On Sparse Solutions of Underdetermined Linear Systems

Ming-Jun Lai
Department of Mathematics
the University of Georgia
Athens, GA 30602

January 17, 2009

Abstract

We first explain the research problem of finding the sparse solution of underdetermined linear systems with some applications. Then we explain three different approaches how to solve the sparse solution: the ℓ_1 approach, the orthogonal greedy approach, and the ℓ_q approach with $0 < q \leq 1$. We mainly survey recent results and present some new or simplified proofs. In particular, we give a good reason why the orthogonal greedy algorithm converges and why it can be used to find the sparse solution. About the restricted isometry property (RIP) of matrices, we provide an elementary proof to a known result that the probability that the random matrix with iid Gaussian variables possesses the PIP is strictly positive.

1 The Research Problem

Given a matrix Φ of size $m \times n$ with $m \leq n$, let

$$\mathcal{R}_k = \{\Phi\mathbf{x}, \mathbf{x} \in \mathbf{R}^n, \ \|\mathbf{x}\|_0 \leq k\}$$

be the range of Φ of all the k-component vectors, where $\|\mathbf{x}\|_0$ stands for the number of the nonzero components of $\mathbf{x}$.

Throughout this article, Φ is assumed to be of full rank. For a vector $\mathbf{y} \in \mathcal{R}_k$, we solve the following minimization problem

$$\min\{\|\mathbf{x}\|_0, \quad \mathbf{x} \in \mathbb{R}^n, \Phi\mathbf{x} = \mathbf{y}\}. \tag{1}$$

The solution of the above problem is called the sparse solution of $\mathbf{y} = \Phi\mathbf{x}$.

It is clear that the above problem can be solved in a finite time. Indeed, write $\Phi = [\phi_1, \phi_2, \cdots, \phi_n]$ with ϕ_i being a $m \times 1$ vector. One can choose m columns, say $A = [\phi_{i_1}, \cdots, \phi_{i_m}]$ from Φ to form a $m \times m$ linear system: $A\mathbf{z} = \mathbf{y}$. If A is nonsingular, one can find a solution $\mathbf{z}$. By exhausting all $m \times m$ nonsingular submatrices from Φ and solving all such linear system of equations, one can see which solution has the smallest number of nonzero entries.

However, there could be C_m^n such nonsingular linear systems from $\Phi\mathbf{x} = \mathbf{y}$ which need to be solved. For example, a rectangular matrix Φ with entries $(x_j)^i, i = 0, \cdots, m, j = 1, \cdots, n$ for distinct real numbers x_i's. Any $m \times m$ sub-matrix from Φ is of full rank. The number C_m^n grows exponentially fast as m and n go to ∞. For example, when $n = 2m$, $C_m^n \approx 2^n$. A common case $m = 512$ and $n = 1024$ needs to solve at least 2^{512} linear systems of size 512×512. This is impossible to do within a hour using current available computer. That is, the above method to solve Eq. (1) needs non-polynomial time. Are there any other methods to solve the above problem? Before we answer this question, let us see why we want to solve the problem in the next section.

Remark 1.1 *When $m = n$, $\Phi\mathbf{x} = \mathbf{y}$ is a standard linear system and the solution is unique if Φ is of full rank. We have already known the Gaussian elimination method can be used to solve such linear systems.*

Remark 1.2 *When $m > n$, one may not be able to have $\Phi\mathbf{x} = \mathbf{y}$. Instead, one asks to find $\mathbf{x}$ which minimizes the quantity $\|\Phi\mathbf{x} - \mathbf{y}\|_2$, where $\|\cdot\|_2$ is the discrete ℓ_2 norm. This is a standard least squares problem. When Φ is not full rank, one usually solves the following minimal norm solution using standard least squares methods. That is, find $\mathbf{x}$ such that*

$$\min\{\|\mathbf{x}\|_2, \quad \mathbf{x} \in S_A \subset \mathbb{R}^n, \}. \tag{2}$$

and

$$S_A := \{\mathbf{x} \in \mathbb{R}^n, \|\Phi\mathbf{x} - \mathbf{y}\|_2 = \min_{\mathbf{z}} \|\Phi\mathbf{z} - \mathbf{y}\|_2\}. \tag{3}$$

The solution can be found by using the pseudo inverse or using the singular value decomposition.

