CONVERGENCE OF AN INERTIAL PROXIMAL METHOD FOR L1-REGULARIZED LEAST-SQUARES

Patrick R Johnstone and Pierre Moulin

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Email: prjohns2@illinois.edu, moulin@ifp.uiuc.edu

ABSTRACT

A fast, low-complexity, algorithm for solving the ℓ_1-regularized least-squares problem is devised and analyzed. Our algorithm, which we call the Inertial Iterative Soft-Thresholding Algorithm (I-ISTA), incorporates inertia into a forward-backward proximal splitting framework. We show that the iterates of I-ISTA converge linearly to a minimum with a better rate of convergence than the well-known Iterative Shrinkage/Soft-Thresholding Algorithm (ISTA) for solving ℓ_1-regularized least-squares. The improvement in convergence rate over ISTA is significant on ill-conditioned problems and is gained with minor additional computations. We conduct numerical experiments which show that I-ISTA converges more quickly than ISTA and two other computationally comparable algorithms on compressed sensing and deconvolution problems.

Index Terms— Inertial forward-backward proximal splitting, heavy ball method, gradient descent with momentum, compressed sensing, deconvolution.

1. INTRODUCTION

We are interested in problems of the following form:

$$\min_x \{f(x) + r(x)\}, \tag{1}$$

where $x \in \mathbb{R}^n$, f is proper, convex and differentiable and r is proper, convex, lower semi-continuous, but not necessarily differentiable. The ℓ_1-regularized least-squares problem, which we call Problem ℓ_1-LS, corresponds to (1) when $f(x)$ is a quadratic and $r(x)$ is the ℓ_1 norm of x. It is also referred to as basis pursuit de-noising (BPDN). Given $A \in \mathbb{R}^{m\times n}$, $y \in \mathbb{R}^m$ and a regularization parameter $\gamma \geq 0$, Problem ℓ_1-LS is

$$\min_x \left\{ \frac{1}{2}\|y - Ax\|^2 + \gamma \|x\|_1 \right\}, \tag{2}$$

where $\|\cdot\|$ is the Euclidean norm and $\|\cdot\|_1$ is the ℓ_1-norm. Since the objective is coercive and bounded below by 0, problem ℓ_1-LS always has a solution.

Problem ℓ_1-LS has important applications in machine learning, signal processing, statistics and compressed sensing [1, 2, 3, 4, 5]. In compressed sensing, a sparse vector x_0 is measured with a sensing matrix A, to form $y = Ax_0 + e$, where e is measurement noise. We would like to recover x_0 from y, however the number of measurements m is less than n. Nonetheless, under certain conditions on the sparsity level, measurement matrix and γ, a solution to Problem ℓ_1-LS accurately approximates x_0 [6, 7, 8].

This work was supported in part by NSF under grant CCF 11-17980.

Many approaches to solving Problem ℓ_1-LS have been proposed in recent years, such as, interior point methods [9], path-following homotopy methods [8] and active-set identification methods [10]. In this paper we focus on first-order *iterative shrinkage/soft-thresholding* approaches because of their simplicity, computational efficiency and scalability [11]. In particular we focus on *forward-backward splitting* (FBS) [12, 13, 14]. FBS describes a family of first-order iterative methods for solving problems in the form of (1).

Inertial methods involve a history of previous iterates in the update of the next iterate [15, 16]. These methods were inspired by the discretization of differential equations and can accelerate convergence [17, 18]. In this paper, we propose adding an inertial term to ISTA, a well-known FBS algorithm for solving Problem ℓ_1-LS [19].

1.1. Contributions in this Paper

We propose a new method for solving Problem ℓ_1-LS. We prove that the iterates of the algorithm converge linearly to a minimum and determine the parameters which optimize the convergence rate. We show that the algorithm has a better convergence rate than ISTA. Finally, we present numerical simulations on four important instances of Problem ℓ_1-LS showing that our algorithm is faster than ISTA and two other well-known first-order iterative shrinkage and thresholding algorithms.

