Convergence of Common Proximal Methods for ℓ_1-Regularized Least Squares

Shaozhe Tao, Daniel Boley, Shuzhong Zhang
University of Minnesota, Minneapolis MN 55455 USA
{taoxx120,boley,zhangs}@umn.edu

Abstract

We compare the convergence behavior of ADMM (alternating direction method of multipliers), [F]ISTA ([fast] iterative shrinkage and thresholding algorithm) and CD (coordinate descent) methods on the model ℓ_1-regularized least squares problem (aka LASSO). We use an eigenanalysis of the operators to compare their local convergence rates when close to the solution. We find that, when applicable, CD is often much faster than the other iterations, when close enough to the solution. When far from the solution, the spectral analysis implies that one can often get a sequence of iterates that appears to stagnate, but is actually taking small constant steps toward the solution. We also illustrate how the unaccelerated ISTA algorithm can sometimes be faster compared to FISTA when close enough to the solution.

1 Introduction

Many problems in machine learning and data fitting can be cast as a least squares problem with a regularization term to limit overfitting. Using an ℓ_1-norm regularization has been found to be particularly effective in many applications like feature selection [Tibshirani, 1996], compressed sensing [Chen *et al.*, 1998], sparse coding [Gregor and LeCun, 2010], and discovery of graph connectivity [Hsieh *et al.*, 2011]. In this paper, we use a model problem consisting of a linear least squares problem with ℓ_1-regularization, also known as a LASSO problem:

$$\min_{x\in\mathbb{R}^n} \tfrac{1}{2}\|A\mathbf{x}-\mathbf{b}\|_2^2+\lambda\|\mathbf{x}\|_1 \tag{1}$$

where $A\in\mathbb{R}^{m\times n}$ is a short-flat matrix (i.e. $n>m$) with full row rank, $\mathbf{b}$ is a given vector, and λ is a positive scalar. The ℓ_1 regularizer tends to produce a sparse solution, avoiding overfitting while reducing the computational cost [Tibshirani, 1996].

Most general convex solvers such as interior point methods [Ben-Tal and Nemirovski, 2001] do not scale to the large-scale data problems encountered in practice, but many algorithms have been proposed recently to take advantage of the special structure in (1). In this paper we compare four of the most popular of these methods: the Alternating Direction Method of Multipliers (ADMM) [Boyd *et al.*, 2011], Iterative Shrinkage Thresholding Algorithm (ISTA) [Parikh and Boyd, 2014] and its accelerated version Fast ISTA (FISTA) [Beck and Teboulle, 2009], and the cyclic Coordinate Descent method (CD) [Saha and Tewari, 2010].

There has been a recent flurry of activity on bounds for the convergence rates of these methods. A global convergence rate bound of $O(1/k)$ has been shown for ADMM [Eckstein and Bertsekas, 1990; Deng and Yin, 2012; He and Yuan, 2012] and ISTA, while FISTA enjoys a bound of $O(1/k^2)$ [He and Yuan, 2012]. Moreover, for ISTA, there have been recent local convergence results under some strict sparsity conditions, e.g. [Bredies and Lorenz, 2008; Liang *et al.*, 2014]. The convergence of CD in general was shown in [Tseng and Yun, 2009], and an $O(1/k)$ convergence rate for the model LASSO problem was shown in [Saha and Tewari, 2010]. In this paper, we focus on local convergence behavior, by modelling each iteration as a matrix recurrence. We find that each iteration passes through several phases or *regimes* before reaching a final regime with linear convergence, as first found in [Boley, 2013] for ADMM on general LPs and QPs. The analysis shows that ISTA can sometimes be faster than FISTA when close enough to the solution.

The paper is organized as follows. Section 2 gives some basic preliminaries of the LASSO problem. Sections 3 & 4 give the linear convergence analysis of different iterations. Numerical Examples are shown in Section 5 and conclusions are drawn in Section 6.

2 Preliminaries

2.1 Optimality condition of the LASSO problem

The first order KKT optimality conditions for the LASSO problem (1) are

$$A^{\mathrm{T}}(\mathbf{b}-A\mathbf{x})=\lambda\boldsymbol{\nu} \tag{2}$$

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

$$\left\{\begin{matrix}\boldsymbol{\nu}_i=\mathrm{sign}(\mathbf{x}_i) & \text{if } \mathbf{x}_i\neq 0\\ -1\le\boldsymbol{\nu}_i\le+1 & \text{if } \mathbf{x}_i=0\end{matrix}\right\} \quad \text{for} \quad i=1,2,\cdots. \tag{3}$$

Here the "sign" function is defined by $\mathrm{sign}(x)=1,0,-1$ according to whether $x>0, x=0, x<0$, respectively.

2.2 Uniqueness

There are various sufficient and necessary conditions for the uniqueness of the LASSO problem or its variants. For example, [Osborne *et al.*, 1999; Candès and Plan, 2009; Fuchs, 2005] showed different sufficient conditions and [Tibshirani, 2013] studied the necessary conditions for the LASSO problem. In fact, the problem (1) needs to have a unique solution in many situations. For example, in compressed sensing, having non-uniqueness solutions will result in unreliable recovery given the data. We refer readers to [Tibshirani, 2013; Zhang *et al.*, 2012] and references therein for a discussion of the uniqueness of the LASSO problem. In some cases, we will assume uniqueness of the LASSO solution.

3 Convergence of ADMM, ISTA and FISTA

3.1 Auxiliary Variables with Local Monotonic Behavior

In this section, we show that ADMM, ISTA and FISTA can be all transformed into a matrix recurrence form in a similar way. We distinguish the iterates of the different algorithms using the notation $\mathbf{x}$, $\widehat{\mathbf{x}}$ and $\widetilde{\mathbf{x}}$ to denote the iterates of ADMM, ISTA and FISTA respectively.