Remark 1.3 *When each column of Φ is normalized to be 1, Φ is called a dictionary. When $\Phi\Phi^T = I_m$ with identity matrix I_m of size $m \times m$, Φ is called a tight frame. We shall use these two concepts in later sections.*

2 Why do we find the sparse solutions?

In this section we give several reasons why we want to solve the sparse solution of underdetermined systems of linear equations.

2.1 Motivation: Signal and Image Compression

This is the most direct and natural application. Suppose that a signal or an image $\mathbf{y}$ is represented by using a tight frame Φ of size $m \times n$ with $m < n$. We look for a sparse approximation $\mathbf{x}$ satisfying

$$\min\{\|\mathbf{x}\|_0, \quad \mathbf{x} \in \mathbb{R}^n, \|\Phi\mathbf{x} - \mathbf{y}\| \leq \theta\}, \tag{4}$$

where $\theta > 0$ is a tolerance. In particular, for lossless compression, i.e., $\theta = 0$, the above (4) is the our research problem (1).

2.2 Motivation: Compressed Sensing

We are interested in economically recording information about a vector $\mathbf{x}$ in $\mathbb{R}^n$. First of all, we allocate m nonadaptive questions to ask about $\mathbf{x}$. Each question takes the form of a linear functional applied to $\mathbf{x}$. Thus, the information we obtain from the questions is given by

$$\mathbf{y} = \Phi\mathbf{x},$$

where Φ is a matrix of $m \times n$. In general, m is much smaller than n since $\mathbf{x}$ is a compressible data vector. Let Δ be a decoder that provides an approximation $\mathbf{x}^*$ to $\mathbf{x}$ using the information that $\mathbf{y}$ holds. That is, $\Delta\mathbf{y} = \mathbf{x}^* \approx \mathbf{x}$. Typically, the mapping Δ is nonlinear. The central question of compressed sensing is to find a good set of questions and a good decoder (Φ, Δ) so that we can find a good approximation $\mathbf{x}^*$ of $\mathbf{x}$. See, e.g. [Candés'06].

For example, when $\mathbf{x} \in \mathbf{R}^n$ with $\|\mathbf{x}\|_0 \leq k << n$, one wants to know which components of $\mathbf{x}$ are not zero and what values are. We can design some vectors Φ to question $\mathbf{x}$ by inner product. Thus, we get $\mathbf{y} = \Phi\mathbf{x} \in \mathcal{R}_k$. To find $\mathbf{x}$, we solve our research problem (1).

For another example, suppose that a data vector $\mathbf{z}$ is a set of compressible data, i.e., there exists a vector $\mathbf{x} \in \mathbf{R}^n$ with only k nonzero entries such that $\mathbf{z} = A\mathbf{x}$ for an invertible matrix A of $n \times n$. Suppose that the questions can be represented in the form $\mathbf{y} = C\mathbf{z}$ with $m \times n$ matrix C. Since $\mathbf{z} \in \mathcal{R}_k$, i.e., $\mathbf{z} = A\mathbf{x}$, we have

$$\mathbf{y} = CA\mathbf{x}. \tag{5}$$

Certainly it is necessary to question $\mathbf{z}$ m times with $m > k$. That is, we have $k < m < n$.

If C is chosen in the form ΦA^{-1} for some rectangular matrix Φ of size $m \times n$, we need to solve the following minimization problem in order to record the data $\mathbf{z}$ economically.

$$\mathbf{y} = C\mathbf{z} = \Phi A^{-1}A\mathbf{x} = \Phi\mathbf{x}. \tag{6}$$

The problem is to find the sparse representation $\mathbf{x}$ satisfying the above (6) which is the same as (1).

