISTA takes 622 iterations to converge and the flag matrix $D$ changes 42 times in total. After flag matrix $D$ changes 42 times in initial 258 iterations, it reaches the final regime [A] at iteration 259 and converges linearly in 363 steps. Since $\mathbf{N}_{aug}$ varies at each iteration due to varying $\tau$, the convergence rate changes very slightly step by step. The spectral radius of $N$, i.e. upper left part of the operator $\mathbf{N}_{aug}$ in the last step is $\rho(N) = 0.9914$. Actually, the largest eigenvalues of $N$ are a complex conjugate pair, $0.9843 \pm 0.1185i$. They are complex numbers because of the increasing $\tau$, as stated in Proposition 6.5 in Section 6. Hence based on the power method, in the final regime, the convergence to eigenvector for eigenvalue 1 of $\mathbf{N}_{aug}$ will oscillate between the conjugate complex pair. This explains why curves in Figure 1 (FISTA) oscillate in the latter part of the FISTA convergence.

We made three more remarks with our analysis in Section 6 on this example.

1. It costs FISTA many fewer steps (259 iterations) than ISTA (4973 iterations) to get to the final regime. The main reason is that FISTA has much larger constant steps in regime [B] so that it can jump out of that regime more quickly. Though this will lead to more changes of regimes (flag matrix $D$ changes 42 times, 17 more times than ISTA), the overall iteration numbers have been cut down, consistent with Lemma 6.4. One can also notice this in Figure 1 (FISTA) that difference between iterates do not remain constant for many iterations, with the process transitioning into the final regime.

2. Figure 1 (Hybrid F/ISTA) shows the behavior of hybrid F/ISTA idea illustrated in Section 6.2. Particularly for this example, it runs FISTA until it reaches final regime and then switches to ISTA at iteration 260. At step 260, $\tau = 0.9886$, larger

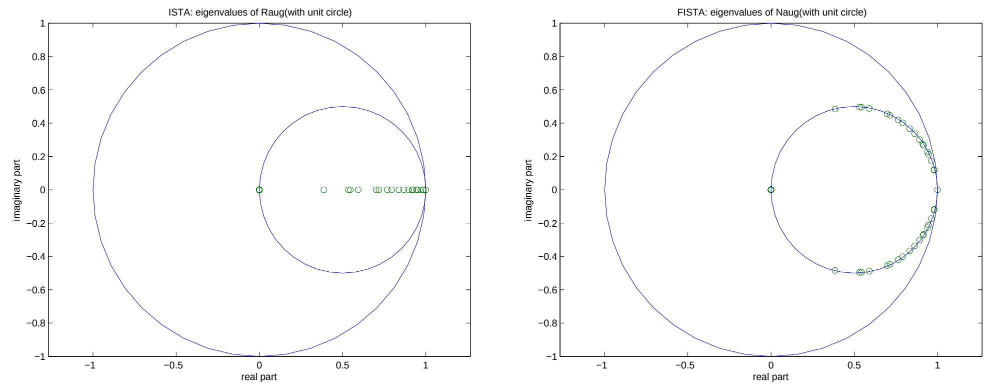


FIG. 3. *ISTA (left) and FISTA (right) on Example 1: Eigenvalues of ISTA operator* $\mathbf{R}_{\rm aug}$ *and FISTA operator* $\mathbf{N}_{\rm aug}$ *on the complex plane during the last regime of the iteration process. The unit circle and* $\mathcal{D}(1/2, 1/2)$ *are shown for reference.*

than ISTA rate 0.9817, predicting that ISTA should converge faster than FISTA. Though Hybrid F/ISTA converges in 661 iterations, more than 622 of FISTA iterations, it doesn't contradict with our analysis. In Figure 1, from the gradient of FISTA and Hybrid F/ISTA curve, one could observe that Hybrid F/ISTA converges faster than the upper bound of FISTA.

3. Figure 1 (Heuristic Algm) shows the behavior of heuristic algorithm established in Section 6.2. It converges in only 212 iterations with the same accuracy. Though it costs extra shrinkage operations and objective value evaluations, the overall cost can still be much less.

Figure 3 shows the eigenvalues of the operators $\mathbf{R}_{aug}$ and $\mathbf{N}_{aug}$ during the final regime. One notices that the eigenvalues for the $\mathbf{R}_{aug}$ from (3.5) lie strictly on the interval $(0, 1)$ and eigenvalues for $\mathbf{N}_{aug}$ lie close to the boundary but strictly inside the circle $\mathcal{D}(1/2, 1/2)$ (except for 0 and 1), consistent with Lemmas 3.3 & 3.5.