1.2. Notation

A sequence a^k is said to *converge linearly* to its limit a^*, with *rate of convergence* q, if

$$\limsup_{k\to\infty} \frac{\|a^{k+1} - a^*\|}{\|a^k - a^*\|} = q \in (0, 1).$$

For any $v \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ and $F \subseteq \{1, 2, \ldots, n\}$, let the following definitions hold. Let v_F be the $|F|$-dimensional vector with elements equal to v on the subset F. Let $\text{supp}(v) \triangleq \{i : v_i \neq 0\}$. Let $\lambda_{\max}$ be the largest eigenvalue of A^TA. Let A_F be the $m \times |F|$ matrix formed by selecting the columns of A corresponding to F. Let $\lambda_1, \lambda_2, \ldots, \lambda_{|F|}$ be the eigenvalues of $A_F^TA_F$ in increasing order of magnitude and let $\lambda^F_{\min} \triangleq \lambda_1$ and $\lambda^F_{\max} \triangleq \lambda_{|F|}$.

We use S_ν to denote the shrinkage/soft-thresholding operator, defined as:

$$S_\nu(a) \triangleq \max\{|a| - \nu, 0\} \operatorname{sgn}(a) \tag{3}$$

where $\operatorname{sgn}(a) = 1$ if $a \geq 0$ and -1 otherwise. $S_\nu(v)$ refers to the result of applying the soft-thresholding operator element-wise to v.

2. PREVIOUS ALGORITHMS

2.1. FBS

FBS is a well-known approach to solving (1). It relies on *proximity operators*. The proximity operator of a convex function r is

$$\text{prox}_r(y) \triangleq \arg\min_z r(z) + \frac{1}{2}\|z-y\|^2. \quad (4)$$

Since r is convex and lower semi-continuous, $\text{prox}_r(y)$ is a well-defined function. For many important choices of r, prox_r is inexpensive to compute [13]. The iterations of FBS are

$$x^{k+1} = \text{prox}_{\tau_k r}\left(x^k - \tau_k \nabla f(x^k)\right), \quad (5)$$

starting at an arbitrary initial point x_0 and with explicit prescriptions on the step-size τ_k.

2.2. ISTA

For Problem ℓ_1-LS, a well-known algorithm is ISTA, see [11, 19, 20, 21]. ISTA belongs to the family of FBS methods [22]. The proximity operator of the ℓ_1-norm is the soft-thresholding operator S_ν given in (3). Rate of convergence results for ISTA are much stronger than for the general FBS family. Let $x^*_{\text{IST}} \triangleq \lim_{k\to\infty} x^k$ and $F \triangleq \text{supp}(x^*_{\text{IST}})$. Assume $\lambda^F_{\min} > 0$, which is a necessary condition in compressed sensing when F corresponds to the support set of x_0 [6]. It was shown in [22] that the iterates of ISTA converge linearly to x^*_{IST} with rate q_0 where

$$q_0 \triangleq \frac{\lambda_{\max} - \lambda^F_{\min}}{\lambda_{\max} + \lambda^F_{\min}}, \quad (6)$$

which is achieved by choosing $\tau_k = 2/(\lambda_{\max} + \lambda^F_{\min})$ for all k. If $\lambda_{\max} \gg \lambda^F_{\min}$, the rate is very close to 1, leading to very slow convergence of ISTA.

2.3. The Heavy Ball Method

The *Heavy Ball with Friction* method (HBF) is an inertial method for minimizing unconstrained strongly-convex quadratic functions [17]. It was inspired by discretizing the differential equation describing the motion of a particle under the effects of friction and a conservative force [15, 16]. The iterates of HBF for minimizing f are

$$x^{k+1} = x^k - \tau\nabla f(x^k) + \beta(x^k - x^{k-1}). \quad (7)$$

where τ and β are scalars. HBF is equivalent to the standard gradient descent algorithm (GD) with an additional inertia term: $\beta(x^k - x^{k-1})$. The inertia term is also called a *momentum* term and has been used in other settings [23, 24, 25]. The parameters τ and β are determined by the condition number of the Hessian and β is always in $[0, 1)$. The effect of the inertia is to smooth out the rapid changes in the gradient descent direction which can occur on ill-conditioned problems, leading to slow convergence of GD. The inertia causes HBF to converge more quickly than GD [17, 23].