ADMM as a Matrix Recurrence

The ADMM is constructed by splitting the primal $\mathbf{x}$ variables into two separate variables such that the minimum with respect to each individual variable can be easily computed, and then imposing an equality constraint between the two variables. A typical splitting for LASSO problem is to use variable $\mathbf{x}$ for the least squares loss function and $\mathbf{z}$ for the l_1-norm regularizer. Then the modified LASSO problem becomes

$$\min_{\mathbf{x}} {}^1\!/_2\|A\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{z}\|_1 \quad \text{s.t. } \mathbf{x}-\mathbf{z}=0 \tag{4}$$

with augmented Lagrangian function for the resulting problem being

$$\mathcal{L}_\rho = {}^1\!/_2\|A\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{z}\|_1 + \boldsymbol{\mu}^{\mathrm{T}}(\mathbf{x}-\mathbf{z}) + {}^\rho\!/_2\|\mathbf{x}-\mathbf{z}\|_2^2 \tag{5}$$

where ρ is a penalty parameter, $\boldsymbol{\mu}$ is the dual variable. Let $\mathbf{u} = \boldsymbol{\mu}/\rho$, then the formal ADMM iterates can be represented as:

$$\begin{cases} \mathbf{x}^{[k+1]} = (A^{\mathrm{T}}A+\rho I)^{-1}[A^{\mathrm{T}}\mathbf{b} + \rho(\mathbf{z}^{[k]} - \mathbf{u}^{[k]})] \\ \mathbf{z}^{[k+1]} = \mathrm{Shr}_{\lambda\!/_\rho}(\mathbf{x}^{[k+1]} + \mathbf{u}^{[k]}) \\ \mathbf{u}^{[k+1]} = \mathrm{Thr}_{\lambda\!/_\rho}(\mathbf{x}^{[k+1]} + \mathbf{u}^{[k]}) \end{cases} \tag{6}$$

where $\mathrm{Shr}_\sigma(s) = (1-\sigma/|s|)_+ s$ and $\mathrm{Thr}_\sigma(s) = s - \mathrm{Shr}_\sigma(s)$.

We replace the iterates $\mathbf{z}^{[k]}$, $\mathbf{u}^{[k]}$ with two equivalent "auxiliary" iterates carrying the same information. One variable, namely $\mathbf{w}^{[k]}$, exhibits smooth behavior, with linear convergence locally around a fixed point, and the other variable $\mathbf{d}^{[k]}$ is a discrete ternary "flag" vector indicating which of the three cases of the shrinkage operator applies to each component. Specifically, for all k, we let the common iterate be $\mathbf{w}^{[k]} = \mathbf{z}^{[k]} + \mathbf{u}^{[k]}$ and $\mathbf{d}^{[k]}$ be a vector defined elementwise as $d_i^{[k]} = \mathrm{sign}(\mathrm{Shr}_{\lambda\!/_\rho}(w_i^{[k]}))$ and the flag matrix $D^{[k]} = \mathrm{diag}(\mathbf{d}^{[k]})$. Then one can derive

$$\mathbf{w}^{[k+1]} = M^{[k]}\mathbf{w}^{[k]} + \mathbf{h}^{[k]}$$

where

$$\begin{aligned} M^{[k]} &= (I-(D^{[k]})^2) + \rho(A^{\mathrm{T}}A+\rho I)^{-1}(2(D^{[k]})^2 - I) \\ \mathbf{h}^{[k]} &= {}^\lambda\!/_\rho\mathbf{d}^{[k]} + (A^{\mathrm{T}}A+\rho I)^{-1}(A^{\mathrm{T}}\mathbf{b} - 2\lambda\mathbf{d}^{[k]}). \end{aligned} \tag{7}$$

The ADMM update with $\mathbf{x}^{[k+1]}$, $\mathbf{z}^{[k+1]}$, $\mathbf{u}^{[k+1]}$ now can be modified in matrix form in terms of $D^{[k+1]}$ and $\mathbf{w}^{[k+1]}$ as below.

```
Algorithm 1: One pass of modified ADMM
Start with $\mathbf{w}^{[k]}$, $D^{[k]}$.
1. $\mathbf{w}^{[k+1]} = M^{[k]}\mathbf{w}^{[k]} + \mathbf{h}^{[k]}$ ($M^{[k]}, \mathbf{h}^{[k]}$ defined by (7)).
2. $D^{[k+1]} = \mathrm{DIAG}(\mathrm{sign}(\mathrm{Shr}_{\lambda\!/_\rho}(\mathbf{w}^{[k+1]})))$.
Result is $\mathbf{w}^{[k+1]}$, $D^{[k+1]}$ for next pass.
```

Note that step 1 of Alg. 1 is written as a homogeneous matrix recurrence, which will be used to characterize its convergence property.

$$\begin{pmatrix}\mathbf{w}^{[k+1]} \\ 1\end{pmatrix} = \mathbf{M}_{\mathbf{aug}}^{[k]}\begin{pmatrix}\mathbf{w}^{[k]} \\ 1\end{pmatrix} = \begin{pmatrix} M^{[k]} & \mathbf{h}^{[k]} \\ 0 & 1\end{pmatrix}\begin{pmatrix}\mathbf{w}^{[k]} \\ 1\end{pmatrix} \tag{8}$$

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

FISTA and ISTA as a Matrix Recurrence

The ISTA and FISTA iteration [Daubechies *et al.*, 2004; Beck and Teboulle, 2009] are given as follows, where $t^{[0]} = t^{[1]} = 1$ and L is the Lipschitz constant equal to $\|A^TA\|_2$.

$$\begin{cases} t^{[k+1]} = \frac{1+\sqrt{1+4(t^{[k]})^2}}{2} \\ \tau^{[k]} = 0 \text{ for ISTA, or } \frac{t^{[k]}-1}{t^{[k+1]}} \text{ for FISTA;} \\ \mathbf{v}^{[k+1]} = \widetilde{\mathbf{x}}^{[k]} + \tau^{[k]}(\widetilde{\mathbf{x}}^{[k]} - \widetilde{\mathbf{x}}^{[k-1]}) \\ \widetilde{\mathbf{x}}^{[k+1]} = \mathrm{Shr}_{\lambda\!/_L}((I - {}^1\!/_L A^{\mathrm{T}}A)\mathbf{v}^{[k+1]} + {}^1\!/_L A^{\mathrm{T}}\mathbf{b}). \end{cases} \tag{9}$$

If $\tau^{[k]} = 0$ (or equivalently $t^{[k]} = 1$) for all k, leading to $\mathbf{v}^{[k+1]} = \widetilde{\mathbf{x}}^{[k]}$, then this reduces to the ISTA updates, specifically,

$$\widehat{\mathbf{x}}^{[k+1]} = \mathrm{Shr}_{\lambda\!/_L}((I - {}^1\!/_L A^{\mathrm{T}}A)\widehat{\mathbf{x}}^{[k]} + {}^1\!/_L A^{\mathrm{T}}\mathbf{b}). \tag{10}$$

