MIST: l_0 SPARSE LINEAR REGRESSION WITH MOMENTUM

Goran Marjanovic [♭] *Magnus O. Ulfarsson* [♮ †] *Alfred O. Hero III* [♯ ‡]

[♭] School of Electrical Eng., University of New South Wales, Sydney, AUSTRALIA
[♮] Dept. Electrical Eng., University of Iceland, Reykjavik, ICELAND
[♯] Dept. Electrical Eng. and Computer Science, University of Michigan, Ann Arbor, MI, USA

ABSTRACT

Significant attention has been given to minimizing a penalized least squares criterion for estimating sparse solutions to large linear systems of equations. The penalty induces sparsity and the natural choice is the so-called l_0 norm. In this paper we develop a Momentumized Iterative Shrinkage Thresholding (MIST) algorithm for minimizing the resulting non-convex criterion and prove its convergence to a local minimizer. Simulations on large data sets show superior performance of the proposed method to other methods.

***Index Terms*—** sparsity, non-convex, l_0 regularization, iterative shrinkage thresholding, momentum

1. INTRODUCTION

In the current age of big data acquisition there has been an ever growing interest in sparse representations, which consists of representing, say, a noisy signal as a linear combination of very few components. This has huge benefits in analysis, processing and storage of high dimensional signals. As a result, sparse linear regression has been widely studied with many applications in signal and image processing, statistical inference as well as machine learning. The linear regression model is given by:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon},$$

where $\mathbf{y}_{d\times 1}$ is a vector of noisy data observations, $\mathbf{x}_{m\times 1}$ is the sparse representation (vector) of interest, $\mathbf{A}_{d\times m}$ is the regression matrix and $\boldsymbol{\epsilon}_{d\times 1}$ is the observation noise. The estimation aim is to choose the simplest model, i.e., the sparsest $\mathbf{x}$, that adequately explains the data $\mathbf{y}$. To estimate $\mathbf{x}$, major attention has been given to minimizing a sparsity Penalized Least Squares (PLS) criterion [1–10]. The least squares term promotes goodness-of-fit of the estimator while the penalty shrinks its coefficients to zero. Here we consider the non-convex l_0 penalty since it is the natural sparsity promoting penalty and induces maximum sparsity. The resulting non-convex l_0 PLS criterion is:

$$F(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_0, \tag{1}$$

where $\lambda > 0$ is the tuning parameter and $\|\mathbf{x}\|_0$ represents the number of non-zeros in $\mathbf{x}$ (l_0 penalty).

† This work was partly supported by the Research Fund of the University of Iceland and the Icelandic Research Fund (130635-051).

‡ This work was partially supported by ARO grant W911NF-11-1-0391.

1.1. Previous Work

Existing algorithms for directly minimizing (1) fall into the category of Iterative Shrinkage Thresholding (IST), and rely on the Majorization-Minimization (MM) type procedures, see [1, 10]. These procedures exploit separability properties of the l_0 PLS criterion, and thus, rely on the minimizers of one dimensional versions of the PLS function: the so-called hard-thresholding operators. Since the convex l_1 PLS criterion has similar separability properties, some MM procedures developed for its minimization could with modifications be applied to minimize (1). Applicable MM procedures include first order methods and their accelerated versions [8, 11–14]. However, when these are applied to the l_0 penalized problem (1) there is no guarantee of convergence, and for [8] there is additionally no guarantee of algorithm stability.

Analysis of convergence of MM algorithms for minimizing the l_0 PLS criterion (1) is rendered difficult due to lack of convexity. As far as we are aware, algorithm convergence for this problem has only been shown for the Iterative Hard Thresholding (IHT) method [1, 10]. Specifically, a bounded sequence generated by IHT was shown to converge to the set of local minimizers of (1) when the singular values of $\mathbf{A}$ are strictly less than one. Convergence analysis of algorithms designed for minimizing the l_q PLS criterion, $q \in (0, 1]$, is not applicable to the case of the l_0 penalized objective (1) because it relies on convex arguments when $q = 1$, and continuity and/or differentiability of the criterion when $q \in (0, 1)$.

This papers contribution is a new MM algorithm with momentum acceleration, called Momentumized IST (MIST), for minimizing the l_0 PLS criterion (1) along with a proof of its convergence to a single local minimizer without any assumptions on $\mathbf{A}$. Simulations on large data sets are carried out, which show that the proposed algorithm outperforms existing methods for minimizing (1), including modified MM methods originally designed for the l_1 PLS criterion.

The paper is organized as follows. Section 2 reviews some of background on MM that will be used to develop the proposed convergent algorithm. The proposed algorithm is given in Section 3, and Section 4 contains the convergence analysis. Lastly, Section 5 and 6 presents the simulations and concluding remarks respectively.

