Least Favorable Compressed Sensing Problems for First Order Methods

Arian Maleki
Department of Electrical and Computer Engineering
Rice University

Richard Baraniuk
Department of Electrical and Computer Engineering
Rice University

Abstract—Compressed sensing (CS) aims to exploit the compressibility of natural signals to reduce the number of samples needed for the accurate reconstruction. Unlike the traditional Shannon-Nyquist sampling, the recovery algorithms in CS, such as the popular \( \ell_1 \)-minimization, are computationally expensive. Therefore, first order methods, such as iterative soft thresholding (IST) and fast iterative soft thresholding algorithm (FISTA) have been explored extensively as cheap methods for solving the \( \ell_1 \)-minimization. However, the theoretical aspects of these algorithms have been mainly studied in the standard framework of convex optimization, called the deterministic framework here. In this paper, we first show that the deterministic approach results in very pessimistic conclusions that are not indicative of the performance of these algorithms in CS. Several papers have considered the statistical convergence rate as an alternative to the deterministic view. However, the theoretical aspects of this convergence rate have remained unexplored in the sparse recovery problem. We study this convergence rate for several standard first order methods both theoretically and empirically. We derive several hallmark properties of the statistical convergence, including universality over the matrix ensemble and the least favorable coefficient distribution. These results may help the researchers for choosing more informative problem instances in their simulations.

I. FIRST ORDER METHODS FOR LASSO

A. The LASSO problem

Consider a simple unconstrained convex optimization problem

\[
\min_{x \in \mathbb{R}^N} f(x),
\]

where \( f \) is a differentiable function. Iterative approaches are used to approximate the optimal point when the exact solution cannot be calculated explicitly. Suppose that the algorithm is iterative and our estimate of the optimal point \( x^* \) at iteration \( t \) is called \( x^t \). First-order methods use the gradient of the function at \( x^0, x^1, \ldots, x^t \) to obtain a new estimate at time \( t + 1 \). These methods are very popular for very large scale problems, due to their inexpensive iterations.

In this paper we are interested in solving the following problem:

\[
(P_\lambda) \quad \min_{x} \frac{1}{2} ||Ax - y||_2^2 + \lambda ||x||_1,
\]

where the matrix \( A \) is drawn from a random ensemble, e.g., iid Gaussian or Bernoulli. With a slight abuse of notation an algorithm is called a first order method for \( P_\lambda \) if it only uses the gradient of the function \( \frac{1}{2} ||Ax - y||_2^2 \) at the previous iterations. Using first order methods for solving the \( \ell_1 \)-minimization problem has been an active area of research in the last decade, with new proposals appearing regularly [11–17]. In this paper we focus on three standard first order algorithms and our goal is to characterize the difficult CS problems for these three algorithms.

B. Iterative soft thresholding (IST)

Without loss of generality we assume that all the algorithms start from \( x^0 = 0 \). The IST iteration for solving \( P_\lambda \) is given by

\[
x^{t+1} = \eta(x^t + \alpha^t A^*(y - Ax^t); \alpha^t \lambda),
\]

where \( \eta \) is the soft thresholding function\(^1\) applied component-wise to the elements of the vector. \( \alpha^t \) is called the step size and ensures the stability of the algorithm. For more information on the history of this algorithm see [9].

C. Fast iterative soft thresholding algorithm (FISTA)

Another first order method that has recently drawn attention is FISTA [10], [15]. This algorithm uses the following iteration:

\[
x^{t+1} = \eta(x^t + \alpha^t A^*(y - Ax^t); \lambda \alpha^t),
\]

\[
z^t = x^t + \frac{t - 1}{t + 2} (x^t - x^{t-1}).
\]

Note that unlike IST, in addition to \( x^t \), \( x^{t-1} \) is used in obtaining \( x^{t+1} \).

D. Approximate Message Passing (AMP)

Now consider the linear measurement model \( y = Ax_o + w \), where \( w \) is the noise of the system and \( x_o \) is the compressible vector. The iterations of AMP for this problem are given by

\[
x^{t+1} = \eta(x^t + A^T z^t; \tau \hat{\sigma}^t),
\]

\[
z^t = y - Ax^t + \frac{|I|^t}{n} z^{t-1},
\]

where \( \hat{\sigma}^t \) is an estimate of the standard deviation of the the vector \( x^t + A^T z^t - x_o \) and \( I^t \) is the active set of the vector \( x^t \) at time \( t \) [13]. For more information on parameter \( \tau \) and its connection to \( \lambda \) refer to Section IV-A. Unlike the other two methods AMP is derived from the statistical framework and therefore is more adapted to CS problems.

