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 ℓ_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 ℓ_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

$$\min_{x \in \mathbb{R}^N} f(x), \tag{1}$$

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:

$$(\mathcal{P}_\lambda) \qquad \min_x \frac{1}{2}\|Ax - y\|_2^2 + \lambda \|x\|_1,$$

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 $\mathcal{P}_\lambda$ 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 ℓ_1-minimization problem has been an active area of research in the last decade, with new proposals appearing regularly [1]–[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 $\mathcal{P}_\lambda$ is given by

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

where η is the soft thresholding function[1] applied componentwise to the elements of the vector. α^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:

$$\begin{aligned} x^{t+1} &= \eta(z^t + \alpha^t A^*(y - Az^t); \lambda \alpha^t), \\ z^t &= x^t + \frac{t-1}{t+2}(x^t - x^{t-1}). \end{aligned}$$

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 comprressible vector. The iterations of AMP for this problem are given by

$$\begin{aligned} x^{t+1} &= \eta(x^t + A^* z^t; \tau \hat{\sigma^t}), \\ z^t &= y - Ax^t + \frac{|I^t|}{n} z^{t-1}, \end{aligned}$$

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 τ and its connection to λ 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 $\mathcal{P}_\lambda$, 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,

$$\mathcal{P}_L^{n,N} = \{\frac{1}{2}\|y-Ax\|_2^2+\lambda\|x\|_1 \ : \ A \in \mathbb{R}^{n\times N}, \ \|A^*A\|_{2,2} \le L\}.$$

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

$$\begin{aligned} MSE^t(\mathcal{P}_L^{n,N}, \mathbf{a}) &= \max_{f\in\mathcal{F}} d(x_{\mathbf{a},f}^t, X_f^*), \\ PE^t(\mathcal{P}_L^{n,N}, \mathbf{a}) &= \max_{f\in\mathcal{F}} [f(x_{\mathbf{a},f}^t) - f(x_f^*)], \end{aligned}$$

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

$$\begin{aligned} MSE^t(\mathcal{P}_L^{n,N}, \mathcal{A}) &= \min_{\mathbf{a}\in\mathcal{A}} MSE^t(\mathcal{P}_L^{n,N}, \mathbf{a}), \\ PE^t(\mathcal{P}_L^{n,N}, \mathcal{A}) &= \min_{\mathbf{a}\in\mathcal{A}} PE^t(\mathcal{P}_L^{n,N}, \mathbf{a}). \end{aligned}$$

Theorem II.1. *Consider the class of functions $\mathcal{P}_L^{n,N}$. There exists a constant C such that for any given $L, n < N$, and for any $t \le \frac{n-1}{2}$,*

$$\begin{aligned} PE^t(\mathcal{P}_L^{n,N}, \mathcal{A}) &\ge CL\frac{\|x^*\|^2}{(t+1)^2}, \\ MSE^t(\mathcal{P}_L^{n,N}, \mathcal{A}) &\ge \frac{1}{25}\|x^*\|_2^2. \end{aligned}$$

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 $\|x^t - x_o\|_2^2$ on the worst possible case we work with average case quantity $\mathbb{E}(\|x^t - x_o\|_2^2)$. 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 ν 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_o [2]. ϵ is the probability that an element of x_o is non-zero. In other words, $x_{oi} \sim (1-\epsilon)\delta_0(x_{oi}) + \epsilon G(x_{oi})$, where $x_{o,i}$ is the i^{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 $t(\alpha) = \inf\{t_0 : \lim_{N\to\infty} \mathbb{E}\|x^t - x_o\|_2^2/\mathbb{E}\|x_o\|_2^2 \le \alpha \ \ \forall t > t_0\}$. $t(\alpha)$ 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 $\alpha < \lim_{t\to\infty} \lim_{N\to\infty} \mathbb{E}\|x^t - x_o\|_2^2/\mathbb{E}\|x_o\|_2^2$. Also according to [13] $\lim_{N\to\infty} \mathbb{E}\|x^t - x_o\|_2^2$ for these problem instances can converge to its final value exponentially fast.

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

Definition III.2. *Two problem instances Θ and Θ' are called equivalent if and only if Θ is less favorable than Θ' 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.[3] Furthermore, the non-zero elements of the vector x_o are sampled randomly from a given distribution G. The observed behavior of $\mathbb{E}\|x^t - x_o\|_2^2/N$ 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

[2] It can be replaced with the distribution function as well. However, for the simplicity of notation we assume that probability density function exists.

[3] We assume that $\mathbb{E}(A_{ij}) = 0$ and $\mathbb{E}(A_{ij}^2) = \frac{1}{n}$.