An extended version of this paper is available [15] with more examples and more detailed proofs.

1.2. Notation

$\mathbf{v}[i]$ is the i-th entry of vector $\mathbf{v}$. $\|\mathbf{M}\|$ is the spectral norm of matrix $\mathbf{M}$. $\mathbb{I}(\cdot)$ is the indicator function = 1 if its argument is true, and 0 otherwise. So, $\|\mathbf{v}\|_0 = \sum_i \mathbb{I}(\mathbf{v}[i] \neq 0)$. sgn$(\cdot)$ is the sign function. $\{\mathbf{x}_k\}_{k\geq 0}$ is a sequence, and $\{\mathbf{x}_{k_n}\}_{n\geq 0}$ a subsequence ($k_n \leq k_{n+1}$).

2. PRELIMINARIES

Denoting the least squares term in (1) by: $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$, the Lipschitz continuity of $\nabla f(\cdot)$ implies:

$$f(\mathbf{z}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{z} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{z} - \mathbf{x}\|_2^2$$

for all $\mathbf{x}, \mathbf{z}, \mu \geq \|\mathbf{A}\|^2$. For the proof see [8, Lemma 2.1]. As a result, the following approximation of the objective function $F(\cdot)$ in (1),

$$Q_\mu(\mathbf{z}, \mathbf{x}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{z} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{z} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{z}\|_0 \quad (2)$$

is a majorizing function, i.e.,

$$F(\mathbf{z}) \leq Q_\mu(\mathbf{z}, \mathbf{x}) \text{ for any } \mathbf{x}, \mathbf{z}, \mu \geq \|\mathbf{A}\|^2. \quad (3)$$

Let $\mathcal{P}_\mu(\mathbf{x})$ be any point in the set $\arg\min_{\mathbf{z}} Q_\mu(\mathbf{z}, \mathbf{x})$, we have:

$$F(\mathcal{P}_\mu(\mathbf{x})) \overset{(3)}{\leq} Q_\mu(\mathcal{P}_\mu(\mathbf{x}), \mathbf{x}) \leq Q_\mu(\mathbf{x}, \mathbf{x}) = F(\mathbf{x}), \quad (4)$$

where the stacking of (3) above the first inequality indicates that this inequality follows from Eq. (3). The proposed algorithm is constructed using the above MM framework with a momentum acceleration designed based on the following:

Theorem 1. Let $\mathbf{B}_\mu = \mu\mathbf{I} - \mathbf{A}^T\mathbf{A}$, where $\mu > \|\mathbf{A}\|^2$, and:

$$\alpha = 2\eta\left(\frac{\boldsymbol{\delta}^T\mathbf{B}_\mu(\mathcal{P}_\mu(\mathbf{x}) - \mathbf{x})}{\boldsymbol{\delta}^T\mathbf{B}_\mu\boldsymbol{\delta}}\right),\ \eta \in [0, 1], \quad (5)$$

where $\boldsymbol{\delta} \neq 0$. Then, $F\left(\mathcal{P}_\mu(\mathbf{x} + \alpha\boldsymbol{\delta})\right) \leq F(\mathbf{x})$.

For the proof see the Appendix.

2.1. Evaluating the Operator $\mathcal{P}_\mu(\cdot)$

Since (2) is non-convex there may exist multiple minimizers of $Q_\mu(\mathbf{z}, \cdot)$ so that $\mathcal{P}_\mu(\cdot)$ may not be unique. We select a single element of the set of minimizers as described below. By simple algebraic manipulations of the quadratic quantity in (2), letting:

$$g(\mathbf{x}) = \mathbf{x} - (1/\mu)\nabla f(\mathbf{x}), \quad (6)$$

it is easy to show that:

$$Q_\mu(\mathbf{z}, \mathbf{x}) = \frac{\mu}{2}\|\mathbf{z} - g(\mathbf{x})\|_2^2 + \lambda\|\mathbf{z}\|_0 + f(\mathbf{x}) - \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|_2^2,$$

and so, $\mathcal{P}_\mu(\cdot)$ is given by:

$$\mathcal{P}_\mu(\mathbf{x}) = \arg\min_{\mathbf{z}}\ \frac{1}{2}\|\mathbf{z} - g(\mathbf{x})\|_2^2 + (\lambda/\mu)\|\mathbf{z}\|_0. \quad (7)$$

For the proposed algorithm we fix $\mathcal{P}_\mu(\cdot) = \mathcal{H}_{\lambda/\mu}(g(\cdot))$, the point to point map defined in the following Theorem.