Similar to ADMM, we use auxiliary variables $\widetilde{\mathbf{w}}^{[k]}, \widetilde{D}^{[k]}$ to replace variable $\widetilde{\mathbf{x}}^{[k]}$ to carry the FISTA iterations. Set $\widetilde{\mathbf{w}}^{[k]} = (I - {}^1\!/_L A^{\mathrm{T}}A)\mathbf{v}^{[k]} + {}^1\!/_L A^{\mathrm{T}}\mathbf{b}$ and the vector $\widetilde{\mathbf{d}}^{[k]}$ is defined elementwise as $\widetilde{d}_i^{[k]} = \mathrm{sign}(\mathrm{Shr}_{\lambda\!/_L}(\widetilde{w}_i^{[k]}))$ and the flag matrix $\widetilde{D}^{[k]} = \mathrm{diag}(\widetilde{\mathbf{d}}^{[k]})$. The iteration can be described as

$$\widetilde{\mathbf{w}}^{[k+1]} = P^{[k]}\widetilde{\mathbf{w}}^{[k]} + Q^{[k-1]}\widetilde{\mathbf{w}}^{[k-1]} + \bar{\mathbf{h}}^{[k]}$$

where we denote

$$\begin{aligned} \tau^{[k]} &= \tfrac{t^{[k]}-1}{t^{[k+1]}} \quad \text{(for FISTA, or 0 for ISTA)} \\ \widetilde{R}^{[k]} &= (I - {}^1\!/_L A^{\mathrm{T}}A)(\widetilde{D}^{[k]})^2 \\ P^{[k]} &= (1+\tau^{[k]})\widetilde{R}^{[k]} \\ Q^{[k-1]} &= -\tau^{[k]}\widetilde{R}^{[k-1]} \\ \bar{\mathbf{h}}^{[k]} &= {}^1\!/_L A^{\mathrm{T}}\mathbf{b} \; + \; (I - {}^1\!/_L A^{\mathrm{T}}A)\times \\ &\quad \left[-(1+\tau^{[k]}){}^\lambda\!/_L\widetilde{\mathbf{d}}^{[k]} + \tau^{[k]}\,{}^\lambda\!/_L\widetilde{\mathbf{d}}^{[k-1]}\right]. \end{aligned} \tag{11}$$