2.4. Inertial Forward-Backward Splitting

Moudafi and Oliny extended HBF to a more general setting which includes Problem (1) [26]. In this paper we will refer to their proposal as Inertial Forward-Backward Splitting (I-FBS). The iterates of I-FBS are

$$x^{k+1} = \text{prox}_{\tau_k r}\left(x^k - \tau_k\nabla f(x^k) + \beta_k(x^k - x^{k-1})\right). \quad (8)$$

In fact, I-FBS can be applied to the more general problem of finding a zero of the sum of two maximal-monotone operators, of which Problem (1) is a special case. Note that I-FBS is not the same as the more famous Fast Iterative Soft-Thresholding Algorithm (FISTA) [27], which computes the gradient at the extrapolated point. FISTA has different convergence guarantees.

2.5. Convergence Properties of I-FBS

To finish the section on previous work, we present the following result on the convergence of I-FBS methods.

Theorem 1 ([18, 26]). *Suppose that, for all k, $0 \le \tau_k \le \overline{\tau} < 2/\lambda_{\max}$, the β_k's are non-decreasing and $0 < \beta_k \le \overline{\beta} < 1/3$. Then the output x^k of I-FBS converges to a solution to Problem ℓ_1-LS.*

While Theorem 1 guarantees the convergence of I-FBS methods under the restrictions on β_k and τ_k, the type of convergence is not known (i.e. linear, quadratic etc.). Furthermore, effective choices within the constraints of the sequences τ_k and β_k are also not known. In the following sections we address these short comings for Problem ℓ_1-LS by introducing I-ISTA.

3. I-ISTA

Our proposed algorithm, I-ISTA, is given in Algorithm 1. It is similar to ISTA but with the addition of the all-important inertia term which accelerates the rate of convergence. It is a member of the I-FBS family of methods since it can be written in the form of (8). Our analysis allows us to provide much stronger guarantees than what is known for the more general I-FBS family along with a more precise prescription for τ and β. We consider fixed τ and β across iterations. It is possible to extend our analysis to the more general case with some loss of readability and clarity.

Algorithm 1 I-ISTA

Require: The tuple (A, y, γ) which is an instance of Problem ℓ_1-LS, a starting point x^0, constants β and τ and a stopping criterion.

1: Set $k = 0$, $x^{-1} = x^0$.
2: **while** stopping criterion is not satisfied **do**
3: $\quad x^{k+1} = S_{\tau\gamma}\left[x^k - \tau A^T(Ax^k - y) + \beta(x^k - x^{k-1})\right]$.
4: $\quad k = k + 1$.
5: **end while**
6: **return** x^k

4. ASYMPTOTIC EQUIVALENCE OF I-ISTA AND HBF

Our first result states that after a finite number of iterations, I-ISTA becomes equivalent to HBF applied to an explicit quadratic function. This fact will allow us to determine the rate of convergence of I-ISTA by considering the behavior of HBF applied to this quadratic. Due to space limitations we cannot include the proof which can be found on pages 22-24 of [28].

Theorem 2 ([28]). *Assume the conditions of Theorem 1 hold for a particular instance of Problem ℓ_1-LS and choices of τ and β, then the sequence x^k converges to some x^*, which is a solution to Problem ℓ_1-LS. Let $E \triangleq supp(x^*)$. Under the strict-complementarity condition (Eq. (4.22) of [22]) there exists a constant $K > 0$ such that for $k > K$, the support set of x^k is E, and, the sequence x^k_E*

for $k > K$ is equal to that generated by HBF applied to minimize a strongly convex quadratic function $\phi(u)$ for $u \in \mathbb{R}^{|E|}$, starting from some initial point. Finally, the Hessian of ϕ is $A_E^T A_E$.

The proof of Theorem 2 proceeds in the spirit of [22]. We show that after finitely many iterations, I-ISTA becomes equivalent to heavy-ball method applied to minimize a quadratic function with Hessian given by $A_E^T A_E$ on the set E. Outside of E, the iterates are 0. The strict complementarity condition of [22] is that, for all $i \notin \text{supp}(x^*)$, $|a_i^T(Ax^* - y)| < \gamma$. This condition can be relaxed [28].