Theorem 2. Let the hard-thresholding (point-to-point) map $\mathcal{H}_h(\cdot)$, $h > 0$, be such that for each $i = 1, \ldots, m$:

$$\mathcal{H}_h(g(\mathbf{v}))[i] = \begin{cases} 0 & \text{if } |g(\mathbf{v})[i]| < \sqrt{2h} \\ g(\mathbf{v})[i]\ \mathbb{I}(\mathbf{v}[i] \neq 0) & \text{if } |g(\mathbf{v})[i]| = \sqrt{2h} \\ g(\mathbf{v})[i] & \text{if } |g(\mathbf{v})[i]| > \sqrt{2h}. \end{cases} \quad (8)$$

Then, $\mathcal{H}_{\lambda/\mu}(g(\cdot)) \in \arg\min_{\mathbf{z}} Q_\mu(\mathbf{z}, \cdot)$.

The proof is in the Appendix.

Clearly Theorem 1 holds with $\mathcal{P}_\mu(\cdot)$ replaced by $\mathcal{H}_{\lambda/\mu}(g(\cdot))$. The motivation for selecting this particular minimizer is Lemma 2 in Section 4.

3. THE ALGORITHM

The proposed MIST algorithm is constructed by repeated application of Theorem 1 where $\boldsymbol{\delta}$ is chosen to be the difference between the current and the previous iterate, i.e.,

$$\mathbf{x}_{k+1} = \mathcal{H}_{\lambda/\mu}\left(\mathbf{w}_k - \frac{1}{\mu}\nabla f(\mathbf{w}_k)\right),\ \mathbf{w}_k = \mathbf{x}_k + \alpha_k\boldsymbol{\delta}_k \quad (9)$$

with α_k given by (5), where $\boldsymbol{\delta}_k = \mathbf{x}_k - \mathbf{x}_{k-1}$. The iteration (9) is an instance of a momentum accelerated IST algorithm, similar to Fast IST Algorithm (FISTA) introduced in [8] for minimizing the convex l_1 PLS criterion. In (9), $\boldsymbol{\delta}_k$ is called the momentum term and α_k is a momentum step size parameter. A more explicit implementation of (9) is given below.

Momentumized IST (MIST) Algorithm

Compute $\bar{\mathbf{y}} = (\mathbf{y}^T\mathbf{A})^T$ off-line. Choose $\mathbf{x}_0$ and let $\mathbf{x}_{-1} = \mathbf{x}_0$. Calculate $\|\mathbf{A}\|^2$ off-line, let $\mu > \|\mathbf{A}\|^2$ and $k = 0$. Then:

(1) If $k = 0$, let $\alpha_k = 0$. Otherwise, compute:

(a) $\mathbf{u}_k = \mathbf{A}\mathbf{x}_k$

(b) $\mathbf{v}_k = (\mathbf{u}_k^T\mathbf{A})^T$

(c) $\mathbf{g}_k = \mathbf{x}_k - \frac{1}{\mu}(\mathbf{v}_k - \bar{\mathbf{y}})$

(d) $\mathbf{p}_k = \mathcal{H}_{\lambda/\mu}(\mathbf{g}_k) - \mathbf{x}_k$

(e) $\boldsymbol{\delta}_k = \mathbf{x}_k - \mathbf{x}_{k-1}$ and $\boldsymbol{\gamma}_k = \mu\boldsymbol{\delta}_k - \mathbf{v}_k + \mathbf{v}_{k-1}$

(f) Choose $\eta_k \in (0, 1)$ and compute:

$$\alpha_k = 2\eta_k\left(\frac{\boldsymbol{\gamma}_k^T\mathbf{p}_k}{\boldsymbol{\gamma}_k^T\boldsymbol{\delta}_k}\right) \quad (10)$$

(2) Using (c), (e) and (f) compute:

$$\mathbf{x}_{k+1} = \mathcal{H}_{\lambda/\mu}\left(\mathbf{g}_k + \frac{\alpha_k}{\mu}\boldsymbol{\gamma}_k\right) \quad (11)$$

(3) Let $k = k + 1$ and go to (1).

Remark 1. Thresholding using (8) is simple, and can always be done off-line. Secondly, note that MIST requires computing only $\mathcal{O}(2md)$ products, which is the same order required when the momentum term $\boldsymbol{\delta}_k$ is not incorporated, i.e., $\eta_k = 0$ for all k. In this case, MIST is a generalization of IHT from [1, 10]. Other momentum methods such as FISTA [8] and its monotone version M-FISTA [11] also require computing $\mathcal{O}(2md)$ and $\mathcal{O}(3md)$ products, respectively.

4. CONVERGENCE ANALYSIS

Here we prove that MIST converges to a local minimizer of $F(\cdot)$.