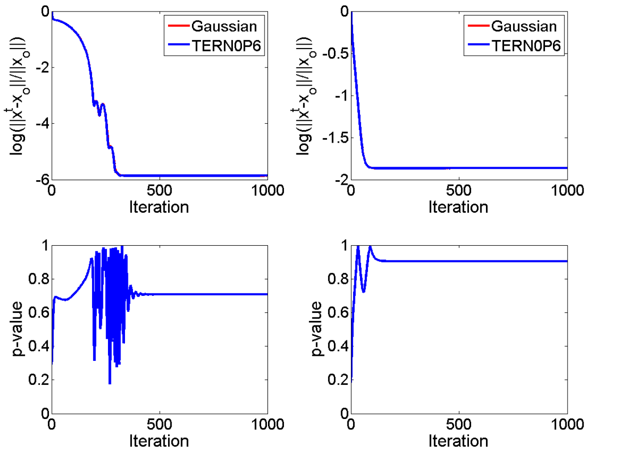

Fig. 1. Checking the matrix universality hypothesis. Top-Left: logarithm of the mean square error of FISTA for two different matrix ensembles defined in Table I at $N = 2000$, $\delta = .5$, $\lambda = .001$. Bottom-Left: The p-values of the Z-test on the null hypothesis of equality of mean square errors. Top-Right: logarithm of the mean square error of IST for two different matrix ensembles defined in Table I at $N = 2000$, $\delta = .5$, $\lambda = .1$. Bottom-right: The p-values. The large p-values confirm that we cannot reject the null hypothesis.

problem instances $(D_A, G, \epsilon; \delta, \nu)$ and $(N(0, 1/n), G, \epsilon; \delta, \nu)$ are equivalent.

We used a vague statement of "well-behaved" since the exact specifications of the universality class are not known yet. Proving the above universality hypothesis is clearly more difficult than the other universality conjectures that are still open in CS [20]. Therefore, following [20] we use an statistical analysis to test the above hypothesis. We tested this hypothesis on several matrix ensembles defined in Table I. For more information on our experimental set up and the statistical tests refer to [19]. Figure 1 shows the result of one of our experiments.

TABLE I
MATRIX ENSEMBLES CONSIDERED IN THE MATRIX UNIVERSALITY HYPOTHESIS TESTS.

Name	Specification
RSE	iid elements equally likely to be $\frac{\pm 1}{\sqrt{n}}$
USE	iid elements $N(0, 1/n)$
TERN	iid elements equally likely to be 0, $\sqrt{3/2n}$, $\sqrt{-3/2n}$
TERN0P6	iid elements taken values 0, $\sqrt{5/2n}, -\sqrt{5/2n}$ with $P(0) = .6$

The other important factor on the performance of the algorithms is the distribution of the input vector. First assume that we are considering a class of problem instances in which $D_A, \epsilon, \delta, \nu$ are fixed and the only thing that can change is G. Call this class $\mathcal{I}(D_A, \epsilon; \delta, \nu)$.

Lemma III.3. *Fix $D_A, G, \epsilon; \delta, \nu$ and suppose $E_G(X^2)$ all the problem instances of the form $(D_A, |\alpha| G(\alpha\mu), \epsilon; \delta, \nu)$ resulting from varying $\alpha \neq 0$ are equivalent for FISTA, IST, and AMP.*

The proof is very simple and skipped. Based on the above lemma define the class $\mathcal{I}^N(D_A, \epsilon; \delta, \nu) = \{(D_A, G, \epsilon; \delta, \nu) \in \mathcal{I}(D_A, \epsilon; \delta, \nu) \ : \ E_G(X^2) = 1\}$.

Theorem III.4. *$G = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$ results in the least favorable problem instance for AMP algorithm on $\mathcal{I}^N(D_A, \epsilon; \delta, \nu)$.*

This theorem which is proved in Section IV-B states that as far as the rate of convergence of the mean square error is concerned the constant amplitude distribution is the least favorable.

Theorem III.5. *If the least favorable problem exists for IST or FISTA, it will necessarily be $G = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$.*

The above theorem confirms that the constant amplitude ensemble results in the largest values of $t(\alpha)$ for at least some values of α. But it is not as strong as Theorem III.4 that we proved for AMP; it does not prove that this is the least favorable distribution. However, the empirical observations confirm that the least favorable distribution exists for these two algorithms and therefore according to the above theorem it is equal to $\frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$.

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.

Name	Specification
3P	iid elements taking value $0, 1, -1$ with $P(0) = 1 - \epsilon$
5P1	iid elements taking values $0, \pm 1, \pm 5$ with $P(1) = P(-1) = .3\epsilon$ and $P(5) = P(-5) = .2\epsilon$
5P2	iid elements taking values $0, \pm 1, \pm 20$ with $P(1) = P(-1) = .3\epsilon$ and $P(20) = P(-20) = .2\epsilon$
U1	iid $U[0, 4]$
G1	iid $N(0, 4)$

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

Theorem III.6. *If $\epsilon < \epsilon'$, ϵ' 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 δ, if $\epsilon < \epsilon'$, ϵ can not be less favorable for FISTA and IST than ϵ'.*

Clearly, the above theorem does not prove that ϵ is less favorable than ϵ' for IST or FISTA. However, empirical observations confirm this.