**Example 2** We consider an example of compressive sensing. The purpose of this example is to show and compare the convergence behavior of different algorithms mentioned in previous sections to support our analysis. Suppose there exists a true sparse signal represented by a $n$-th dimension vector $\mathbf{x}$ with $k$ non-zero elements. We observe the image of $\mathbf{x}_s$ under the linear transformation $A\mathbf{x}_s$, where $A$ is the so-called measurement matrix. Our observation thus should be

$$\mathbf{b} = A\mathbf{x}_s + \epsilon \tag{7.1}$$

where $\epsilon$ is the observation noise. The goal is to recover the sparse vector $\mathbf{x}_s$ from the measurement matrix $A$ and observation $\mathbf{b}$. For this example, we let $A \in \mathbb{R}^{m\times n}$ be Gaussian matrix whose elements are *i.i.d.* distributed as $\mathcal{N}(0, 1)$ with $m = 128$ and $n = 1024$, $\epsilon$ be a vector whose elements are *i.i.d.* distributed as $\mathcal{N}(0, \sigma^2)$ with $\sigma = 10^{-3}$. The original true signal for the problem is generated by choosing the locations of $\mathbf{x}$'s $k(= 10)$ nonzeros uniformly at random, then setting those locations to values drawn from $\mathcal{N}(0, 2^2)$. We solve this compressive sensing problem by model (1.1) with $\lambda = 1$ and illustrate the convergence behavior of four methods: ISTA, FISTA, Hybrid F/ISTA and the heuristic algorithm. Figure 2 (right) shows ISTA's convergence behavior. Figures 4 respectively show the error and difference of the iterates of other three methods.

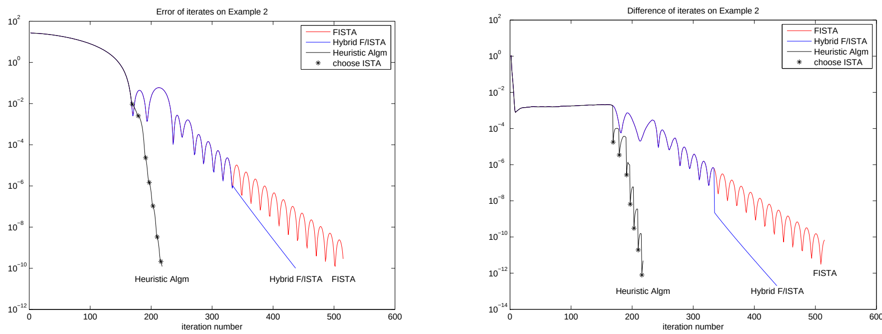


FIG. 4. *Left: Error of iterates of FISTA, Hybrid F/ISTA and Heuristic Algorithm. Right: Difference of iterates of FISTA, Hybrid F/ISTA and Heuristic algorithms. The star * on the Heuristic Alg. curve marks the iterations where the ISTA iterate was selected.*

**ISTA:** It costs 3763 iterations to reach the final regime, finally converging in altogether 3882 iterations. The linear convergence rate is the second largest eigenvalue of $\mathbf{R}_{aug}$, which equals to 0.9584, well separated from 1.

**FISTA:** It costs 333 iterations to reach the final regime and converges in totally 515 iterations. The linear convergence rate at pass $k$, as shown in Theorem 5.5, depends on the second largest eigenvalue of $\mathbf{N}_{aug}^{[k]}$. The linear rate is 0.9746 with $\tau = 0.9911$ at step 333 and the rate is 0.9761 with $\tau = 0.9942$ at step 515.

The iteration number for FISTA obviously is shorter than ISTA. It can be seen that in Figure 2 (right, curve B) that the difference of ISTA iterates remain at a constant number for many iterations. This is because they are in the constant regimes such that the difference between consecutive iterates are converging to a constant vector. From Figure 4 (Right), one can observe that the difference of FISTA iterates doesn't stagnate for as many iterations as ISTA because it has a larger constant step size, as predicted in Section 6.1.

**Hybrid F/ISTA:** By the time FISTA reaches the final regime, $\tau = 0.9911$ which is already greater than ISTA rate 0.9584, predicting that switching to ISTA at this point would be advantageous by Proposition 6.5. Particularly, one runs FISTA iterates until the arrival of the final regime at step 334. Then one switches to ISTA until convergence so that a faster linear rate is obtained. The algorithm of this idea converges only in 437 iterations with the same accuracy compared to 515 iterations of FISTA. One can observe the acceleration in Figure 4 (Left).

**Heuristic Algorithm:** Finally, we test our heuristic algorithm developed in Section 6.2 on this example. Basically, at each iteration, the algorithm compares the objective value by running ISTA and FISTA, and update the iterate with the lower value. From our analysis in Section 6, the heuristic algorithm should mostly run FISTA before final linear regime and switch to ISTA very often toward the end. We indeed observe this phenomenon in Figure 4. The heuristic algorithm converges only in 218 iterations. Though it loses the theoretical global $O(1/k^2)$ rate, it has a better practical performance. This, combined with the Hybrid F/ISTA idea, is consistent with our analysis of switching iterations to ISTA towards the end.