Theorem 3. Suppose $\{\mathbf{x}_k\}_{k\geq 0}$ is a bounded sequence generated by the MIST algorithm. Then $\mathbf{x}_k \to \mathbf{x}_\bullet$ as $k \to \infty$, where $\mathbf{x}_\bullet$ is a local minimizer of (1).

The proof requires several lemmas. In Lemma 1 and 2 it is assumed that MIST reaches a fixed point only in the limit, i.e., $\mathbf{x}_{k+1} \neq \mathbf{x}_k$ for all k. This implies that $\boldsymbol{\delta}_k \neq 0$ for all k.

Lemma 1. $\mathbf{x}_{k+1} - \mathbf{x}_k \to 0$ as $k \to \infty$.

The following lemma motivates Theorem 2 and is crucial for the subsequent convergence analysis.

Lemma 2. Assume the result in Lemma 1. If $\mathbf{x}_{k_n} \to \mathbf{x}_\bullet$ as $n \to \infty$, then:

$$\mathcal{H}_{\lambda/\mu}\left(\mathbf{w}_{k_n} - \frac{1}{\mu}\nabla f(\mathbf{w}_{k_n})\right) \to \mathcal{H}_{\lambda/\mu}\left(\mathbf{x}_\bullet - \frac{1}{\mu}\nabla f(\mathbf{x}_\bullet)\right), \quad (12)$$

where $\mathbf{w}_{k_n} = \mathbf{x}_{k_n} + \alpha_{k_n}\boldsymbol{\delta}_{k_n}$.

Lemma 3. Suppose $\mathbf{x}_\bullet$ is a fixed point of MIST. Letting $\mathcal{Z} = \{i : \mathbf{x}_\bullet[i] = 0\}$ and $\mathcal{Z}^c = \{i : \mathbf{x}_\bullet[i] \neq 0\}$,

(C_1) If $i \in \mathcal{Z}$, then $|\nabla f(\mathbf{x}_\bullet)[i]| \leq \sqrt{2\lambda\mu}$.

(C_2) If $i \in \mathcal{Z}^c$, then $\nabla f(\mathbf{x}_\bullet)[i] = 0$.

(C_3) If $i \in \mathcal{Z}^c$, then $|\mathbf{x}_\bullet[i]| \geq \sqrt{2\lambda/\mu}$.

Lemma 4. Suppose $\mathbf{x}_\bullet$ is a fixed point of MIST. Then there exists $\epsilon > 0$ such that $F(\mathbf{x}_\bullet) < F(\mathbf{x}_\bullet + \mathbf{d})$ for any $\mathbf{d}$ satisfying $\|\mathbf{d}\|_2 \in (0, \epsilon)$. In other words, $\mathbf{x}_\bullet$ is a strict local minimizer of (1).

Lemma 5. The limit points of $\{\mathbf{x}_k\}_{k\geq 0}$ are fixed points of MIST.

Due to lack of space only a sketch of the proofs is provided in the Appendix. Detailed proofs are given in [15].

5. SIMULATIONS

Here we demonstrate the performance advantages of the proposed MIST algorithm in terms of convergence speed. The methods used for comparison are the well known MM algorithms: ISTA and FISTA from [8], as well as M-FISTA from [11], where the soft-thresholding map is replaced by the hard-thresholding map. In this case, ISTA becomes identical to the IHT algorithm from [1, 10], while FISTA and M-FISTA become its accelerated versions, which exploit the ideas in [16].

A popular compressed sensing scenario is considered with the aim of reconstructing a length m sparse signal $\mathbf{x}$ from d observations ($d < m$). A relatively high dimensional example is considered: $d = 2^{13} = 8192$ and $m = 2^{14} = 16384$, and $\mathbf{x}$ contains 150 randomly placed ± 1 spikes (0.9% non-zeros). $\mathbf{A}_{d\times m}$ contains independent samples from the standard Gaussian distribution, and the standard deviation of the observation noise $\boldsymbol{\epsilon}$ is $\sigma = 3, 6, 10$.

The Signal to Noise Ratio (SNR) is defined by: SNR $= 10\log_{10}\left(\|\mathbf{A}\mathbf{x}\|_2^2/(\sigma^2 d)\right)$. For example plots of $\mathbf{A}\mathbf{x}$ and $\boldsymbol{\epsilon}$ when $\sigma = 3$ (SNR=12), $\sigma = 6$ (SNR=6) and $\sigma = 10$ (SNR=1.7) see Figures 1, 2 and 3 in [15] respectively.