Therefore, the FISTA update (9) can be written using the new auxiliary variables $\widetilde{\mathbf{w}}$ and $\widetilde{D}$ as follows.

```
Algorithm 2: 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. Set $t^{[k+1]} = \frac{1+\sqrt{1+4(t^{[k]})^2}}{2}$ so that $\tau^{[k]} = \frac{t^{[k]}-1}{t^{[k+1]}}$.
2. Set $\widetilde{\mathbf{w}}^{[k+1]} = P^{[k]}\widetilde{\mathbf{w}}^{[k]} + Q^{[k-1]}\widetilde{\mathbf{w}}^{[k-1]} + \bar{\mathbf{h}}^{[k]}$
      (with $P^{[k]}, Q^{[k-1]}, \bar{\mathbf{h}}^{[k]}$ defined by (11)).
3. Set $\widetilde{D}^{[k+1]} = \textsc{Diag}(\text{sign}(\text{Shr}_{\lambda/L}(\widetilde{\mathbf{w}}^{[k+1]})))$.
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 2 of above procedure can also be formulated as a homogeneous matrix recurrence analogous to (8) for ADMM with a larger (approximately double) dimension:

$$\begin{pmatrix} \widetilde{\mathbf{w}}^{[k+1]} \\ \widetilde{\mathbf{w}}^{[k]} \\ 1 \end{pmatrix} = \begin{pmatrix} N^{[k]} & \widetilde{\mathbf{h}}^{[k]} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{w}}^{[k]} \\ \widetilde{\mathbf{w}}^{[k-1]} \\ 1 \end{pmatrix} \tag{12}$$

where we denote $N^{[k]} = \begin{pmatrix} P^{[k]} & Q^{[k-1]} \\ I & 0 \end{pmatrix}$ and $\widetilde{\mathbf{h}}^{[k]} = \begin{pmatrix} \bar{\mathbf{h}}^{[k]} \\ 0 \end{pmatrix}$ and $\mathbf{N}_{\mathbf{aug}}^{[k]} = \begin{pmatrix} N^{[k]} & \widetilde{\mathbf{h}}^{[k]} \\ 0 & 1 \end{pmatrix}$ in the remainder of this paper. For ISTA, Alg. 2 reduces to

$$\begin{cases} 1. \quad \begin{pmatrix} \widehat{\mathbf{w}}^{[k+1]} \\ 1 \end{pmatrix} = \begin{pmatrix} R^{[k]} & \widehat{\mathbf{h}}^{[k]} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{w}}^{[k]} \\ 1 \end{pmatrix} \\ 2. \quad \widehat{D}^{[k+1]} = \textsc{Diag}(\text{sign}(\text{Shr}_{\lambda/L}(\widehat{\mathbf{w}}^{[k+1]}))) \end{cases} \tag{13}$$

where $R^{[k]} = (I - {}^1\!/_L A^{\mathrm{T}}A)(\widehat{D}^{[k]})^2$ and $\widehat{\mathbf{h}}^{[k]} = -(I - {}^1\!/_L A^{\mathrm{T}}A){}^\lambda\!/_L\widehat{\mathbf{d}}^{[k]} + {}^1\!/_L A^{\mathrm{T}}\mathbf{b}$. We denote $\mathbf{R}_{\mathbf{aug}}^{[k]}$ as $\begin{pmatrix} R^{[k]} & \widehat{\mathbf{h}}^{[k]} \\ 0 & 1 \end{pmatrix}$, the augmented matrix of $R^{[k]}$, in this paper.

3.2 Spectral Properties and Four Regimes

It is seen that $\mathbf{M}_{\mathbf{aug}}^{[k]}$, $\mathbf{R}_{\mathbf{aug}}^{[k]}$ and $\mathbf{N}_{\mathbf{aug}}^{[k]}$ play key roles in the convergence. We summarize their respective properties.

Lemma 3.1 *Suppose $D^{[k-1]} = D^{[k]} = D^{[k+1]}$, (and same for $\widehat{D}^{[k]}$ and $\widetilde{D}^{[k]}$). Then the iteration matrices $M^{[k]}$* (7)*, $R^{[k]}$* (13)*, and $N^{[k]}$* (11) *&* (12) *(for ADMM, ISTA and FISTA, respectively) have the following properties:*
(a). $\|M^{[k]}\|_2 \le 1$, $\|R^{[k]}\|_2 \le 1$, $\|N^{[k]}\|_2 \le 1$.
(b). *All eigenvalues or $M^{[k]}, R^{[k]}, N^{[k]}$ lie in the closed disk in the complex plane with center ${}^1\!/_2$ and radius ${}^1\!/_2$, denoted as $\mathcal{D}({}^1\!/_2, {}^1\!/_2)$, and the eigenvalues of $R^{[k]}$ are real.*
(c). *The eigenvalue 1, if it exists, must have a complete set of eigenvectors (no Jordan blocks larger than 1×1).*

Proof. We temporarily omit the pass number ${}^{[k]}$. For M (ADMM) we have:
(a). Observe $2D^2 - I$ is an orthogonal matrix by $(2D^2 - I)(2D^2 - I) = I$. Hence, $\|M\|_2 = \|M(2D^2 - I)\|_2 = \|D^2 - I + ({}^1\!/_\rho A^{\mathrm{T}}A + I)^{-1}\|_2 \le 1$.
(b). $\|M - {}^1\!/_2 I\|_2 = \|({}^1\!/_\rho A^{\mathrm{T}}A + I)^{-1} - {}^1\!/_2 I\|_2 \le {}^1\!/_2$. Thus the eigenvalues of $M - {}^1\!/_2 I$ lie in the closed disk $\mathcal{D}(0, {}^1\!/_2)$. The eigenvalues of M lie in the disk $\mathcal{D}({}^1\!/_2, {}^1\!/_2)$, which is entirely in the open right half plane plus the origin.

The proof of (c) and all the cases for ISTA and FISTA follow the same lines as in [Boley, 2013; Tao *et al.*, 2015] and hence are omitted. □

Lemma 3.1 gives rise to the four possible "regimes" associated with the ADMM, ISTA and FISTA iterations, depending on the flag matrix and the eigenvalues of operators $\mathbf{M}_{\mathbf{aug}}$, $\mathbf{R}_{\mathbf{aug}}$, $\mathbf{N}_{\mathbf{aug}}$. 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.

When flag matrix remains unchanged from iteration k to $k+1$, ($D^{[k]} = D^{[k+1]}$ (or $\widehat{D}^{[k]} = \widehat{D}^{[k+1]}$, $\widetilde{D}^{[k]} = \widetilde{D}^{[k+1]}$)):