2.3 Motivation: Error Correcting Codes

Let $\mathbf{z}$ be a vector encoded $\mathbf{x}$ by a redundant linear system A of size $m \times n$ with $m > n$. That is, $\mathbf{z} = A\mathbf{x}$ is transmitted through a noisy channel. The channel corrupts some random entries of $\mathbf{z}$, resulting a new vector $\mathbf{w} = \mathbf{z} + \mathbf{v}$. Finding the vector $\mathbf{v}$ is equivalent to correcting the errors.

To this end, we extend A to a square matrix B of size $m \times m$ by adding $A^\perp$, i.e, $B = [A; A^\perp]$. Assume that A satisfies $A^T A = I_n$, where I_n is the identity matrix of n. Then we can choose $A^\perp$ such that $BB^T = I_m$ the identity matrix of size m. Clearly,

$$B^T\mathbf{w} = B^T\mathbf{z} + B^T\mathbf{v} = \begin{bmatrix} \mathbf{x} \\ 0 \end{bmatrix} + \begin{bmatrix} A^T\mathbf{v} \\ (A^\perp)^T\mathbf{v} \end{bmatrix}.$$

Let $\mathbf{y} = (A^\perp)^T\mathbf{v}$ which is the last $m - n$ entries of $B^T\mathbf{w}$. Since $\mathbf{z}$ is in the codeword space V which is a linear span of columns of the matrix A, $(A^\perp)^T\mathbf{v}$

is not in the codeword space and is the only information about $\mathbf{v}$ available to the receiver.

If the receiver is able to solve the minimization problem Eq.(1) with $\Phi = (A^{\perp})^T$. That is, find the sparsest solution $\mathbf{v}$ such that $\mathbf{y} = (A^{\perp})^T\mathbf{v}$. Then we can get the correct $\mathbf{x}$. Thus, this error correcting problem is again equivalent to the sparsest solution problem Eq. (1). See [Candes and Tao'05] and [Candes, Romberg, Tao'06] for more detail.

2.4 Motivation: Cryptography

Although large prime numbers are currently used for secure data transmission, it is possible to use underdetermined systems of linear equations instead. The ideas can be described as follows. Suppose that we have a class of matrices Φ of size $m \times n$ with $m < n$ which admit a computationally efficient algorithm for solving the minimization Eq.(1) for any given $\mathbf{y}$ which is in the range $\mathcal{R}_k$. Let Ψ be an invertible random matrix of size $m \times m$ and $A = \Psi\Phi$. Suppose that a receiver wants to get a secret data vector $\mathbf{x}$ from a customer, e.g., a vector consisting of credit card number, expiration date, and the name on the credit card. The receiver sends to the customer the matrix A in a public channel. After receiving A the customer computes $\mathbf{z} = A\mathbf{x}$ and sends $\mathbf{z}$ to the receiver in a public channel. As we mentioned above, finding the sparse solution $\mathbf{x}$ from $\mathbf{z}$ using matrix A is non-polynomial time. With overwhelming probability, such $\mathbf{x}$ can not be found by other parties.

However, the receiver is able to get $\mathbf{x}$ by solving $\mathbf{y} = \Psi^{-1}\mathbf{z} = \Phi\mathbf{x}$ which is our research problem Eq. (1). By changing Ψ frequently enough, the receiver is able to get the secured data every time while the hacker is impossible to decode the data.