Empirical Finding 3. For a fixed value of δ 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

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

$X \sim (1-\epsilon)\delta_0(\mu) + \epsilon G(\mu)$ 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 Ψ function. It is proved that the Ψ 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 ($\theta_\infty = \tau\sqrt{\frac{m_\infty}{\delta} + \nu}$) and equilibrium detection rate ($\mathrm{EqDR} = \mathbb{P}\{|\eta(Y_\infty; \theta_\infty| \neq 0\}$, where $Y_\infty = X + \sqrt{\frac{m_\infty}{\delta} + \nu}Z$). m_∞ represents the fixed point of Ψ 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(λ) have statistically equivalent observables. In particular the MSE have the same value when τ and λ satisfy the following relation:*

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

Lemma IV.2. *[21] Define τ^0, so that $\mathrm{EqDR}(\tau) \leq \delta$ when $\tau > \tau^0$. For each λ there is a unique value of $\tau(\lambda) \in [\tau_0, \infty)$ such that*

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

The above discussion is mainly used in the calibration of λ 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 $G_\epsilon(\mu) \triangleq \epsilon G(\mu) + (1-\epsilon)\delta_0(\mu)$ and $F_{\epsilon,1} \triangleq \frac{\epsilon}{2}\delta_1(\mu) + \frac{\epsilon}{2}\delta_{-1}(\mu) + (1-\epsilon)\delta_0(\mu)$. Our goal is to prove that $F_{\epsilon,1}$ is less favorable than G_ϵ. For a given τ in AMP for $F_{\epsilon,1}$, we choose $\tau_G = \tau$. Let m_G^t be the mean square error on G_ϵ problem instance at time t and m_F^t represent the same thing for $F_{\epsilon,1}$. Suppose that $MSE_G^t \leq MSE_F^t$ at time t, our goal is to first prove that $m_G^{t+1} \leq m_F^{t+1}$. We have

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

Also let us define,

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

Our claim is that,

$$m_G^{t+1} \leq \tilde{m}_F^{t+1} \leq m_F^{t+1}.$$

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 $\mathbb{E}_X(\eta(X + \sqrt{\frac{MSE_G^t}{\delta} + \nu}Z; \tau\sqrt{\frac{MSE_G^t}{\delta} + \nu}) - X)$ 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 any $\nu > 0$ if*

$$\lim_{t\to\infty}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{F_{\epsilon,1}}\mathbb{E}_{x_o}\|\hat{x}_\lambda - x_o\|_2^2 \leq \epsilon,$$

then there exists a corresponding λ_G such that,

$$\lim_{t\to\infty}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_G\mathbb{E}_{x_o}\|\hat{x}_{\lambda_G} - x_o\|_2^2$$
$$= \lim_{t\to\infty}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{F_{\epsilon,1}}\mathbb{E}_{x_o}\|\hat{x}_\lambda - x_o\|_2^2,$$

Proof: According to Lemma IV.2, there exist a value of τ for which the asymptotic mean square error of the state evolution frameworks is the same as the asymptotic mean square error of LASSO(λ). Here is our approach for solving this lemma. The first thing that we want to prove is that there exists a value of $\underline{\tau}$ for which the asymptotic mean square error of AMP($\underline{\tau}$) on G is below the MSE of AMP(τ) on $F_{\epsilon,1}$. Then we prove there exists a value of $\overline{\tau}$ for which the asymptotic mean square error of AMP($\overline{\tau}$) is larger than the mean square error of AMP(τ) on $F_{\epsilon,1}$. Finally we will use implicit function theorem to prove that there exists a value of τ_G for which the asymptotic mean square error of AMP(τ_G) on G is exactly the same as the MSE of AMP(τ) on $F_{\epsilon,1}$. Choice $\overline{\tau} = \infty$ is clear. Therefore we focus on constructing a choice for $\underline{\tau}$. The claim is that $\underline{\tau} = \tau$. This is due to the concavity of $\mathbb{E}_X(\eta(X + \sqrt{m/\delta Z}; \beta) - X)^2$ 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 λ on the problem instance $(D_A, \frac{1}{2}\delta_1(\mu) + \frac{1}{2}\delta_{-1}\mu, \epsilon; \delta, \nu)$ we can calculate λ_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 λ that generates the same mean square error on $F_{\epsilon,1}$ and there may be more than one value of λ_G 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* $r_s(\mu, \tau; \sigma) = \mathbb{E}(\eta(\mu + \sigma Z; \tau\sigma) - \mu)^2$*, we have*

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

Lemma IV.5. *Consider the following risk function,*

$$R(\epsilon) = \mathbb{E}_{X \sim F_{\epsilon,1/\sqrt{\epsilon}}} \mathbb{E}_X(\eta(X + \sigma Z; \tau\sigma) - X)^2;$$

$R(\epsilon)$ *is an increasing function of* ϵ.