5.1. Results

All algorithms are initialized with $\mathbf{x}_0 = \mathbf{0}$, and are terminated when $|F(\mathbf{x}_k) - F(\mathbf{x}_{k-1})|/F(\mathbf{x}_k) < 10^{-10}$. In the MIST algorithm we let $\mu = \|\mathbf{A}\|^2 + 10^{-15}$ and $\eta_k = 1 - 10^{-15}$. All experiments were run in MATLAB 8.1 on an Intel Core i7 processor with 3.0GHz CPU and 8GB of RAM.

As $\mathbf{x}$ is generally unknown to the experimenter we also report results of using a model selection method to select λ. Since classical methods tend to select a model with many spurious components when m is large and d is comparatively smaller [17], we use the Extended Bayesian Information Criterion (EBIC) model selection method proposed in [17]. The EBIC criterion is defined by $\text{EBIC}(\lambda) = \log\left(\frac{\|\mathbf{y}-\mathbf{A}\widehat{\mathbf{x}}\|_2^2}{d}\right) + \left(\frac{\log d}{d} + 2\gamma\frac{\log m}{d}\right)\|\widehat{\mathbf{x}}\|_0$, where $\widehat{\mathbf{x}} = \widehat{\mathbf{x}}(\lambda)$ is the estimator of $\mathbf{x}$ produced by a particular algorithm, and the chosen λ in (1) is the minimizer of the EBIC. As suggested in [17], $\gamma = 1 - 1/(2\kappa)$, where κ is the solution of $m = d^\kappa$, i.e., $\kappa \approx 1.08$.

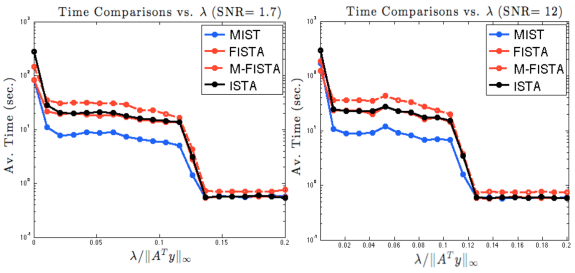

Fig. 1: Average time vs. λ over 5 instances when SNR=1.7 and 12. For the comparison, 20 equally spaced values of λ are considered, where $10^{-4}\|\mathbf{A}^T\mathbf{y}\|_\infty \leq \lambda \leq 0.2\|\mathbf{A}^T\mathbf{y}\|_\infty$. As it can be seen, except in the scenario when $\lambda = 10^{-4}\|\mathbf{A}^T\mathbf{y}\|_\infty$ the MIST algorithm outperforms the others. The smallest averaged $\arg\min_\lambda \text{EBIC}(\lambda)$ from the four algorithms is $\lambda = 0.03\|\mathbf{A}^T\mathbf{y}\|_\infty$ for SNR=1.7 and $\lambda = 0.017\|\mathbf{A}^T\mathbf{y}\|_\infty$ for SNR=12. The comparisons for SNR=6 are similar to the ones above, see Figure 11 in [15].

Based on a large number of experiments we noticed that MIST, FISTA and ISTA usually outperformed M-FISTA in terms of run time. This could be due to the fact that M-FISTA requires computing a larger number of products, see Remark 1, and the fact that it is a monotone version of a severely non-monotone FISTA. The high non-monotonicity could possibly be due to non-convexity of the objective function $F(\cdot)$.

6. CONCLUSION

We have developed a momentum accelerated MM algorithm, MIST, for minimizing the l_0 penalized least squares criterion for linear regression problems. We have provided sketch proofs of the convergence of MIST to a local minimizer without imposing any assumptions on the regression matrix $\mathbf{A}$. Simulations on large data sets have shown that the MIST algorithm outperforms other popular MM algorithms in terms of run time and number of iterations.

7. APPENDIX

Sketch Proof of Theorem 1: Let $\mathbf{w} = \mathbf{x} + \beta\boldsymbol{\delta}$. The quantities $f(\mathbf{w})$, $\nabla f(\mathbf{w})^T(\mathbf{z} - \mathbf{w})$ and $\|\mathbf{z} - \mathbf{w}\|_2^2$ are quadratic functions, and by simple linear algebra they can easily be expanded in terms of $\mathbf{z}$, $\mathbf{x}$ and $\boldsymbol{\delta}$. Using these expansions and the definition of $Q_\mu(\cdot,\cdot)$ it can also be shown that: $Q_\mu(\mathbf{z}, \mathbf{w}) = Q_\mu(\mathbf{z}, \mathbf{x}) + \Phi_\mu(\mathbf{z}, \boldsymbol{\delta}, \beta)$, where $\Phi_\mu(\mathbf{z}, \boldsymbol{\delta}, \beta) = \frac{1}{2}\beta^2\boldsymbol{\delta}^T\mathbf{B}_\mu\boldsymbol{\delta} - \beta\boldsymbol{\delta}^T\mathbf{B}_\mu(\mathbf{z} - \mathbf{x})$. Observing that $\boldsymbol{\delta}^T\mathbf{B}_\mu\boldsymbol{\delta} > 0$, let:

$$\beta = 2\eta\frac{\boldsymbol{\delta}^T\mathbf{B}_\mu(\mathbf{z} - \mathbf{x})}{\boldsymbol{\delta}^T\mathbf{B}_\mu\boldsymbol{\delta}}, \ \eta \in [0, 1]. \quad (13)$$