\(^1\)The soft thresholding function is defined as \( \eta(x; \lambda) = (x - \lambda)_+ \text{sign}(x) \).
II. DRAWBACK OF THE DETERMINISTIC FRAMEWORK FOR CS

To compare the performance of different first order methods for solving \( P_\lambda \), researchers have considered several different approaches. The simplest and most popular approach is the deterministic framework. In this framework, we consider the following class of functions,

\[ P_{L}^{n,N} = \{ \frac{1}{2} \| y - Ax \|_2^2 + \lambda \| x \|_1 : A \in \mathbb{R}^{n \times N}, \| A^* A \|_{2,2} \leq L \}. \]

We use the notation \( x_{a,f}^t \) for the estimate of algorithm \( a \) on the function \( f \in P_{L}^{n,N} \) at time \( t \). Further, assume that \( X_t^* \) is the set of all the points that achieve the minimum of \( f \). Define \( d(x_{a,f}^t, X_t^*) = \inf_{x \in X_t^*} \| x_{a,f}^t - x \|_2^2 \). In the deterministic framework we are interested in the performance of an algorithm on the least favorable function in the class, i.e.,

\[
MSE^t(P_{L}^{n,N}, a) = \max_{f \in F} d(x_{a,f}^t, X_t^*),
PE^t(P_{L}^{n,N}, a) = \max_{f \in F} |f(x_{a,f}^t) - f(x_t^*)|,
\]

where \( X_t^* \) is the set of minimizers of \( f \) and \( x_t^* \in X_t^* \). In this analysis, we are mainly interested in the decay rate \( MSE^t(P_{L}^{n,N}, a) \) and \( PE^t(P_{L}^{n,N}, a) \) as \( t \) grows. To see the main drawback of the deterministic framework we consider the class \( A \) of all the first order methods and define the following two minimax errors:

\[
MSE^t(P_{L}^{n,N}, A) = \min_{a \in A} MSE^t(P_{L}^{n,N}, a),
PE^t(P_{L}^{n,N}, A) = \min_{a \in A} PE^t(P_{L}^{n,N}, a).
\]

**Theorem II.1.** Consider the class of functions \( P_{L}^{n,N} \). There exists a constant \( C \) such that for any given \( L, n < N \), and for any \( t \leq \frac{1}{n \lambda^2} \),

\[
PE^t(P_{L}^{n,N}, A) \leq C L \| x^* \|_2 \left( \frac{n}{n + 1} \right)^2.
MSE^t(P_{L}^{n,N}, A) \geq \frac{1}{25} \| x^* \|_2^2.
\]

The proof of the above theorem is similar to the proof of Theorem 2.1.7 [18] and therefore for the sake of brevity is skipped here. See [19] for details. Since in CS we are mainly interested in very high dimensional problems and we are often interested in the convergence of \( x^t \) to \( x^* \), the deterministic framework is not indicative of what we observe in practice; \( x^t \) of several first order methods, such as AMP, converges to \( x^* \) extremely fast.

Due to this drawback, average case analysis and simulations have played an important role in comparisons. In this approach we exploit the randomness of \( A \) and instead of calculating \( \| x^t - x^* \|_2^2 \) on the worst possible case we work with average case quantity \( E(\| x^t - x^* \|_2^2) \). Majority of the papers published on the LASSO solvers use Monte Carlo simulations to estimate and compare this quantity. However, despite the extensive number of simulations in the literature, our understanding of the average convergence rate is very limited. This has led to ad-hoc problem selection methods and ad-hoc comparisons.

Our goal in this paper is to study the properties of the statistical convergence rate on three standard first order methods FISTA, IST and AMP mentioned in Section I. We show which matrix ensembles and which coefficient ensembles are more difficult for these algorithms. This provides a guideline for choosing difficult problem instances for each algorithm. We will also see the interesting properties of the constant amplitude coefficient ensemble that makes it a difficult coefficient ensemble for these algorithms.