2.5 Motivation: Recovery of Loss Data

Let $\mathbf{z}$ be an image and $\widetilde{\mathbf{z}}$ be a partial image of $\mathbf{z}$. That is, $\mathbf{z}$ loses some data to become $\widetilde{\mathbf{z}}$. Suppose that we know the location where the data are lost. We would like to recover the original image from the partial image $\widetilde{\mathbf{z}}$. Let Φ be a tight wavelet frame such that $\mathbf{x} = \Phi\mathbf{z}$ is the most sparse representation for $\mathbf{z}$. Let Ψ be the residual matrix from Φ by dropping off the columns corresponding to the unavailable entries, i.e., the missing data locations. Note that $\Phi^T\Phi = I_m$ and hence, $\Psi^T\Psi = I_\ell$ with $\ell < m$.

It follows that $\Psi^T \mathbf{x} = \Psi^T \Phi \mathbf{z} = \widetilde{\mathbf{z}}$ by the orthonormality of columns of Φ. Thus, we need to find the sparsest solution $\mathbf{x}$ from the given $\widetilde{\mathbf{z}}$ such that $\widetilde{\mathbf{z}} = \Psi^T \mathbf{x}$ which is exactly the same problem as our research problem(1). Once we have $\mathbf{x}$, we can find $\mathbf{z}$ which is $\mathbf{z} = \Phi \mathbf{x}$. See [Aharon, Elad, Bruckstein'06] for numerical experiments.

3 The ℓ_1 Approach

Although the problem in Eq. (1) needs a non-polynomial time to solve (cf. [Natarajan95]) in general, it can be much more effectively solved by using many other methods, e.g., ℓ^1 minimization approach, reweighted ℓ^1 method, OGA(orthogonal greedy algorithm), and the ℓ_q approach. Let us review these approaches in the following subsections and following sections.

The ℓ_1 minimization problem is the following

$$\min\{\|\mathbf{x}\|_1, \quad \mathbf{x} \in \mathbb{R}^n, \Phi \mathbf{x} = \mathbf{y}\}, \tag{7}$$

where $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$ for $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$. The solution $\Delta_1 \Phi \mathbf{x}$ is called the ℓ_1 solution of $\mathbf{y} = \Phi \mathbf{x}$. Since the ℓ_1 minimization problem is equivalent to the linear programming, this converts the problem into a tractable computational problem. (See [Lai and Wenston'04] for a justification of the equivalence and a computational algorithm for ℓ_1 minimization.) A matlab ℓ_1 minimization program is available on-line.

But one has to study when the (P1) solution (the solution of Eq. (7)) is also the (P0) solution (the solution of Eq. (1)). There are two concepts: mutual coherence(MC) and restricted isometric property (RIP) of the matrix Φ to help describe the situation.

3.1 The Mutual Coherence

Let us begin with the *spark* of matrix A, the smallest possible number σ such that there exists σ columns from A that are linearly dependent. It is clear that $\sigma(A) \leq \text{rank}(A) + 1$. The following theorem is belong to [Donoho and Elad'03].

Theorem 3.1 *A representation* $\mathbf{y} = \Phi \mathbf{x}$ *is necessarily the sparsest possible if* $\|\mathbf{x}\|_0 < \textit{spark}(\Phi)/2$.

Proof. Suppose that there are two sparse solutions $\mathbf{x}^1$ and $\mathbf{x}^2$ with $\|\mathbf{x}^1\|_0 \leq k$ and $\|\mathbf{x}^2\|_0 \leq k$ solving $\mathbf{y} = \Phi\mathbf{x}$. Then $\Phi(\mathbf{x}^1 - \mathbf{x}^2) = 0$. So $\|(\mathbf{x}^1 - \mathbf{x}^2)\|_0 \leq 2k$ but, $\|(\mathbf{x}^1 - \mathbf{x}^2)\|_0 \geq \text{spark}(\Phi)$. It follows that $k \geq \text{spark}(\Phi)/2$. Hence, when $k < \text{spark}(\Phi)/2$, the sparsest solution is unique. That is, if one find a solution $\mathbf{x}$ of $\Phi\mathbf{x} = \mathbf{y}$ with $\|\mathbf{x}\|_0 < \text{spark}(\Phi)/2$, then $\mathbf{x}$ is the sparse solution. ■