5. RATE OF CONVERGENCE OF I-ISTA

We will now choose τ and β to improve upon the rate of convergence of ISTA, which is q_0 given in (6). As in Theorem 2, let $E \triangleq \text{supp}(\lim_{k\to\infty} x^k)$. Let $\kappa_E \triangleq \lambda^E_{\max}/\lambda^E_{\min}$ and let $\kappa' \triangleq \lambda_{\max}/\lambda^E_{\min}$.

Theorem 3. *Let C be some constant such that $0 < C < 1/3$. Let*

$$\beta^* \triangleq \max\left\{\left(\frac{\sqrt{\kappa_E}-1}{\sqrt{\kappa_E}+1}\right)^2, \left(1-\sqrt{\frac{2}{\kappa'+1}}\right)^2\right\}, \quad (9)$$

and

$$\tau^* \triangleq \frac{2}{\lambda_{\max}+\lambda^E_{\min}}. \quad (10)$$

If $\beta = \min\{\beta^*, C\}$, $\tau = \tau^*$, the strict-complementarity condition (Eq. (4.22) of [22]) holds, and $\lambda^E_{\min} > 0$, then the iterates of I-ISTA with parameters τ and β converge linearly to a solution of Problem ℓ_1-LS. Since $\lambda^E_{\min} > 0$, ISTA converges to the same minimizer. The rate of convergence of I-ISTA is $q(\tau,\beta)$ with $q(\tau,\beta) < q_0$, where q_0 is the rate of convergence of ISTA. Furthermore if $\beta^* < C$, $q(\tau,\beta) = \sqrt{\beta^*}$.

Proof. According to Theorem 2, there exists a K such that for $k > K$, the iterations of I-ISTA satisfy

$$\begin{bmatrix} x_E^{k+1} - x_E^* \\ x_E^k - x_E^* \end{bmatrix} = M \begin{bmatrix} x_E^k - x_E^* \\ x_E^{k-1} - x_E^* \end{bmatrix} \quad (11)$$

where

$$M \triangleq \begin{bmatrix} (1+\beta)I_{|E|\times|E|} - \tau A_E^T A_E & -\beta I_{|E|\times|E|} \\ I_{|E|\times|E|} & \mathbf{0}_{|E|\times|E|} \end{bmatrix} \quad (12)$$

and $x_i^k = 0$ for all $i \notin E$. Which implies that for $k > K$,

$$\|x^k - x^*\| = \|x_E^k - x_E^*\| \le C(q(\tau,\beta)+\epsilon_k)^k, \quad (13)$$

where $q(\tau,\beta)$ is greater than or equal to the maximum magnitude of all eigenvalues of M and $\lim_{k\to\infty}\epsilon_k = 0$ [17].

Let the eigenvalues of M be ρ_j, for $j = 1, 2, \ldots, 2|E|$. These are equal to the eigenvalues of the 2×2 matrices:

$$\begin{bmatrix} 1+\beta-\tau\lambda_i & -\beta \\ 1 & 0 \end{bmatrix}, \quad (14)$$

for $i = 1, 2, \ldots, |E|$. Recall that $\lambda_1, \lambda_2, \ldots, \lambda_{|E|}$ are the eigenvalues of $A_E^T A_E$, in order of increasing magnitude. Therefore the ρ_j are the $2|E|$ roots of the equations

$$\rho^2 - \rho(1+\beta-\tau\lambda_i) + \beta = 0, \quad i = 1, 2, \ldots, |E|. \quad (15)$$

Let $\rho_i^+(\tau,\beta) \triangleq \rho_{2i-1}$ and $\rho_i^-(\tau,\beta) \triangleq \rho_{2i}$ for $i = 1, 2, \ldots, |E|$. Then

$$\rho_i^{\pm}(\tau,\beta) = \frac{(1+\beta-\tau\lambda_i) \pm \sqrt{(1+\beta-\tau\lambda_i)^2 - 4\beta}}{2}. \quad (16)$$

where we have collected the $2|E|$ eigenvalues into pairs of solutions for each of the $|E|$ quadratic equations. Now if

$$(1+\beta-\tau\lambda_i)^2 - 4\beta \le 0 \quad (17)$$

then $|\rho_i^+(\tau,\beta)| = |\rho_i^-(\tau,\beta)| = \sqrt{\beta}$.