III. MAIN CONTRIBUTIONS

In this paper we consider the sparse recovery problem in the presence of noise, i.e., let \( y = Ax_o + z \), where \( z \) is iid Gaussian noise with variance \( \nu \) and \( x_o \) is the signal to be recovered. We denote a problem instance by \( \Theta = (D_A, G, \epsilon; \delta, \nu) \). \( D_A \) represents the distribution from which the matrix is drawn. \( G \) is the probability density function of non-zero elements of \( x^* \), \( \epsilon \) is the probability that an element of \( x_o \) is non-zero. In other words, \( x_o \sim (1 - \epsilon) \delta_0(x_o) + \epsilon G(x_o) \), where \( x_o \) is the \( \tau \)th element of the vector \( x_o \). Finally \( \delta = n/N \) where \( n \) is the number of measurements and \( N \) is the dimension of the vector \( x_o \). Suppose that \( x^t \) is a sequence resulting from one of the algorithms. We define the mean-time-to-converge \( t(\alpha) = \inf \{ t_0 : \lim_{N \to \infty} E[\| x^t - x_o \|_2^2 / E[\| x_o \|_2^2] \leq \alpha \text{ for } t > t_0 \} \}. t(\alpha) \) depends on both the problem instance and the algorithm. It is also worth mentioning that if \( x^t \) does not converge to \( x_o \) in the mean square sense, then \( t(\alpha) \) will be infinite for \( \alpha < \lim_{N \to \infty} \lim_{\tau \to \infty} E[\| x^t - x_o \|_2^2 / E[\| x_o \|_2^2] \] . Also according to [13] \( \lim_{N \to \infty} E[\| x^t - x_o \|_2^2] \) for these problem instances can converge to its final value exponentially fast.

**Definition III.1.** The problem instance \( \Theta \) is called less favorable than the problem instance \( \Theta' \) for IST, FISTA, and AMP algorithm if and only if for every \( \lambda_0 \) (or \( \tau_0 \)) there exists \( \lambda_0' \) (or \( \tau_0' \)) such that for every \( \alpha > 0 \), \( t_{\lambda_0}(\alpha) > \frac{1}{\lambda_0'}(\alpha) \).

**Definition III.2.** Two problem instances \( \Theta \) and \( \Theta' \) are called equivalent if and only if \( \Theta \) is less favorable than \( \Theta' \) and vice versa.

**Matrix universality hypothesis:** Suppose that the elements of \( n \times N \) measurement matrix are chosen iid at random from a “well-behaved” probability distribution. Furthermore, the non-zero elements of the vector \( x_o \) are sampled randomly from a given distribution \( G \). The observed behavior of \( E[\| x^t - x_o \|_2^2 / N] \) for the FISTA algorithm (or IST or AMP) will exhibit the same behavior as the Gaussian ensemble with large \( N \). In other words, under the above assumptions the

\[ E(A_{ij}) = 0 \quad \text{and} \quad E(A_{ij}^2) = \frac{1}{n}. \]
The large p-values confirm that we cannot reject the null hypothesis.

**Empirical Finding 2:** Under the above assumptions $G = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$ is less favorable problem instance for FISTA and IST than the other ensembles defined in Table II.

We believe that the above finding holds on a very wide range of distributions, although we have only checked it for a limited number of distributions.

**TABLE II**

COEFFICIENT ENSEMBLES CONSIDERED IN COEFFICIENT ENSEMBLE ANALYSIS EXPERIMENTS.

<table>
<thead>
<tr>
<th>Name</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>3P</td>
<td>iid elements taking value 0, 1, -1 with $P(0) = 1 - \epsilon$</td>
</tr>
<tr>
<td>5P1</td>
<td>iid elements taking values $0, \pm 1, \pm 5$ with $P(1) = P(-1) = .3\epsilon$ and $P(5) = P(-5) = .2\epsilon$</td>
</tr>
<tr>
<td>5P2</td>
<td>iid elements taking values $0, \pm 1, \pm 20$ with $P(1) = P(-1) = .3\epsilon$ and $P(20) = P(-20) = .2\epsilon$</td>
</tr>
<tr>
<td>U1</td>
<td>iid $U[0, 4]$</td>
</tr>
<tr>
<td>G1</td>
<td>iid $N(0, 4)$</td>
</tr>
</tbody>
</table>

The final important factor that we want to consider is the sparsity level $c$. Fix all the other parameters and vary $\epsilon$. Call the resulting set of problem instances $\mathcal{I}(D_A, G; \epsilon, \nu)$.