Regime [A]. The spectral radius of $M^{[k]}$ (or $R^{[k]}$, $N^{[k]}$) is strictly less than 1. If close enough to the optimal solution (if it exists), the result is linear convergence to that solution. The convergence rate depends on the second largest eigenvalue of $\mathbf{M}_{\mathbf{aug}}$ (or $\mathbf{R}_{\mathbf{aug}}$, $\mathbf{N}_{\mathbf{aug}}$, resp.) according to the theory of power method for the matrix eigenvalue problem [Golub and Loan, 2013].

Regime [B]. Also known as "constant step regime", $M^{[k]}$ (or $R^{[k]}$, $N^{[k]}$) has an eigenvalue equal to 1 but which yields a 2×2 Jordan block $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ for eigenvalue 1 for the augmented matrix. Then the iteration process tends to a constant step (eigenvector of $M^{[k]}$ (or $R^{[k]}$, $N^{[k]}$) for eigenvalue 1). The iteration will continue in this way until the discrete flag matrix D changes. Such a change is guaranteed to occur due to the global convergence of the algorithms.

Regime [C]. $M^{[k]}$ (or $R^{[k]}$, $N^{[k]}$) has an eigenvalue equal to 1, but the augmented matrix still has a complete set of eigenvectors for eigenvalue 1 (this eigenvalue has no 2×2 Jordan block). The iterates will converge to an eigenvector for eigenvalue 1 with a linear rate as in Regime [A]. This cannot occur in the final regime if the original optimization problem has a unique solution.

When flag matrix changes at next iteration, i.e. $D^{[k]} \neq D^{[k+1]}$, then we have:

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

3.3 Local Linear Convergence

Now we present part of our main results below. Essentially, when close enough to the optimal solution, we can give a guarantee that eventually the flag matrix will not change. Then the iterations of ADMM, ISTA and FISTA eventually behave like the power method for eigenvalues such that they converge linearly to the optimal solution. The proof of the following theorem follows [Boley, 2013] (for ADMM) and [Tao *et al.*, 2015] (for F/ISTA).

Theorem 3.2 *Suppose the LASSO problem* (1) *has a unique solution and this solution has strict complementarity: that is for every index i, $\mathbf{w}_i^* \neq \pm{}^\lambda\!/_\rho$ for ADMM, $\widehat{\mathbf{w}}_i^* \neq \pm{}^\lambda\!/_L$ for ISTA, $\widetilde{\mathbf{w}}_i^* \neq \pm{}^\lambda\!/_L$ for FISTA. Then eventually the ADMM iteration used specifically for the LASSO problem* (6)*, the ISTA* *iteration* (10)*, and the FISTA iteration* (9) *all reach a stage where the iteration converges linearly to that unique solution.*

The eigenanalysis in terms of regimes established in [Tao *et al.*, 2015] allows one not only to study the local convergence behavior but also to explain the whole iteration behavior. The following two propositions show that how FISTA can be viewed as an accelerated ISTA process and FISTA may slow down compared to ISTA when close enough to the optimal solution.

Proposition 3.3 *In regime [B], the constant step vector for ISTA is* $\mathbf{v}$*, where* $\mathbf{v} = R\mathbf{v}$ *is a scaled eigenvector of* R *for eigenvalue* 1*, while the constant step vector for FISTA is* $\frac{1}{1-\tau^{[k]}}\mathbf{v}$*. Since* $\tau^{[k]} \to 1$*, the constant stepsize of FISTA is much larger, which yields a speedup.*

Proposition 3.4 *In regime [A] or [C], let* $\beta_{\max}, \gamma_{\max}$ *be the largest eigenvalue of* R *and* N *respectively, then FISTA is faster than ISTA if* $1 > \beta_{\max} > \tau^{[k]} > 0$ *but slower if* $1 > \tau^{[k]} > \beta_{\max} > 0$*. Since* $\beta_{\max}$ *is a fixed value for one specific regime, with the* τ *growing to* 1*, ISTA will be faster than FISTA toward the end. Besides, when* $\tau > \beta_{\max}$*,* $\gamma_{\max}$ *must be one of a complex conjugate pair.*

The proof follows the same lines as [Tao *et al.*, 2015]. In practice, if $\beta_{\max}$ is well separated from 1, then it is advantageous to make FISTA iterations switch to ISTA once it reaches the final regime. This idea is implemented in Section 5 called Hybrid F/ISTA.

4 Convergence of CD

4.1 Local linear convergence

The cyclic CD is widely used due to its easy update rule. Essentially it goes through and updates all of the components in a cyclic fashion instead of updating them simultaneously as the gradient descent method. In this section, we establish a natural relationship between the iterations of the cyclic CD and of the Gauss-Seidel method when close enough to the optimal solution of problem (1) so that linear convergence is guaranteed eventually under some mild conditions.

We denote $\mathbf{y}$ as the iterates of CD, $\mathbf{y}_i^{[k]}$ as the i-th coordinate of $\mathbf{y}^{[k]}$ and A_i as the i-th column of A. All coordinates other than i are denoted as $-i$. The CD updates for problem (1) is as follows.