(17) holds for every i if and only if (iff)

$$\frac{(1-\sqrt{\beta})^2}{\lambda_i} \le \tau \le \frac{(1+\sqrt{\beta})^2}{\lambda_i}, \quad i = 1, 2, \ldots, |E|. \quad (18)$$

Now τ can be chosen to satisfy condition (18) iff

$$\frac{(1-\sqrt{\beta})^2}{\lambda^E_{\min}} \le \frac{(1+\sqrt{\beta})^2}{\lambda^E_{\max}}. \quad (19)$$

(19) holds iff

$$\beta \ge \left(\frac{\sqrt{\kappa_E}-1}{\sqrt{\kappa_E}+1}\right)^2. \quad (20)$$

Furthermore, we would like to be able to choose $\tau = 2/(\lambda_{\max} + \lambda^E_{\min})$. So we enforce:

$$\frac{(1-\sqrt{\beta})^2}{\lambda^E_{\min}} \le \frac{2}{\lambda_{\max}+\lambda^E_{\min}} \le \frac{(1+\sqrt{\beta})^2}{\lambda^E_{\max}}, \quad (21)$$

which is true if

$$\beta \ge \left(1-\sqrt{\frac{2}{\kappa'+1}}\right)^2. \quad (22)$$

Now $\beta = \beta^*$ satisfies both conditions (20) and (22). Also $\tau^* < 2/\lambda_{\max}$. If $\beta^* < C < 1/3$, then the choice $\beta = \beta^*$ and $\tau = \tau^*$ satisfies the conditions of Theorem 1. Therefore the iterates of I-ISTA with this choice converge linearly to a minimum with rate $q(\tau^*,\beta^*) = \sqrt{\beta^*}$. It can be verified that $\sqrt{\beta^*} < q_0$.

Now consider the case where $\beta^* > C$. We must examine the eigenvalues, $\rho_i^{\pm}(\tau,\beta)$ in more detail. For all i,

$$1-\tau\lambda_i > 0 \implies \rho_i^+(\tau,0) = 1-\tau\lambda_i \text{ and } \rho_i^-(\tau,0) = 0,$$

and

$$1-\tau\lambda_i < 0 \implies \rho_i^+(\tau,0) = 0 \text{ and } \rho_i^-(\tau,0) = 1-\tau\lambda_i.$$

Now, by considering the derivative with respect to β, one can see that $\rho_i^+(\tau,\beta)$ is monotone decreasing in β for $\beta < \beta^*$, for all i and for all $\tau > 0$. Similarly, $\rho_i^-(\tau,\beta)$ is monotone increasing in β. Therefore $\max\left\{|\rho_i^+(\tau,\beta)|, |\rho_i^-(\tau,\beta)|\right\}$ is monotone decreasing in β for $\beta \le \beta^*$, and

$$q(\tau,\beta) = \max_i \left\{\max\left\{|\rho_i^+(\tau,\beta)|, |\rho_i^-(\tau,\beta)|\right\}\right\}$$

is monotone decreasing in β for $\beta < \beta^*$. Finally, the choice $\tau = \tau^*$ and $\beta = C$ satisfies the conditions of Theorem 1. Thus the iterates of I-ISTA converge linearly with rate of convergence equal to $q(\tau^*, C) < q(\tau^*, 0) = q_0$. ∎

Remarks on Theorem 3

The function ϕ defined in Theorem 2 has Lipschitz continuous gradient. Consequently, Theorem 3 also implies linear convergence of $F(x^k)$ to the optimal value of Problem ℓ_1-LS, with the same rate.

Empirical studies suggest that the restriction of β to be less than $1/3$ in Theorem 1 is not a necessary condition of convergence. Extending the theoretical guarantee to all $\beta \in [0, 1)$ is an important direction of future research.