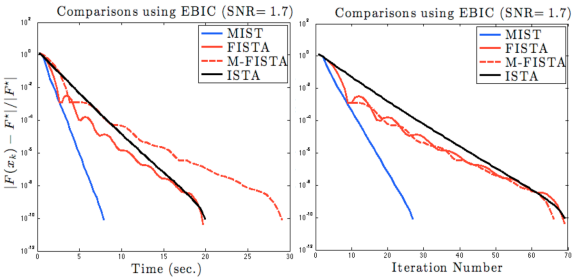

Fig. 2: Algorithm comparisons based on relative error $|F(\mathbf{x}_k) - F^\star|/|F^\star|$ where $F^\star$ is the final value of $F(\cdot)$ obtained by each algorithm at its termination, i.e., $F^\star = F(\mathbf{x}_k)$, where $F(\mathbf{x}_k)$ satisfies the termination criterion. All the algorithms use a common λ which is chosen to be the smallest λ from the averaged $\arg\min_\lambda$ EBIC(λ) obtained by each algorithm (over 10 instances). As it can be seen, in the high noise environment (SNR=1.7) the MIST algorithm outperforms the rest, both in terms of time and iteration.

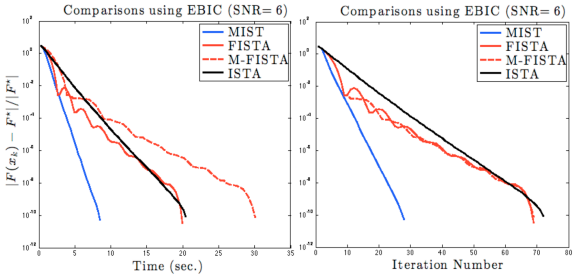

Fig. 3: Similar comparisons as in Fig. 2 except that SNR=6. As it can be seen, in the intermediate noise environment the MIST algorithm outperforms the others, both in terms of time and iteration number. The algorithm comparisons for SNR=12 are very similar to the ones here, see Figure 4 in [15].

Then, one has:

$$\begin{aligned} Q_\mu(\mathcal{P}_\mu(\mathbf{w}),\mathbf{w}) &= \min_{\mathbf{z}} Q_\mu(\mathbf{z},\mathbf{w}) \le Q_\mu(\mathbf{z},\mathbf{w}) \\ &= Q_\mu(\mathbf{z},\mathbf{x}) + \Phi_\mu(\mathbf{z},\boldsymbol{\delta},\beta) \\ &\overset{(13)}{=} Q_\mu(\mathbf{z},\mathbf{x}) - 2\eta(1-\eta)\frac{[\boldsymbol{\delta}^T\mathbf{B}_\mu(\mathbf{z}-\mathbf{x})]^2}{\boldsymbol{\delta}^T\mathbf{B}_\mu\boldsymbol{\delta}} \quad (14) \\ &\le Q_\mu(\mathbf{z},\mathbf{x}), \quad (15) \end{aligned}$$

which holds for any $\mathbf{z}$. So, letting $\mathbf{z} = \mathcal{P}_\mu(\mathbf{x})$ implies:

$$F(\mathcal{P}_\mu(\mathbf{w})) \overset{(4)}{\le} Q_\mu(\mathcal{P}_\mu(\mathbf{w}),\mathbf{w}) \overset{(15)}{\le} Q_\mu(\mathcal{P}_\mu(\mathbf{x}),\mathbf{x}) \overset{(4)}{\le} F(\mathbf{x}),$$

which completes the sketch proof. For the complete proof see [15, proof of Theorem 1]. □

Sketch Proof of Theorem 2: Looking at (7), $\mathcal{P}_\mu(\cdot)[i] = \arg\min_{\mathbf{z}[i]} \frac{1}{2}(\mathbf{z}[i]-g(\cdot)[i])^2 + (\lambda/\mu)\mathbb{I}(\mathbf{z}[i]\neq 0)$. The result then easily follows by considering [18, Theorem 1], see [15, proof of Theorem 2]. □