```
Algorithm 3: One pass of CD
Start with $\mathbf{y}^{[k]}$.
for coordinate $i = 1, 2, \cdots, n$,
    Set $\mathbf{y}_i^{[k+1]} = \text{Shr}_\lambda \Big( A_i\mathbf{b} - \sum_{j=1}^{i-1}(A^{\mathrm{T}}A)_{ij}\mathbf{y}_j^{[k+1]}$
     $- \sum_{j=i+1}^{|n|}(A^{\mathrm{T}}A)_{ij}\mathbf{y}_j^{[k]} \Big) / \|A_i^T A_i\|$.
end for
Result is $\mathbf{y}^{[k+1]}$ to the next pass.
```

Our result is motivated by the following key observation.

Lemma 4.1 *Suppose the solution to problem* (1) *has strict complementarity as in Theorem 3.2, then there exists a* K *such that for all* $k > K$*, the CD iterate* $\mathbf{y}^{[k]}$ *is close enough to the optimal solution* $\mathbf{y}^*$ *that sign*$(\mathbf{y}^{[k]})$ *is fixed.*

Proof. We first define the following index set based on $\mathbf{y}^*$

$$\mathcal{E} = \{i \in \{1, \cdots, n\} : \mathbf{y}_i^* \neq 0\} \tag{14}$$

and $\overline{\mathcal{E}}$ as the complement set of $\mathcal{E}$. Consequently,

$$\mathbf{y}^* = \begin{bmatrix} \mathbf{y}_{\mathcal{E}}^* \\ \mathbf{0} \end{bmatrix} \text{ (with } \mathbf{y}_{\mathcal{E}}^* \text{ all non-zero).}$$

Based on the notation, it also can be seen that $A_{\mathcal{E}}$ is composed of the columns corresponding to the nonzero element of $\mathbf{y}_{\mathcal{E}}^*$. For the simplicity of our analysis and without loss of generality, we split matrix A based on $\mathcal{E}$ and permute the columns so that $A = [A_{\mathcal{E}}, A_{\overline{\mathcal{E}}}]$. Then the optimality condition of problem (1) then can be rewritten as

$$\begin{array}{ll} \text{(a)} & A_{\mathcal{E}}^T\mathbf{b} - A_{\mathcal{E}}^T A_{\mathcal{E}}\mathbf{y}_{\mathcal{E}}^* - A_{\mathcal{E}}^T A_{\overline{\mathcal{E}}}\mathbf{y}_{\overline{\mathcal{E}}}^* = \lambda\mathbf{d}_{\mathcal{E}} \\ \text{(b)} & A_{\overline{\mathcal{E}}}^T\mathbf{b} - A_{\overline{\mathcal{E}}}^T A_{\mathcal{E}}\mathbf{y}_{\mathcal{E}}^* - A_{\overline{\mathcal{E}}}^T A_{\overline{\mathcal{E}}}\mathbf{y}_{\overline{\mathcal{E}}}^* = \lambda\mathbf{d}_{\overline{\mathcal{E}}}, \end{array}$$

where $\mathbf{d}_{\mathcal{E}}$ is composed of elements $\{\pm 1\}$ based on sign$(\mathbf{y}_{\mathcal{E}}^*)$, and $\mathbf{d}_{\overline{\mathcal{E}}}$ is composed of elements strictly between -1 and $+1$ by the assumption of strict complementarity. Let δ be the largest entry in absolute value found in $\mathbf{d}_{\overline{\mathcal{E}}}$ so that

$$\text{Shr}_\lambda \left(A_{\overline{\mathcal{E}}}^T\mathbf{b} - A_{\overline{\mathcal{E}}}^T A_{\mathcal{E}}\mathbf{y}_{\mathcal{E}}^* - A_{\overline{\mathcal{E}}}^T A_{\overline{\mathcal{E}}}\mathbf{y}_{\overline{\mathcal{E}}}^* \right) = 0.$$

Now consider the case of $\overline{\mathcal{E}}$. Let $\mathbf{y}^{[k]}$ be a vector very close to $\mathbf{y}^*$ such that $\|\mathbf{y}^{[k]} - \mathbf{y}^*\|_\infty < \epsilon_1$. Then $|\mathbf{y}_l - \mathbf{y}_l^*| < \epsilon_1$ $\forall l \in \mathcal{E}$ and $|\mathbf{y}_j| < \epsilon_1$ $\forall j \in \overline{\mathcal{E}}$. If ϵ_1 is sufficiently small, then (b) above induces

$$\|A_{\overline{\mathcal{E}}}^T A_{\mathcal{E}}\mathbf{y}_{\mathcal{E}}^{[k]} - A_{\overline{\mathcal{E}}}^T A_{\overline{\mathcal{E}}}\mathbf{y}_{\overline{\mathcal{E}}}^{[k]} - A_{\overline{\mathcal{E}}}^T\mathbf{b}\|_\infty < \lambda(\delta + c_1\epsilon_1) < \lambda,$$

so that $\text{Shr}_\lambda(\mathbf{y}_{\overline{\mathcal{E}}}^{[k]}) = 0$. Here c_1 is some constant magnification factor depending only on A. Since $\mathbf{y}^{[k]}$ converges to $\mathbf{y}^*$, it would imply that the future iterates, starting from $\mathbf{y}^{[k]}$, would have zeros in the $\overline{\mathcal{E}}$ part.

Next consider the case of $\mathcal{E}$. Let $\epsilon_2 = \min_{\mathcal{E}}\{|\mathbf{y}_{\mathcal{E}}^*|\} - c_2$ for a positive constant c_2 sufficiently small to make $\epsilon_2 > 0$. And for each $l \in \mathcal{E}$, we define a ball around each $\mathbf{y}_l^*$ that $\mathcal{B}(\mathbf{y}_l^*) = \{\mathbf{y}_l : \|\mathbf{y}_l - \mathbf{y}_l^*\|_\infty \leq \epsilon_2\}$.

According to Theorem 16 in [Saha and Tewari, 2010], cyclic CD converges to LASSO problem at the rate of $O(1/k)$. Hence there must exist an iteration number K that for all $k > K$,

$$\|\mathbf{y}^{[k]} - \mathbf{y}^*\|_\infty \leq \epsilon = \min\{\epsilon_1, \epsilon_2\}$$

which implies $\mathbf{y}_{\overline{\mathcal{E}}}^{[k]} = 0$ and $\mathbf{y}_l^{[k]} \in \mathcal{B}(\mathbf{y}_l^*)$, $\forall l \in \mathcal{E}$. In other words, for each component i, $\mathbf{y}_i^{[k]}$ must fall in one of three cases: $\mathbf{y}_i < 0$, $\mathbf{y}_i = 0$ and $\mathbf{y}_i > 0$ and never jump out to another case. □

Theorem 4.2 *Suppose the optimal solution* $\mathbf{y}^*$ *of problem* (1) *is sparse such that* $A_{\mathcal{E}}^{\mathrm{T}} A_{\mathcal{E}}$ *is positive definite, with* $\mathcal{E}$ *defined in* (14)*. Then eventually the cyclic coordinate descent iteration reaches a stage where it converges linearly to the solution.*

Proof. From Lemma 4.1, when close enough to the optimal solution, sign$(\mathbf{y}^{[k]})$ is fixed. So $\mathbf{y}_{\overline{\mathcal{E}}}^{[k]}$ remained zero for future iterations. As for $\mathbf{y}_{\mathcal{E}}^{[k]}$, we can simplify the resulting CD updates in Alg. 3 by eliminating $\mathbf{y}_j^{[k]}$ $(\forall j \in \overline{\mathcal{E}})$ so that $\forall l \in \mathcal{E}$,

$$\mathbf{y}_l^{[k+1]} = \Big(A_l\mathbf{b} - \lambda\widehat{\mathbf{d}}_l^{[k]} - \textstyle\sum_{j=1}^{l-1}(A^{\mathrm{T}}A)_{lj}\mathbf{y}_j^{[k+1]} \\ - \textstyle\sum_{j=l+1}^{|\mathcal{E}|}(A^{\mathrm{T}}A)_{lj}\mathbf{y}_j^{[k]}\Big) / \|A_l^T A_l\|. \tag{15}$$

Assuming λ is large enough such that optimal solution $\mathbf{y}^*$ is sparse and $A_{\mathcal{E}}^{\mathrm{T}}A_{\mathcal{E}}$ is positive definite, then (15) is equivalent to the Gauss-Seidel method applied to

$$A_{\mathcal{E}}^{\mathrm{T}}A_{\mathcal{E}}\mathbf{y}_{\mathcal{E}} = (A^{\mathrm{T}}\mathbf{b})_{\mathcal{E}} - \lambda\mathbf{d}_{\mathcal{E}} \tag{16}$$

where elements of $\mathbf{d}_{\mathcal{E}}$ is fixed and equal to either 1 or -1. The iteration must converge linearly by the theory of Gauss-Seidel method [Golub and Loan, 2013]. □