6. PRACTICAL CONSIDERATIONS FOR I-ISTA

The optimal choice of τ and β depends on κ_E and κ', which depend on the support set of the limit. These quantities could be estimated based on known properties of A and the estimated sparsity level. The estimates could be iteratively updated using the support set of the current x^k. Alternatively, we recommend the following simple rule of thumb based on the optimal choices determined in Theorem 3. Set $\tau = 2/\lambda_{\max}$ and

- $0.9 < \beta < 1$, if A_E is expected to be poorly conditioned (i.e. $\kappa_E \gg 100$),
- $0.5 < \beta \leq 0.9$, if A_E is expected to be moderately conditioned (i.e. $\kappa_E \approx 100$) and
- $0 \leq \beta \leq 0.5$, if A_E is expected to be well-conditioned (i.e. $\kappa_E \approx 1$).

The more information we have about κ' and κ_E, the closer we can get to choosing the optimal β. Strictly we should choose $\tau = \frac{2}{\lambda_{\max}} - \epsilon$ with an arbitrarily small ϵ, however in practice ϵ can be set to 0.

Empirical studies suggest that too much inertia can slow convergence in the early iterations of I-ISTA. We believe this is because the support of x^k changes dramatically in the early iterations and too much inertia can interfere with this process. However a small amount of inertia does appear to help. Consequently, we suggest using an increasing sequence of inertia parameters. We used $\beta_k = \max\{0, \beta - 1/k\}$, with β chosen using the rule of thumb above.

7. SIMULATIONS

We present the results of four simulations that compare the performance of ISTA, I-ISTA, FISTA [27] and TwIST [29]. We consider these two iterative methods alongside ISTA and I-ISTA because they have comparable computational complexity per iteration. They are also inertial methods and involve the previous two iterates in the update of the next iterate. We now outline the four experiments.

Experiment 1 *(compressed sensing)*: the matrix A is 500×1000 and had entries drawn i.i.d. from $\mathcal{N}(0, 1/500)$. $\gamma = 5 \times 10^{-3}$.

Experiment 2 *(compressed sensing)*: the matrix A is 500×1000 and had values drawn i.i.d. from a scaled version of the Rademacher distribution, taking value $\pm 1/\sqrt{500}$, with equal probability. $\gamma = 1 \times 10^{-3}$.

Experiment 3 *(deconvolution)*: F is a 1049×1049 Toeplitz matrix with columns given by incrementally shifted copies of a length-50 constant rectangular window, H is the inverse Haar wavelet matrix truncated to size 1049×1049 and $A = FH$. $\gamma = 2$.

Experiment 4 *(deconvolution)*: F is a 1069×1069 Toeplitz matrix with columns given by incrementally shifted copies of a length-70 triangular window. H is the inverse Haar wavelet matrix truncated to size 1069×1069 and $A = FH$. $\gamma = 2$.

In all Experiments, $y = Ax_0 + e$, where e had i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. In Experiment 1, $\sigma^2 = 0.05^2$, and in Experiments 2, 3 and 4, $\sigma^2 = 0.01^2$. In Experiments 1, 3 and 4, x_0 is 250-sparse and in Experiment 2, x_0 is 100-sparse. The non-zero entries of x_0 are drawn i.i.d. $\mathcal{N}(0, 1)$ and the support is drawn uniformly at random.

Let $F(x) \triangleq \frac{1}{2}\|y - Ax\|^2 + \gamma\|x\|_1$. We used the interior point method described in [9] to compute the optimal value F^* of Problem ℓ_1-LS , to a *relative duality gap* of 10^{-8}. Let $\hat{F}$ be the value computed by the interior point method. Note that a relative duality gap of 10^{-8} does not imply $\hat{F} - F^* = 10^{-8}$. For a sequence of estimates x^k, let $e^k \triangleq |F(x^k) - \hat{F}|$. Note that e^k will saturate to $\hat{F} - F^*$. The results of Experiments 1 through 4 can be seen in Figs. 1 and 2. We plotted the log of e^k against iteration number to better illustrate the various convergence rates achieved by the algorithms.