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

Proof of Theorem III.6

Proof: According to Lemma III.3, $F_{\epsilon,1}$ is equivalent to $F_{\epsilon,1/\sqrt{\epsilon}}$ and $F_{\epsilon',1}$ is equivalent to $F_{\epsilon',1/\sqrt{\epsilon'}}$. Therefore we compare these two distributions. We use the mathematical induction to prove the above theorem. Let m_ϵ^t be the mean squared error of AMP(τ) at iteration t on ϵ. Suppose that $m_\epsilon^t > m_{\epsilon'}^t$ and the goal is to prove that $m_\epsilon^{t+1} > m_{\epsilon'}^{t+1}$. Define

$$\tilde{m}_\epsilon^{t+1} = \mathbb{E}_{F_{\epsilon,\frac{1}{\sqrt{\epsilon}}}} \mathbb{E}_X(\eta(X + \sqrt{\frac{m_{\epsilon'}^t}{\delta} + \nu}Z; \tau\sqrt{\frac{m_{\epsilon'}^t}{\delta} + \nu}) - X)^2.$$

We claim $m_\epsilon^{t+1} > \tilde{m}_\epsilon^{t+1} > m_{\epsilon'}^{t+1}$. 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. ■

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.

Acknowledgment

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

References

[1] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. on Pure and Applied Math.*, 75:1412–1457, 2004.

[2] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. and Sim.*, 4(4):1168–1200, 2005.

[3] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM J. on Imag. Sci.*, 1(1):143–168, 2008.

[4] M. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. of Sel. Topics of Sig. Proc.*, 1(4):586–598, 2007.

[5] M. Figueiredo and R. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Trans. on Imag. Proc.*, 12(8):906–916, 2003.

[6] M. Figueiredo, J. Bioucas-Dias, and R. Nowak. Majorization-minimization algorithms for wavelet-based image restoration. *IEEE Trans. on Imag. Proc.*, 16(12):2980–2991, 2007.

[7] S. Wright and M. Figueiredo R. Nowak. Sparse reconstruction by separable approximation. *IEEE Trans. on Sig. Proc.*, 57(7):2479–2493, 2009.

[8] M. Elad, B. Matalon, J. Shtok, and M. Zibulevsky. A wide-angle view at iterated shrinkage algorithms. *Proc. SPIE (Wavelet XII)*, August 2007.

[9] E. Hale, W. Yin, and Y. Zhang. Fixed point continuation method for ℓ_1 minimization with application to compressed sensing. *Rice University Technial Report TR07-07*, 2007.

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

[11] J. Bioucas-Dias and M. Figueiredo. A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Trans. on Imag. Proc.*, 16:2992–3004, 2007.

[12] A. Maleki. Coherence analysis of iterative thresholding algorithms. *Preprint*, 2010.

[13] D. L. Donoho, A. Maleki, and A. Montanari. Message passing algorithms for compressed sensing. *Proc. of Nat. Acad. of Sci.*, 106(45):18914–18919, 2009.

[14] A. Maleki and D. L. Donoho. Optimally tuned iterative thresholding algorithm for compressed sensing. *IEEE J. of Sel. Areas in Sig. Proc.*, April 2010.

[15] Y. Nesterov. Gradient methods for minimizing composite objective function. *CORE Report*, 2007.

[16] S. Becker, J. Bobin, and E. Candès. Nesta: a fast and accurate first-order method for sparse recovery. 2010. submitted for publication.

[17] K. Bredies and D. Lorenz. Linear convergence of iterative soft-thresholding. *J. of Four. Anal. and App.*, 14:813–837, 2008.

[18] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87. Kluwer Academic Publishers, 2004.

[19] A. Maleki and R. Barraniuk. Difficult compressed sensing problem for first order methods. *Prepreint*. http://www.stanford.edu/∼arianm/isit11.pdf.

[20] D. L. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with applications in modern signal processing and data analysis. *Phil. Trans. of Roy. Soc. A*, 367(1906):4273–4293, 2009.

[21] D. L. Donoho, A. Maleki, and A. Montanari. Noise sensitivity phase transition. *IEEE Transactions on Information Theory*, 2010. submitted.

[22] M. Bayati and A. Montanri. The dynamics of message passing on dense graphs, with applications to compressed sensing. *Preprint*, 2010.