**Theorem III.6.** If $\epsilon < \epsilon'$, $\epsilon'$ results in a less favorable problem instance for AMP algorithm.

The proof of this theorem is given in Section IV-C. The above theorem formalizes the intuition that as we relax the sparsity, the problems become more difficult.

**Theorem III.7.** For a fixed value of $\delta$, if $\epsilon < \epsilon'$, $\epsilon$ can not be less favorable for FISTA and IST than $\epsilon'$.

Clearly, the above theorem does not prove that $\epsilon$ is less favorable than $\epsilon'$ for IST or FISTA. However, empirical observations confirm this.
Empirical Finding 3. For a fixed value of $\delta$ if $\epsilon < \epsilon'$, $F_{\epsilon,1}$ is more favorable for FISTA and IST than $F_{\epsilon',1}$.

For the sake of brevity we remove the experimental results that led to the above empirical findings from the paper and refer the reader to [19].

IV. PROOFS OF THE MAIN RESULTS

A. AMP-LASSO calibration

Here we briefly summarize some of the recent advances in predicting the asymptotic performance of the LASSO algorithm that helps us in our analysis. In [13] the authors proposed the state evolution framework as a theoretical framework to predict the performance of AMP algorithm. According to this framework, if $m_t$ is the mean square error at iteration $t$, the mean square error at iteration $t+1$ is $m_{t+1} = \Psi(m_t)$ where

$$ \Psi(m_t) = \mathbb{E}(\eta(X + \sqrt{\frac{m_t}{\delta}} + \nu Z; \tau \sqrt{\frac{m_t}{\delta}} + \nu) - X)^2. $$

$X \sim (1 - \epsilon)\delta_0(\mu) + \epsilon G(\mu)$ and $Z \sim N(0, 1)$ are two independent random variables. Clearly SE predicts the performance of the algorithm in the asymptotic setting $N \to \infty$. The final mean square error of the AMP algorithm also corresponds to the stable fixed points of the $\Psi$ function. It is proved that the $\Psi$ function is concave and therefore it has just one stable fixed point.

In addition to MSE, other observables of AMP algorithm can be calculated through the state evolution framework such as, equilibrium threshold ($\theta_\infty = \tau \sqrt{\frac{\delta}{\delta} + \nu}$) and equilibrium detection rate (EqDR = $\mathbb{P}\{\eta(Y_\infty; \theta_\infty) \neq 0\}$, where $Y_\infty = X + \sqrt{\frac{\delta}{\delta} + \nu} Z$). $m_\infty$ represents the fixed point of $\Psi$ function. The final component that will be used in our arguments is the equivalence between LASSO and AMP solutions. The following finding taken from [21], which has been recently proved in case of Gaussian measurement matrices in [22], explains the equivalence. For more information on the details of the conditions necessary for these theorems refer to [21] and [22].

Theorem IV.1. [21], [22] For each $\lambda \in [0, \infty)$ we find that AMP($\tau(\lambda)$) and LASSO($\lambda$) have statistically equivalent observables. In particular the MSE have the same value when $\tau$ and $\lambda$ satisfy the following relation:

$$ \lambda = \theta_\infty(\tau(1 - \text{EqDR}(\tau)/\delta). $$

Lemma IV.2. [21] Define $\tau_0$, so that EqDR($\tau$) $\leq \delta$ when $\tau > \tau_0$. For each $\lambda$ there is a unique value of $\tau(\lambda) \in [\tau_0, \infty)$ such that

$$ \lambda = \theta_\infty(\tau(1 - \text{EqDR}(\tau)/\delta). $$

The above discussion is mainly used in the calibration of $\lambda$ for two different distributions.