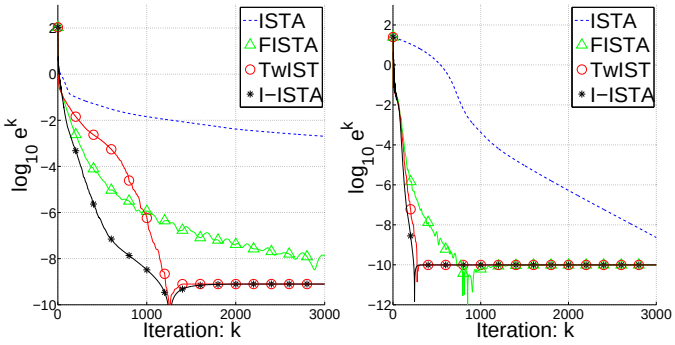

Fig. 1: $\log_{10} e^k$ versus iteration number k, compressed sensing experiments. Left: Experiment 1, right: Experiment 2.

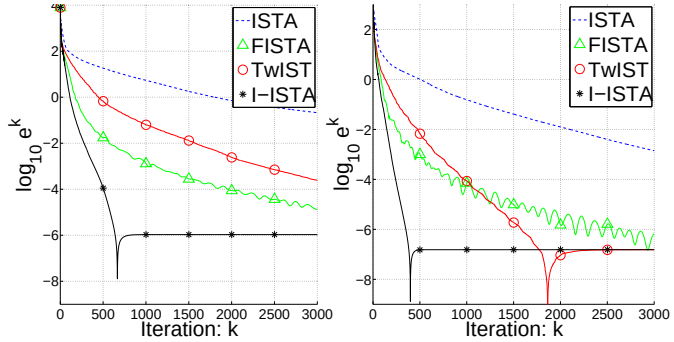

Fig. 2: $\log_{10} e^k$ versus iteration number k, deconvolution experiments. Left: Experiment 3, Right: Experiment 4.

For ISTA, we used $\tau = 2/\lambda_{\max}$. For FISTA we used the non-monotone algorithm with parameters as specified in [27]. For TwIST we used the monotone version with backtracking and parameters as specified in [29]. For I-ISTA, in all experiments we used $\tau = 2/\lambda_{\max}$ and $\beta_k = \max\{0, \beta - 1/k\}$. For Experiment 1, we used $\beta = 0.99$, for Experiment 2, $\beta = 0.9$. and for Experiments 3 and 4, we used $\beta = 0.95$. All algorithms were initialized at $x^0 = 0$.

Discussion of Simulation Results: The results of all the experiments show that ISTA is much slower than the other three accelerated methods on these sorts of problems. This is not surprising as the solution to Problem ℓ_1-LS was not very sparse in any of the experiments and ISTA is only expected to be competitive in the highly sparse regime. FISTA is slower than I-ISTA. We believe this is because the parameters of FISTA are chosen without taking into account the expected conditioning of A_E. The speeds of I-ISTA and TwIST are similar on Experiment 2. However, in Experiments 1, 3 and 4, I-ISTA was faster. Choosing β intelligently based on knowledge of the operator A allows I-ISTA to outperform the other methods. Furthermore the performance of TwIST on instances of Problem ℓ_1-LS where A is singular - as is the case in all four experiments - is hard to predict as there is no performance guarantee in this case [29].

Acknowledgments Thanks to A. Emad, Prof. O. Milenkovic and Prof. A. Nedich for many illuminating and helpful discussions.

8. REFERENCES

[1] Robert Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society. Series B*

(Methodological), pp. 267–288, 1996.

[2] Scott Shaobing Chen, David L Donoho, and Michael A Saunders, "Atomic decomposition by basis pursuit," *SIAM journal on scientific computing*, vol. 20, no. 1, pp. 33–61, 1998.

[3] Alfred M Bruckstein, David L Donoho, and Michael Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009.

[4] Kiryung Lee, Yoram Bresler, and Marius Junge, "Oblique pursuits for compressed sensing," *Information Theory, IEEE Transactions on*, vol. 59, no. 9, pp. 6111–6141, 2013.