Remark: we note here that $A_{\mathcal{E}}^{\mathrm{T}}A_{\mathcal{E}}$ being positive definite is a mild assumption in practice. Since the optimal solution is sparse, $|\mathcal{E}| < m$ can be satisfied easily for λ not too small, and in many applications this is sufficient to guarantee that all columns of $A_{\mathcal{E}}$ are linearly independent. If not, one can increase λ to increase the sparsity.

4.2 Comparison with ISTA

In this part, we show that CD should converge faster than ISTA when both CD and ISTA iterations are in their final regimes from the viewpoint of preconditioning. The next lemma shows the equivalence of the ISTA iteration and the classical Richardson iteration [Kelley, 1995].

Lemma 4.3 *When ISTA iteration reaches the final regime, the regime of linear convergence, then the ISTA iteration is equivalent to the Richardson iteration for solving the linear system* (16).

Proof. When ISTA reaches the final regime, according to Theorem 3.2 $\widehat{\mathbf{w}} = (I - {}^1\!/_L A^{\mathrm{T}}A)\widehat{\mathbf{x}} + {}^1\!/_L A^{\mathrm{T}}\mathbf{b}$ would fall in into three cases: $\widehat{\mathbf{w}} < -L$, $-L < \widehat{\mathbf{w}} < L$, $\widehat{\mathbf{w}} > L$, and never jump out to another case. Hence the shrinkage operator in updating step (10) is fixed so that ISTA reduces to

$$\widehat{\mathbf{x}}_{\mathcal{E}}^{[k+1]} = (I - {}^1\!/_L A_{\mathcal{E}}^{\mathrm{T}}A_{\mathcal{E}})\widehat{\mathbf{x}}_{\mathcal{E}}^{[k]} + {}^1\!/_L((A^{\mathrm{T}}\mathbf{b})_{\mathcal{E}} - \lambda\widehat{\mathbf{d}}_{\mathcal{E}}) \tag{17}$$

where $\widehat{\mathbf{d}}_{\mathcal{E}} = \mathrm{sign}(\mathbf{x}_{\mathcal{E}}^*)$. And $\widehat{\mathbf{x}}_{\overline{\mathcal{E}}}^{[k+1]} = 0$. The resulting ISTA updates (17) is exactly Richardson iteration [Kelley, 1995] for solving (16). □

Combining the result of Theorem 4.2 and Lemma 4.3, if both CD and ISTA reach their final regimes, then CD is equivalent to Gauss-Seidel iteration and ISTA is equivalent to Richardson iteration for solving the same linear system (16). Let T and U be the diagonal and strict upper triangular part of $A_{\mathcal{E}}^{\mathrm{T}}A_{\mathcal{E}}$ and L is the lipschitz constant. Gauss-Seidel iteration can be written as the preconditioned Richardson iteration as below (and hence generally faster):

$$\begin{aligned}\mathbf{y}^{[k+1]} &= (I - (T+U^{\mathrm{T}})^{-1}A_{\mathcal{E}}^{\mathrm{T}}A_{\mathcal{E}})\mathbf{y}^{[k]} \\ &\quad + (T+U^{\mathrm{T}})^{-1}((A^{\mathrm{T}}\mathbf{b})_{\mathcal{E}} - \lambda\mathbf{d}_{\mathcal{E}})\end{aligned} \tag{18}$$

with preconditioner $L(T+U^{\mathrm{T}})^{-1}$ [Kelley, 1995].

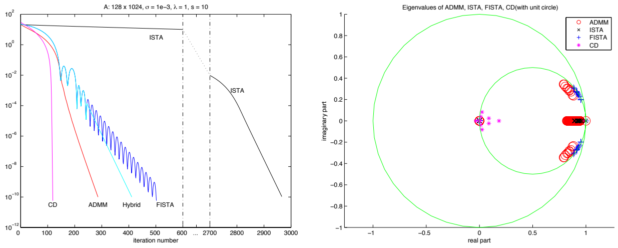

Figure 1: (Left): Convergence behavior in terms of error of iterates of ADMM, ISTA, FISTA, Hybrid F/ISTA and CD for the instance marked $\boxed{**}$ in Table 1. (Right): Spectrum of operators of all iterations on the left during the final regime. The unit circle and $\mathcal{D}({}^1\!/_2, {}^1\!/_2)$ on the complex plane are shown for reference.

5 Numerical Examples

We consider examples of compressed sensing to show different convergence behaviors to support our analysis. Suppose there exists a true sparse signal represented by a n-dimensional vector $\mathbf{x}_s$ with s 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

$$\mathbf{b} = A\mathbf{x}_s + \epsilon \tag{19}$$

where ϵ is the observation noise. The goal is to recover the sparse vector $\mathbf{x}_s$ from the measurement matrix A and observation $\mathbf{b}$. This is a form of feature selection, where we discover the most relevant features. We let $A \in \mathbb{R}^{m\times n}$ be Gaussian matrix whose elements are *i.i.d* distributed as $\mathcal{N}(0,1)$, ϵ be a vector whose elements are *i.i.d* distributed as $\mathcal{N}(0,\sigma^2)$ with $\sigma = 10^{-3}$.

The numerical results are summarized in Table 1. We compare the convergence behavior of ADMM, ISTA, FISTA and CD in terms of the total number of iterations (Total #), number of iterations to reach final linear regime (Final #) and the linear rate in the final regime (Rate), with different λ and sparsity s. Based on Propositions 3.3 & 3.4, we also report the behavior of a hybrid F/ISTA that follows FISTA during initial iterations to reach the final regime and then switch to ISTA until convergence. In general, after comparison with the linear rates, we see that, CD is often much faster than the other iterations during the final linear regime.

Fig. 1 illustrates the methods' behavior for the instance marked $\boxed{**}$ in Table 1, to show how the theorems established before explains the behaviors in practice. Fig. 1(Left) shows the convergence behavior of all algorithms. Observe that all of them pass through a few transitions in the early part of the iterations and then settle on their final linear regimes. Fig. 1(Right) shows the eigenvalues during the final regimes for all algorithms. One can see that the eigenvalues of operators $\mathbf{M_{aug}}, \mathbf{R_{aug}}, \mathbf{N_{aug}}$ lie strictly inside the circle $\mathcal{D}({}^1\!/_2, {}^1\!/_2)$ (except for 0 & 1), consistent with Lemma 3.1. Based on this instance, we also observe the following:

(a). It costs FISTA many fewer steps (245 iterations) than ISTA (2823 iterations) to get to the final regime. The main reason is that FISTA has much larger constant steps in regime

Problem Setting	Total#	Final#	Rate	Total#	Final#	Rate	Total#	Final#	Rate
$m = 64, n = 512$	$s = 7, \lambda = 0.3$			$s = 7, \lambda = 1$			$s = 7, \lambda = 5$		
ADMM	758	711	0.897	271	216	0.897	116	44	0.899
ISTA	6824	6749	0.958	2143	2062	0.958	583	438	0.954
FISTA	541	402	0.975	393	226	0.973	289	84	0.960
Hybrid F/ISTA	512	402	0.958	313	226	0.958	216	84	0.954
CD	399	394	0.141	124	119	0.141	28	25	0.076
$m = 64, n = 512$	$s = 14, \lambda = 0.3$			$s = 14, \lambda = 1$			$s = 14, \lambda = 5$		
ADMM	943	804	0.9632	447	268	0.963	286	87	0.962
ISTA	10000	-	-	9321	7713	0.996	2927	1561	0.994
FISTA	1226	961	0.996	1367	546	0.995	1054	758	0.995
Hybrid F/ISTA	2193	961	0.996	2211	546	0.996	1532	758	0.994
CD	1011	952	0.878	350	294	0.878	102	63	0.822
$m = 128, n = 1024$	$s = 10, \lambda = 0.3$			$\boxed{**}$ $s = 10, \lambda = 1$ $\boxed{**}$			$s = 10, \lambda = 5$		
ADMM	611	498	0.946	286	151	0.946	193	33	0.946
ISTA	9390	9273	0.959	2966	2823	0.959	776	596	0.959
FISTA	693	507	0.976	502	245	0.973	298	99	0.965
Hybrid F/ISTA	622	507	0.959	411	245	0.959	266	99	0.959