Sketch Proof of Lemma 1: From Theorem 1, $0 \le F(\mathbf{x}_{k+1}) \le F(\mathbf{x}_k)$, so the sequence $\{F(\mathbf{x}_k)\}_k$ is bounded meaning it has a finite limit, say, $F_\bullet$. As a result:

$$F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}) \to F_\bullet - F_\bullet = 0 \quad (16)$$

Next, recall that $\mathbf{w}_k = \mathbf{x}_k + \alpha_k\boldsymbol{\delta}_k$ and $g(\cdot) = (\cdot) - \frac{1}{\mu}\nabla f(\cdot)$. So, using (14) in the proof of Theorem 1, the MM inequalities in (4) and the definition of α_k in (10), it can easily be shown that:

$$Q_\mu(\mathbf{x}_{k+1},\mathbf{w}_k) \le F(\mathbf{x}_k) - \sigma_k\alpha_k^2\boldsymbol{\delta}_k^T\mathbf{B}_\mu\boldsymbol{\delta}_k, \quad (17)$$

where $\sigma_k = (1-\eta_k)/\eta_k > 0$. Using the fact that: $Q_\mu(\mathbf{x}_{k+1},\mathbf{w}_k) = F(\mathbf{x}_{k+1}) + \frac{1}{2}(\mathbf{x}_{k+1}-\mathbf{w}_k)^T\mathbf{B}_\mu(\mathbf{x}_{k+1}-\mathbf{w}_k)$, which easily follows from basic linear algebra, it can be shown that (17) implies:

$$F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}) \ge \rho\sigma_k\alpha_k^2\|\boldsymbol{\delta}_k\|_2^2 + \frac{\rho}{2}\|\mathbf{x}_{k+1}-\mathbf{w}_k\|_2^2, \quad (18)$$

where $\rho > 0$ is the smallest eigenvalue of $\mathbf{B}_\mu \succ 0$. So, both terms on the right hand side in (18) are ≥ 0 for all k. As a result, due to (16) we can use the pinching argument on (18) to establish that $\mathbf{x}_{k+1} - \mathbf{w}_k = \boldsymbol{\delta}_{k+1} - \alpha_k\boldsymbol{\delta}_k \to 0$ and $\alpha_k\boldsymbol{\delta}_k \to 0$ as $k \to \infty$. Consequently, $\boldsymbol{\delta}_k \to 0$ as $k \to \infty$, which completes the sketch proof. For the complete proof see [15, proof of Lemma 1]. □

Proof of Lemma 2: See [15, proof of Lemma 2].

Sketch Proof of Lemma 3: The fixed points are obtained by setting $\mathbf{x}_{k+1} = \mathbf{x}_k = \mathbf{x}_{k-1} = \mathbf{x}_\bullet$. So, any fixed point $\mathbf{x}_\bullet$ satisfies the fixed point equation: $\mathbf{x}_\bullet = \mathcal{H}_{\lambda/\mu}\left(\mathbf{x}_\bullet - \frac{1}{\mu}\nabla f(\mathbf{x}_\bullet)\right)$. The result is then easily established by substituting the definition of $\mathcal{H}_{\lambda/\mu}(g(\cdot))$ from (8) into the stated fixed point equation and solving the resulting equation. For the complete proof see [15, proof of Lemma 3]. □

Sketch Proof of Lemma 4: Letting $\mathcal{Z} = \{i : \mathbf{x}_\bullet[i] = 0\}$ and $\mathcal{Z}^c = \{i : \mathbf{x}_\bullet[i] \neq 0\}$, it can easily be shown that $F(\mathbf{x}_\bullet + \mathbf{d}) = F(\mathbf{x}_\bullet) + \phi(\mathbf{d})$, where:

$$\begin{aligned} \phi(\mathbf{d}) &= \frac{1}{2}\|\mathbf{A}\mathbf{d}\|_2^2 + \mathbf{d}^T\nabla f(\mathbf{x}_\bullet) + \lambda\|\mathbf{x}_\bullet + \mathbf{d}\|_0 - \lambda\|\mathbf{x}_\bullet\|_0 \\ &\ge \sum_{i\in\mathcal{Z}} \underbrace{\mathbf{d}[i]\nabla f(\mathbf{x}_\bullet)[i] + \lambda\mathbb{I}(\mathbf{d}[i]\neq 0)}_{=\phi_{\mathcal{Z}}(\mathbf{d}[i])} \\ &\quad + \sum_{i\in\mathcal{Z}^c} \underbrace{\mathbf{d}[i]\nabla f(\mathbf{x}_\bullet)[i] + \lambda\mathbb{I}(\mathbf{x}_\bullet[i]+\mathbf{d}[i]\neq 0) - \lambda}_{=\phi_{\mathcal{Z}^c}(\mathbf{d}[i])} \end{aligned}$$