Next we introduce the concept of mutual coherence of matrix Φ. Assume that each column of Φ is normalized. That is, Φ is a dictionary. Let $G = \Phi^T\Phi$ which is a square matrix of size $n \times n$. Write $G = (g_{ij})_{1\leq i,j\leq n}$, the mutual coherence of Φ is

$$M = M(\Phi) = \max_{\substack{1\leq i,j\leq n\\ i\neq j}} |g_{ij}|.$$

Clearly, $M \leq 1$. We would like to have matrix Φ such that its mutual coherence M is as small as possible.

However, $M(\Phi)$ can not be too small. We have

Lemma 3.2 *If $n \geq 2m$, then $M(\Phi) \geq (2m)^{-1/2}$.*

Proof. Indeed, let $\lambda_i, i = 1, \cdots, n$ be eigenvalues of G. Since G is positive semi-definite, all $\lambda_i \geq 0$. Since the rank of G is equal to m, only m nonzero λ_i. Since $\sum_i \lambda_i$ is equal to the trace of G which is n since $g_{ii} = 1$. That is,

$$n = \sum_i \lambda_i \leq \sqrt{m}\sqrt{\sum_i \lambda_i^2}. \tag{8}$$

On the other hand, using a property of the Frobenius norm of G, we have

$$\sum_i \lambda_i^2 = \|G\|_F^2 = \sum_{1\leq i,j\leq n} (g_{ij})^2. \tag{9}$$

It follows from Eq. (8) and (9) that

$$(n^2 - n)M(\Phi)^2 + n \geq \sum_{1\leq i,j\leq n} (g_{ij})^2 \geq \frac{n^2}{m}$$

That is, $M(\Phi) \geq \sqrt{\frac{n-m}{m(n-1)}}$. In particular, when $n \geq 2m$, we have $M(\Phi) \geq (2m)^{-1/2}$. That is, $M(\Phi) \in ((2m)^{-1/2}, 1]$. ■

With $M(\Phi)$, we can prove the following (cf. [Donoho and Elad03])

Theorem 3.3 *Let Spark*(Φ) *be the spark of* Φ *and* $M(\Phi)$ *be the coherence of* Φ*. Then*

$$\textit{Spark}(\Phi) > 1/M(\Phi).$$

Next we need the following lemma.

Lemma 3.4 *Let* $k < 1/M + 1$*. For any* $S \subset \{1, \cdots, n\}$ *with* $\#(S) \leq k$ *and* Φ_S *be the matrix consisting of the* k *columns of* Φ *with column indices in* S*. Then the* k*th singular value of* Φ_S *is bounded below by* $(1 - M(k-1))^{1/2}$ *and above by* $(1 + M(k-1))^{1/2}$*.*

Proof. For any vector $\mathbf{v} \in \mathbf{R}^n$ with support on S, we have

$$\mathbf{v}^T G \mathbf{v} = \mathbf{v}_S \Phi_S^T \Phi_S \mathbf{v}_S = \|\mathbf{v}\|^2 + \sum_{\substack{i \neq j \\ i,j \in S}} v_i g_{ij} v_j.$$

Since

$$\begin{aligned} \|\sum_{\substack{i \neq j \\ i,j \in S}} v_i g_{ij} v_j\| &\leq M \sum_{\substack{i \neq j \\ i,j \in S}} |v_i v_j| \\ &\leq M \left(\sum_{i,j \in S} |v_i v_j| - \|\mathbf{v}\|_2^2 \right) \\ &\leq M \|\mathbf{v}\|_2^2 (k-1), \end{aligned}$$

we have

$$\mathbf{v}_S \Phi_S^T \Phi_S \mathbf{v}_S \geq \|\mathbf{v}\|_2^2 - M(k-1)\|\mathbf{v}\|_2^2 = (1 - M(k-1))\|\mathbf{v}\|_2^2.$$