B. Coefficient distribution $G$

Proof of Theorem III.4: Suppose that an arbitrary distribution $G$ is given and let $E_G(X^2) = 1$. Define $G_{\epsilon}(\mu) \equiv \epsilon G(\mu) + (1 - \epsilon)\delta_0(\mu)$ and $F_{\epsilon,1} \equiv \frac{1}{2}\delta_1(\mu) + \frac{1}{2}\delta_{-1}(\mu) + (1 - \epsilon)\delta_0(\mu)$. Our goal is to prove that $F_{\epsilon,1}$ is less favorable than $G_{\epsilon}$. For a given $\tau$ in AMP for $F_{\epsilon,1}$, we choose $\tau_G = \tau$. Let $m_{G_{\epsilon}}^t$ be the mean square error on $G_{\epsilon}$ problem instance at time $t$ and $m_{F_{\epsilon,1}}^t$ represent the same thing for $F_{\epsilon,1}$. Suppose that $MSE_{G_{\epsilon}}^t \leq MSE_{F_{\epsilon,1}}^t$ at time $t$, our goal is to first prove that $m_{G_{\epsilon}}^{t+1} \leq m_{F_{\epsilon,1}}^{t+1}$. We have

$$ m_{G_{\epsilon}}^{t+1} = E_G \mathbb{E}_X (\eta(X + \sqrt{\frac{m_G^t}{\delta}} + \nu Z; \tau \sqrt{\frac{m_G^t}{\delta}} + \nu) - X)^2. $$

Also let us define,

$$ m_{F_{\epsilon,1}}^{t+1} = E_{F_{\epsilon,1}} \mathbb{E}_X (\eta(X + \sqrt{\frac{m_{F_{\epsilon,1}}^t}{\delta}} + \nu Z; \tau \sqrt{\frac{m_{F_{\epsilon,1}}^t}{\delta}} + \nu) - X)^2. $$

Our claim is that,

$$ m_{G_{\epsilon}}^{t+1} \leq m_{F_{\epsilon,1}}^{t+1} \leq m_{F_{\epsilon,1}}^{t+1}. $$

The second inequality is a simple result of the fact that the mean square error is non-decreasing function of standard deviation of the noise. The first inequality is the result of Jensen inequality, since $\mathbb{E}_X (\eta(X + \sqrt{\frac{MSE_{F_{\epsilon,1}}^t}{\delta}} + \nu Z; \tau \sqrt{\frac{MSE_{F_{\epsilon,1}}^t}{\delta}} + \nu) - X)$ is a concave function of $X^2$. The proof of the main theorem is now a very simple induction. Since the mean square error for the two distributions are the same at iteration one.

The following lemma plays an important role in our discussion of Theorem III.5.

Lemma IV.3. Consider the distribution $G$ with $E_G(X^2) = 1$. For every $\lambda \geq 0$ and for every $\nu > 0$ if

$$ \lim_{t \to \infty} \lim_{N \to \infty} \frac{1}{N} E_{F_{\epsilon,1}} \mathbb{E}_{x_o} \|\hat{x}_\lambda - x_o\|^2 \leq \epsilon, $$

then there exists a corresponding $\lambda_G$ such that

$$ \lim_{t \to \infty} \lim_{N \to \infty} \frac{1}{N} E_{G} \mathbb{E}_{x_o} \|\hat{x}_{\lambda_G} - x_o\|^2 \leq \epsilon. $$

Proof: According to Lemma IV.2, there exist a value of $\tau$ for which the asymptotic mean square error of the state evolution frameworks is the same as the asymptotic mean square error of LASSO($\lambda$). Here is our approach for solving this lemma. The first thing that we want to prove is that there exists a value of $\tau$ for which the asymptotic mean square error of AMP($\tau$) on $G$ is below the MSE of AMP($\tau$) on $F_{\epsilon,1}$. Then we prove there exists a value of $\tau$ for which the asymptotic mean square error of AMP($\tau$) is larger than the mean square error of AMP($\tau$) on $F_{\epsilon,1}$. Finally we will use implicit function theorem to prove that there exists a value of $\tau_G$ for which the asymptotic mean square error of AMP($\tau$) on $G$ is exactly the same as the MSE of AMP($\tau$) on $F_{\epsilon,1}$. Choice $\tau = \infty$ is clear. Therefore we focus on constructing a choice for $\tau$.