Since $\phi_{\mathcal{Z}}(0) = \phi_{\mathcal{Z}^c}(0) = 0$, all that needs to be shown is that there exists $\delta > 0$ such that $\phi_{\mathcal{Z}}(\mathbf{d}[i]) > 0$ and $\phi_{\mathcal{Z}^c}(\mathbf{d}[i]) > 0$ for any $|\mathbf{d}[i]| \in (0,\delta)$. Letting $\delta = \lambda/\sqrt{2\lambda\mu}$, it can easily be shown that $\phi_{\mathcal{Z}}(\mathbf{d}[i]) > 0$ by considering C_1 in Lemma 3. Using C_2 and C_3 in Lemma 3 we have $\phi_{\mathcal{Z}^c}(\mathbf{d}[i]) = 0$ since $\nabla f(\mathbf{x}_\bullet[i]) = 0$, $|\mathbf{x}_\bullet[i]| \ge \sqrt{2\lambda/\mu}$ and $|\mathbf{d}[i]| < \lambda/\sqrt{2\lambda\mu} < \sqrt{2\lambda/\mu}$, completing the sketch proof. For a detailed proof see [15, proof of Lemma 4]. □

Proof of Lemma 5: See [15, proof of Lemma 5].

Proof of Theorem 3: By Lemma 1 and Ostrowski's result [19, Theorem 26.1], the bounded $\{\mathbf{x}_k\}_{k\ge 0}$ converges to a closed and connected set, i.e., the set of limit points form a closed and connected set. But, by Lemma 5 these limit points are fixed points, which by Lemma 4 are strict local minimizers. So, since the local minimizers form a discrete set the connected set of limit points can only contain one point, and so, the entire $\{\mathbf{x}_k\}_{k\ge 0}$ must converge to a single local minimizer. □

8. REFERENCES

[1] T. Blumensath, M. Yaghoobi, and M. E. Davies, "Iterative hard thresholding and l_0 regularisation," *IEEE ICASSP*, vol. 3, pp. 877–880, 2007.

[2] R. Mazumder, J. Friedman, and T. Hastie, "SparseNet: Coordinate descent with non-convex penalties," *J. Am. Stat. Assoc.*, vol. 106, no. 495, pp. 1–38, 2011.

[3] K. Bredies, D. A. Lorenz, and S. Reiterer, "Minimization of non-smooth, non-convex functionals by iterative thresholding," *Tech. Rept.*, 2009, http://www.uni-graz.at/~bredies/publications.html.

[4] P. Tseng, "Convergence of block coordinate descent method for nondifferentiable minimization," *J. Optimiz. Theory App.*, vol. 109, no. 3, pp. 474–494, 2001.

[5] M. Nikolova, "Description of the minimisers of least squares regularized with l_0 norm. uniqueness of the global minimizer," *SIAM J. Imaging. Sci.*, vol. 6, no. 2, pp. 904–937, 2013.

[6] G. Marjanovic and V. Solo, "On exact l_q denoising," *IEEE ICASSP*, pp. 6068–6072, 2013.

[7] ——, "l_q sparsity penalized linear regression with cyclic descent," *IEEE T. Signal Proces.*, vol. 62, no. 6, pp. 1464–1475, 2014.

[8] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[9] S. J. Wright, R. D. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE T. Signal Proces.*, vol. 57, no. 7, pp. 2479–2493, 2009.

[10] T. Blumensath and M. Davies, "Iterative thresholding for sparse approximations," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 629–654, 2008.

[11] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring," *IEEE T. Image Process.*, vol. 18, no. 11, pp. 2419–2134, 2009.

[12] M. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE T. Image Process.*, vol. 16, no. 12, pp. 2980–2991, 2007.

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

[14] A. Chambolle and C. Dossal, "On the convergence of the iterates of "FISTA"," 2014, https://tel.archives-ouvertes.fr/X-CMAP/hal-01060130v3.

[15] G. Marjanovic, M. O. Ulfarsson, and A. O. Hero III, "MIST: l_0 Sparse Linear Regression with Momentum," arXiv: http://arxiv.org/abs/1409.7193.

[16] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$," *Soviet Math. Doklady*, vol. 27, pp. 372–376, 1983.

[17] J. Chen and Z. Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.

[18] G. Marjanovic and A. O. Hero, "On l_q estimation of sparse inverse covariance," *IEEE ICASSP*, 2014.

[19] A. M. Ostrowski, *Solutions of Equations in Euclidean and Banach Spaces*. New York: Academic Press, 1973.