Similarly,

$$\begin{aligned} \mathbf{v}^T G \mathbf{v} &= \mathbf{v}_S \Phi_S^T \Phi_S \mathbf{v}_S = \|\mathbf{v}\|^2 + \sum_{\substack{i \neq j \\ i,j \in S}} v_i g_{ij} v_j \\ &\leq \|\mathbf{v}\|_2^2 + M(k-1)\|\mathbf{v}\|_2^2 = (1 + M(k-1))\|\mathbf{v}\|_2^2. \end{aligned}$$

These complete the proof. ■

We first show that if $k < (1 + 1/M)/2$ and for any $\mathbf{y} \in \mathcal{R}_k$, the sparse solution of Eq. (1) is unique.

Lemma 3.5 *Suppose* $k < (1+1/M)/2$*. For any* $\mathbf{y} \in \mathcal{R}_k$*, the sparse solution of Eq. (1) is unique.*

Proof. Let $\mathbf{y} = \Phi\mathbf{x}_0 \in \mathcal{R}_k$. Let $\mathbf{x}_1$ be a solution of Eq. (1) with $\|\mathbf{x}_1\|_0 \le k$. Then $\|\mathbf{x}_0 - \mathbf{x}_1\|_0 \le 2k$. By Lemma 3.4, we have

$$(1 - M(2k-1))\|\mathbf{x}_0 - \mathbf{x}_1\|_2^2 \le \|\Phi(\mathbf{x}_0 - \mathbf{x}_1)\|_2^2 = 0.$$

Since $1 - M(2k-1) \neq 0$, it follows that $\mathbf{x}_0 = \mathbf{x}_1$. ■

In order to fully understand the computation of Eq. (7), we next introduce a variance of Eq. (7): solve the following minimization problem

$$\min\{\|\mathbf{x}\|_1, \quad \mathbf{x} \in \mathbb{R}^n, \|\Phi\mathbf{x} - \mathbf{y}\|_2 \le \delta\}, \tag{10}$$

where $\mathbf{y} = \Phi\mathbf{x}_0 + z$ with $\|z\|_2 \le \epsilon < \delta$ and $\mathbf{x}_0 \in \mathbb{R}^n$ is a vector with k nonzero entries, that is, $\Phi\mathbf{x}_0 \in \mathcal{R}_k$. Hence, we consider the case that $\mathbf{y}$ has some measurement error and compute a solution $\mathbf{x}$ within accuracy δ.

Theorem 3.6 *Let M be the mutual coherence of Φ. Suppose that*

$$k < (1/M + 1)/4.$$

For any $\mathbf{x}_0$ with $\|\mathbf{x}_0\|_0 \le k$, let $\widehat{\mathbf{x}}_{\epsilon,\delta}$ be the solution of Eq. (10). Then

$$\|\widehat{\mathbf{x}}_{\epsilon,\delta} - \mathbf{x}_0\|_2^2 \le \frac{(\epsilon + \delta)^2}{1 - M(4k-1)}.$$

Proof. Write $w = \widehat{\mathbf{x}}_{\epsilon,\delta} - \mathbf{x}_0$. Clearly, $\|\widehat{\mathbf{x}}_{\epsilon,\delta}\|_1 = \|w + \mathbf{x}_0\|_1 \le \|\mathbf{x}_0\|_1$ when computing the ℓ_1 minimization. Let $S \subset \{1, 2, \cdots, n\}$ be the index set where $\mathbf{x}_0$ is supported. Since $\|w + \mathbf{x}_0\|_1 \ge \|\mathbf{x}_0\|_1 - \sum_{i \in S}|w_i| + \sum_{i \in \hat{S}}|w_i|$, we have $\sum_{i \in \hat{S}}|w_i| \le \sum_{i \in S}|w_i|$ or

$$\|w\|_1 \le 2\sum_{i \in S}|w_i| \le 2\sqrt{k}\|w\|_2, \tag{11}$$

where $\hat{S}$ denotes the complement set of S in $\{1, 2, \cdots, n\}$.