CD	383	377	0.184	119	114	0.184	32	24	0.184
$m = 128, n = 1024$	$s = 25, \lambda = 0.3$			$s = 25, \lambda = 1$			$s = 25, \lambda = 5$		
ADMM	1760	1541	0.980	807	566	0.980	544	225	0.980
ISTA	10000	-	-	10000	-	-	7315	5857	0.996
FISTA	2024	2021	0.997	1777	1724	0.997	1528	863	0.996
Hybrid F/ISTA	2306	2021	0.996	2245	1724	0.996	2017	863	0.996
CD	2821	2754	0.881	893	820	0.883	237	159	0.892

Table 1: Examples of compressed sensing with different problem settings. λ is the parameter in problem (1) and s is the number of non-zero elements of optimal solution. (Total #): the total number of iterations with maximum 10^4. (Final #): number of iterations before reaching final linear regime. (Rate): The linear rate (eigenvalue) in the final regime. $\boxed{**}$: The instance illustrated in Fig. 1.

[B] so that it can more quickly reach the end of the stagnating regime [B], as suggested by Proposition 3.3. In fact, one can show that the difference between consecutive iterates of ISTA remains a constant for many iterations while FISTA does not. Moreover, since the rate of ISTA is 0.959, well separated from 1, Proposition 3.4 predicts switching to ISTA in the final regime should converge faster than standard FISTA. Indeed, hybrid F/ISTA converges in 411 iterations compared to 502 iterations for FISTA.

(b). FISTA oscillates in the final regime. This is because the second largest eigenvalue of FISTA operator $\mathbf{N_{aug}}$ is a pair of complex conjugates (cf. Lemma 3.1 & Proposition 3.4). Hence, according to the theory of the power method, the convergence will oscillate between the two conjugate complex numbers. Since all the eigenvalues of ISTA operator $\mathbf{R_{aug}}$ are real, ISTA iterates do not oscillate.

(c). Implied by the observation in Section 4.2, in the final regime, CD is equivalent to a preconditioned ISTA-like Richardson method for solving the same linear system. In this instance, we can see the preconditioning plays a big role and the eigenvalues of CD are much smaller than ISTA shown in Fig. 1(Right).

(d). We remark that the cost per iteration is about the same ($O(n^2)$) for all the methods. ISTA, FISTA, and CD require the equivalent of a matrix-vector product. ADMM requires an LU-factorization of $(A^{\mathrm{T}}A + \rho I)$, but only once at the beginning, and requires only a forward-back substitution in each iteration. Hence iteration counts closely reflect total times.

6 Conclusion

In this paper, we show the locally linear convergence of ADMM, ISTA, FISTA and CD applied to the LASSO problem. We model ADMM, ISTA and FISTA as the matrix recurrence form and connect them with the power method. We also establish a connection between CD method and Gauss-Seidel method. By spectral analysis, we show that all of the iterations normally pass through several regimes of different types and eventually settle on a "linear regime" in which the iterates converge linearly.

Such analysis provides a way to study the behavior through the whole iteration process. We explain why ISTA often appears to stagnate during the initial iterations, and why FISTA oscillates towards the end. Besides, we illustrate how the unaccelerated ISTA can sometimes be faster when close enough to the solution compared to FISTA and propose the idea of switching to ISTA in certain circumstances.

Acknowledgements

This research was supported in part by NSF grant IIS-1319749.

References

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

[Ben-Tal and Nemirovski, 2001] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. SIAM, 2001.

[Boley, 2013] D. Boley. Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs. *SIAM J. on Optimization*, 23(4):2183–2207, 2013.

[Boyd *et al.*, 2011] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[Bredies and Lorenz, 2008] K. Bredies and D. A. Lorenz. Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, 14(5-6):813–837, 2008.

[Candès and Plan, 2009] E. J. Candès and Y. Plan. Near-ideal model selection by l1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

[Chen *et al.*, 1998] S. Chen, D. L. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. on Scientific Computing*, 20:33–61, 1998.

[Daubechies *et al.*, 2004] 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(11):1413–1457, 2004.

[Deng and Yin, 2012] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Rice CAAM technical report TR12-14*, 2012.

[Eckstein and Bertsekas, 1990] J. Eckstein and D. P. Bertsekas. An alternating direction method for linear programming. MIT Lab. for Info. and Dec. Sys. report LIDS-P-1967, 1990.

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

[Golub and Loan, 2013] G. H. Golub and V. Loan. *Matrix Computations*. Johns Hopkins Univ. Press, 4th edition, 2013.

[Gregor and LeCun, 2010] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. ICML, pages 399–406, 2010.

[He and Yuan, 2012] B. He and X. Yuan. On the $O(1/t)$ convergence rate of alternating direction method. *SIAM J. on Optimization*, 22(4):1431–1448, 2012.

[Hsieh *et al.*, 2011] C. J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar. Sparse inverse covariance matrix estimation using quadratic approximation. *NIPS*, pages 2330–2338, 2011.

[Kelley, 1995] C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*. Frontiers in Applied Mathematics. SIAM, 1995.

[Liang *et al.*, 2014] J. Liang, J. Fadili, and G. Peyr. Local linear convergence of forward–backward under partial smoothness. In *NIPS*, pages 1970–1978. 2014.

[Osborne *et al.*, 1999] M. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 1999.

[Parikh and Boyd, 2014] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[Saha and Tewari, 2010] A Saha and A Tewari. On the finite time convergence of cyclic coordinate descent methods. *CoRR*, abs/1005.2146, 2010.

[Tao *et al.*, 2015] S. Tao, D. Boley, and S. Zhang. Local linear convergence of ista and fista on the lasso problem. `arXiv:1501.02888 [math.OC]`, 2015.

[Tibshirani, 1996] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[Tibshirani, 2013] R. J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7(0):1456–1490, 2013.

[Tseng and Yun, 2009] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009.

[Zhang *et al.*, 2012] H. Zhang, W. Yin, and L. Cheng. Necessary and sufficient conditions of solution uniqueness in ℓ_1 minimization. *CoRR*, abs/1209.0652, 2012.