**8. Conclusion.** In this paper, we show the local linear convergence of ISTA and FISTA, applied to the model LASSO problem. Extending the same techniques as in [3], both algorithms can be modeled as a matrix recurrence and thus the associated spectra can be used to analyze their convergence behaviors. It is shown that the method normally passes through several regimes of four types and eventually settles on a "linear regime" in which the iterates converge linearly with the rate depending on the absolute value of the second largest eigenvalue of the matrix recurrence.

In addition, we provide a way to analyze every type of the regime. Such analysis in terms of regimes allows one to study the aspect of acceleration of FISTA. It is well known that FISTA is faster than ISTA according the worst case complexity bound. Our analysis gives another way to show how both methods behave during the whole iterations. It turns out that FISTA is not always faster than ISTA in regime [A] and [C], depending on the continually growing stepsize. But in general FISTA is faster because of its acceleration in regime [B]. A heuristic algorithm is developed based on this observation and exhibits very good numerical performance. Inspired by the theory developed in this paper, the behavior of the heuristic algorithm needs a more complete analysis beyond the scope of the present paper.

**Acknowledgment** We thank the two anonymous referees for their helpful and essential comments.

REFERENCES


[1] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sciences, 2 (2009), pp. 183–202.

[2] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS-SIAM Series on optimization, Society for Industrial and Applied Mathematics (SIAM), 2001.

[3] D. Boley, *Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs*, SIAM Journal on Optimization, 23 (2013), pp. 2183–2207.

[4] K. Bredies and D. Lorenz, *Linear convergence of iterative soft-thresholding*, Journal of Fourier Analysis and Applications, 14 (2008), pp. 813–837.

[5] E. J. Candès and Y. Plan, *Near-ideal model selection by $\ell_1$ minimization*, The Annals of Statistics, 37 (2009), pp. 2145–2177.

[6] S. Chen, D. L. Donoho, and M. Saunders, *Atomic decomposition by basis pursuit*, SIAM Journal on Scientific Computing, 20 (1998), pp. 33–61.

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

[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Annals of Statistics, 32 (2004), pp. 407–499.

[9] L. Fuchs, *Recovery of exact sparse representations in the presence of bounded noise*, IEEE Trans. on I.T, (2005), pp. 3601–3608.

[10] G. H. Gene and V. Loan, *Matrix Computations (4th Ed.)*, Johns Hopkins University Press, Baltimore, MD, USA, 2013.

[11] K. Gregor and Y. LeCun, *Learning fast approximations of sparse coding*, in Proceedings of the 27th Annual International Conference on Machine Learning, ICML '10, 2010, pp. 399–406.

[12] A. S. Householder, *The Theory of Matrices in Numerical Analysis*, A Blaisdell book in the pure and applied sciences, Blaisdell, New York, Toronto, London, 1964. Autre tirage : 1965.

[13] J. Liang, J. Fadili, and G. Peyré, *Local linear convergence of forward–backward under partial smoothness*, in Advances in Neural Information Processing Systems 27, 2014, pp. 1970–1978.

[14] Y. Nesterov, *A method of solving a convex programming problem with convergence rate $o(1/k^2)$*, Soviet Mathematics Doklady, 27 (1983), pp. 372–376.

[15] B. O'Donoghue and E. Candès, *Adaptive restart for accelerated gradient scheme*, Found. Comp. Math., 15 (2013), pp. 715–732.

[16] J. M. Ortega, *Numerical Analysis : A Second Course*, Society for Industrial and Applied Mathematics, 1990.

[17] M. R. Osborne, B. Presnell, and B. A. Turlach, *On the LASSO and its dual*, Journal of Computational and Graphical Statistics, 9 (1999), pp. 319–337.
[18] P. Moulin P. Johnstone, *A Lyapunov analysis of FISTA with local linear convergence for sparse optimization*, arXiv:1502.02281, (2015).
[19] N. Parikh and S. Boyd, *Proximal algorithms*, Foundations and Trends in Optimization, 1 (2014), pp. 127–239.
[20] S. Tao, D. Boley, and S. Zhang, *Convergence of common proximal methods for $\ell_1$ regularized least squares*, in International Joint Conference on Artificial Intelligence, 2015.
[21] H. Taylor, S. Bank, and J. McCoy., *Deconvolution with the $\ell_1$ norm*, Geophysics, 44 (1979), pp. 39–52.
[22] R. Tibshirani, *Regression shrinkage and selection via the LASSO*, Journal of the Royal Statistical Society. Series B (Methodological), 58 (1996), pp. 267–288.
[23] R. J. Tibshirani, *The LASSO problem and uniqueness*, Electronic Journal of Statistics, 7 (2013), pp. 1456–1490.
[24] H. Zhang, W. Yin, and L. Cheng, *Necessary and sufficient conditions of solution uniqueness in $\ell_1$ minimization*, CoRR, abs/1209.0652 (2012).