On the other hand, $\|\Phi\widehat{\mathbf{x}}_{\epsilon,\delta} - \mathbf{y}\|_2 \le \delta$ and $y = \Phi\mathbf{x}_0 + z$ imply that $\|\Phi w + z\|_2 \le \delta$. That is, $\|\Phi w\|_2 \le \|\Phi w + z\|_2 + \epsilon \le \delta + \epsilon$.

Finally, $\|\Phi w\|_2^2 = \|w\|_2^2 + w^T(G - I)w \ge \|w\|_2^2 - M(\|w\|_1^2 - \|w\|_2^2) \ge (1 + M)\|w\|_2^2 - M4k\|w\|_2^2$ by the estimate (11) above. It follows that

$$\|w\|_2^2 \le \frac{1}{1 + M - 4Mk}\|\Phi w\|_2^2 \le \frac{(\epsilon + \delta)^2}{1 - M(4k-1)}$$

by the estimate in the previous paragraph. This concludes the result in this theorem. ■

Next we look at the (ϵ, δ) variance of the (P0) problem: to solve the following minimization problem

$$\min\{\|\mathbf{x}\|_0, \quad \mathbf{x} \in \mathbb{R}^n, \|\Phi\mathbf{x} - \mathbf{y}\|_2 \leq \delta\}, \tag{12}$$

where $\mathbf{y} = \Phi\mathbf{x}_0 + z$ with $\|z\|_2 \leq \epsilon$ and $\Phi\mathbf{x}_0 \in \mathcal{R}_k$. Then we can prove

Theorem 3.7 *Let M be the mutual coherence of Φ. Suppose that*

$$k < (1/M + 1)/2.$$

For any $\mathbf{x}_0$ with $\|\mathbf{x}_0\|_0 \leq k$, let $\widetilde{\mathbf{x}}_{\epsilon,\delta}$ be the solution of Eq. (12). Then

$$\|\widetilde{\mathbf{x}}_{\epsilon,\delta} - \mathbf{x}_0\|_2^2 \leq \frac{(\epsilon + \delta)^2}{1 - M(2k - 1)}.$$

Proof. By Lemma 3.4, we have

$$\begin{aligned} \|\widetilde{\mathbf{x}}_{\epsilon,\delta} - \mathbf{x}_0\|_2^2 &\leq \frac{1}{1 - M(2k-1)} \|\Phi(\widetilde{\mathbf{x}}_{\epsilon,\delta} - \mathbf{x}_0)\|_2^2 \\ &= \frac{1}{1 - M(2k-1)} \|\Phi\widetilde{\mathbf{x}}_{\epsilon,\delta} - \mathbf{y} + \mathbf{z}\|_2^2 \leq \frac{(\epsilon + \delta)^2}{1 - M(2k-1)}. \end{aligned}$$

This completes the proof. ■

Both theorems above were proved in [Donoho, Elad and Temlyakov'06]. In particular, the proof of Theorem 3.7 above is a much simplified version of the one in [Donoho, Elad and Temlyakov'06]. It is easy to see that there is a gap between the requirements of k. That is, one is to require $k < (1+1/M)/4$ by any ℓ_1 method and the other is to require $k < (1 + 1/M)/2$ by an ℓ_0 method. Thus, the ℓ_1 method is not optimal yet. It is interesting to know how we can increase k when using the ℓ_1 method.