The claim is that $\tau = \tau_G$. This is due to the concavity of $E_X (\eta(X + \sqrt{m/\delta} Z; \beta) - X)^2$ in terms of $X^2$ and the Jensen inequality. Therefore for the sake of brevity we skip to write the complete argument here.

The interesting fact about the above proof is that it is also constructive. For a given value of $\lambda$ on the problem instance $(D_{\lambda}, \frac{1}{2}\delta_1(\mu) + \frac{1}{2}\delta_{-1}(\mu), \epsilon; \delta, \nu)$ we can calculate $\lambda_G$ that gives
us the same mean square error. This is the value that has been used for deriving the empirical findings.

Note: There might be more than one value of \( \lambda \) that generates the same mean square error on \( F_{e,1} \) and there may be more than one value of \( \lambda_C \) with the same mean square error. In these cases we compare the fastest achievable rates for each case.

Proof of Theorem III.5: The proof of this theorem is also clear from our discussion of the last two theorems. To prove this theorem we focus on the final mean square error of the algorithms and using the equivalence framework we can prove that the mean square error of the algorithm is highest when the distribution is \( \frac{1}{2} \delta_1(\mu) + \frac{1}{2} \delta_{-1}(\mu) \). For more details on the proof, refer to [19].

C. Sparsity level

To simplify the notation in this section we mainly focus on \( G = \frac{1}{2} \delta_1(\mu) + \frac{1}{2} \delta_{-1}(\mu) \). However the results can be easily extended to other distributions as well.

Lemma IV.4. For the risk of the soft thresholding function defined as \( r_s(\mu, \tau; \sigma) = E(\eta(\mu + \sigma Z; \tau \sigma) - \mu)^2 \), we have

\[
\frac{d}{d\mu} r_s(\mu, \tau; \sigma) = 2 \mu \int_{-\tau - \mu}^{\tau + \mu} \phi_Z(z) dz.
\]

Lemma IV.5. Consider the following risk function,

\[
R(\epsilon) = E_{X \sim F_{e,1}, \sqrt{\epsilon}} E_X(\eta(X + \sigma Z; \tau \sigma) - X)^2;
\]

\( R(\epsilon) \) is an increasing function of \( \epsilon \).

See [19] for the proof of the above two lemmas.

Proof of Theorem III.6

Proof: According to Lemma III.3, \( F_{e,1} \) is equivalent to \( F_{e,1}/\sqrt{\epsilon} \) and \( F_{e,1} \) is equivalent to \( F_{e,1}/\sqrt{\tau} \). Therefore we compare these two mathematical uses. We use the mathematical induction to prove the above theorem. Let \( m_t^\epsilon \) be the mean squared error of AMP(\( \tau \)) at iteration \( t \) on \( \epsilon \). Suppose that \( m_t^\epsilon > m_{t-1}^\epsilon \) and the goal is to prove that \( m_{t+1}^\epsilon > m_{t+1}^{\epsilon+1} \). Define

\[
m_{t+1}^\epsilon = E_{F_{e,1}/\sqrt{\epsilon}} E_X(\eta(X + \frac{m_t^\epsilon}{\delta} + \nu Z; \tau \sqrt{\frac{m_t^\epsilon}{\delta} + \nu}) - X)^2.
\]

We claim \( m_{t+1}^\epsilon > m_{t+1}^{\epsilon+1} > m_{t+1}^\epsilon \). The first inequality is due to the fact that the mean square error is a non-decreasing function of the standard deviation of the noise. The second inequality however is the result of Lemma IV.5. The base of the induction is correct since both algorithms start with the same mean square error and therefore the proof is complete.

\section{V. Conclusion}

We studied the statistical convergence rate of three standard first order methods. By empirical study of these rates we showed that a class of measurement matrices are equivalent. Also, we showed that the constant amplitude distribution is the least favorable distribution for FISTA and IST. Our theoretical results also confirmed that the least favorable distribution exists for the AMP algorithm and is equal to the constant amplitude. It also showed that less sparse signals under the same non-zero distribution are less favorable.

\section{Acknowledgment}

The authors would like to thank David L. Donoho, Andrea Montanari, and Rahul Mazumder for helpful discussions.

\section{References}