[5] Wei Dai and Olgica Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *Information Theory, IEEE Transactions on*, vol. 55, no. 5, pp. 2230–2249, 2009.

[6] Emmanuel J. Candès and Terence Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[7] David L Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.

[8] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al., "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[9] S-J Kim, Kwangmoo Koh, Michael Lustig, Stephen Boyd, and Dimitry Gorinevsky, "An interior-point method for large-scale ℓ_1-regularized least squares," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 1, no. 4, pp. 606–617, 2007.

[10] Patrick R Gill, Albert Wang, and Alyosha Molnar, "The in-crowd algorithm for fast basis pursuit denoising," *Signal Processing, IEEE Transactions on*, vol. 59, no. 10, pp. 4595–4605, 2011.

[11] Michael Elad, "Why simple shrinkage is still relevant for redundant representations?," *Information Theory, IEEE Transactions on*, vol. 52, no. 12, pp. 5559–5569, 2006.

[12] Pierre-Louis Lions and Bertrand Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

[13] Patrick L Combettes and Jean-Christophe Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212. Springer, 2011.

[14] Osman Güler, "On the convergence of the proximal point algorithm for convex minimization," *SIAM Journal on Control and Optimization*, vol. 29, no. 2, pp. 403–419, 1991.

[15] R. Ioan Bot, E. R. Csetnek, and S. László, "An inertial forward-backward algorithm for the minimization of the sum of two nonconvex functions," *ArXiv e-prints*, Oct. 2014.

[16] Hédy Attouch, Juan Peypouquet, and Patrick Redont, "A dynamical approach to an inertial forward-backward algorithm for convex minimization," *SIAM Journal on Optimization*, vol. 24, no. 1, pp. 232–256, 2014.

[17] Boris T Polyak, *Introduction to Optimization*, Optimization Software Inc., 1987.

[18] Felipe Alvarez and Hedy Attouch, "An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping," *Set-Valued Analysis*, vol. 9, no. 1-2, pp. 3–11, 2001.

[19] Ingrid Daubechies, Michel Defrise, and Christine De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on pure and applied mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[20] Mário AT Figueiredo and Robert D Nowak, "An EM algorithm for wavelet-based image restoration," *Image Processing, IEEE Transactions on*, vol. 12, no. 8, pp. 906–916, 2003.

[21] Pierre Moulin and Juan Liu, "Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors," *Information Theory, IEEE Transactions on*, vol. 45, no. 3, pp. 909–919, 1999.

[22] Elaine T. Hale, Wotao Yin, and Yin Zhang, "Fixed-point continuation for ℓ_1-minimization: methodology and convergence," *SIAM J. on Optimization*, vol. 19, no. 3, pp. 1107–1130, Oct. 2008.

[23] Ning Qian, "On the momentum term in gradient descent learning algorithms," *Neural networks*, vol. 12, no. 1, pp. 145–151, 1999.

[24] Patrick R Johnstone, Amin Emad, Olgica Milenkovic, and Pierre Moulin, "RFIT: A new algorithm for matrix rank minimization," in *The 5th Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS 2013)*, Lausanne, Switzerland, July 2013.

[25] Patrick R Johnstone, "Inertial iterative thresholding with applications to sparse and low-rank signal recovery," M.S. thesis, University of Illinois at Urbana-Champaign, USA, August 2014. Available at: http://hdl.handle.net/2142/50628.

[26] A. Moudafi and M. Oliny, "Convergence of a splitting inertial proximal method for monotone operators," *Journal of Computational and Applied Mathematics*, vol. 155, no. 2, pp. 447 – 454, 2003.

[27] Amir Beck and Marc Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Img. Sci.*, vol. 2, no. 1, pp. 183–202, Mar. 2009.

[28] Patrick R. Johnstone and Pierre Moulin, "A Lyapunov Analysis of FISTA with Local Linear Convergence for Sparse Optimization," *ArXiv e-print arXiv:1502.02281*, Feb. 2015.

[29] José M Bioucas-Dias and Mário AT Figueiredo, "A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration," *Image Processing, IEEE Transactions on*, vol. 16, no. 12, pp. 2992–3004, 2007.