3.2 RIP

Another approach is to use the so-called Restricted Isometry Property(RIP) of the matrix Φ. Letting $0 < k < m$ be an integer and A_T be a submatrix of A

which consists of columns of A whose column indices are in $T \subset \{1, 2, \cdots, n\}$, the k restricted isometry constant δ_k of A is the smallest quantity such that

$$(1 - \delta_k)\|\mathbf{x}\|_2^2 \leq \|A_T\mathbf{x}\|_2^2 \leq (1 + \delta_k)\|\mathbf{x}\|_2^2 \tag{13}$$

for all subset T with $\#(T) \leq k$. If a matrix A has such a constant $\delta_k > 0$ for some k, A possesses RIP. With this concept, it is easy to see that if $\delta_{2k} < 1$, then the solution of Eq. (1) is unique. Indeed, if there were two solutions $\mathbf{x}^1$ and $\mathbf{x}^2$ such that

$$\Phi(\mathbf{x}^1 - \mathbf{x}^2) = 0,$$

then we choose the index set T which contains the indices of the nonzero entries of $\mathbf{x}^1 - \mathbf{x}^2$ and see that $\#(T) \leq 2k$ which implies

$$(1 - \delta_{2k})\|\mathbf{x}^1 - \mathbf{x}^2\|_2^2 \leq \|\Phi_T(\mathbf{x}^2 - \mathbf{x}^2)\|_2^2 = 0.$$

It follows that $\|\mathbf{x}^1 - \mathbf{x}^2\|_2 = 0$ when $\delta_{2k} < 1$. That is, the solution is unique. Furthermore,

Theorem 3.8 ([Candes, Romberg, and Tao'06]) *Suppose that $k \geq 1$ such that*

$$\delta_{3k} + 3\delta_{4k} < 2$$

and let $\mathbf{x} \in \mathbf{R}^n$ be a vector with $\|\mathbf{x}\|_0 \leq k$. Then for $\mathbf{y} = \Phi\mathbf{x}$, the solution of Eq. (7) is unique and equal to $\mathbf{x}$.

This result is recently simplified slightly in the following way:

Theorem 3.9 ([Candes'08]) *Suppose that $k \geq 1$ such that*

$$\delta_{2k} < \sqrt{2} - 1.$$

Let $\mathbf{x} \in \mathbf{R}^n$ be a vector with $\|\mathbf{x}\|_0 \leq k$. Then for $\mathbf{y} = A\mathbf{x}$, the solution of Eq. (7) is unique and equal to $\mathbf{x}$. In fact

$$\|\mathbf{x} - \mathbf{x}^*\|_2 \leq \frac{2(1+\rho)}{1-\rho}\|A\mathbf{x} - A\mathbf{x}^*\|_2 + \frac{2}{1-\rho}\|\mathbf{x} - \mathbf{x}_T^*\|_1.$$

where $\mathbf{x}^*$ is the (P1) solution (the solution of Eq. (7) and $\mathbf{x}_T^*$ is the vector of the k largest components of $\mathbf{x}^*$. Here $\rho = \frac{\delta_{2k}}{\sqrt{2}-1}$.

As we have already known that $\delta_{2k} < \sqrt{2} - 1 < 1$ which implies that the sparse solution is unique. The above result mainly explains that the (P1) solution (the solution of Eq. (7)) is equal to the (P0) solution (the solution of Eq. (1)).

The results are consequences of the following Theorem 5.2 and hence we omit the proofs of the above two theorems here.

Let us discuss what kind of matrices Φ satisfies the RIP. So far there is no explicit construction of matrices of any size which possess the RIP. Instead, there are a couple of constructions based on random matrices which satisfy the RIP with overwhelming probability. In [Candés, Romberg, and Tao'06], the following results were proved using the measure concentration technique (cf. [Ledoux'01]).

Theorem 3.10 *Suppose that $A = [a_{ij}]_{1\le i\le m, 1\le j\le n}$ be a matrix with entries a_{ij} being iid Gaussian random variables with mean zero and variance $1/\sqrt{m}$. Then the probability*

$$\mathcal{P}\left(\left|\|\Phi\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \le \epsilon\|\mathbf{x}\|_2^2\right) \ge 1 - \binom{n}{k}(1+2/\epsilon)^k e^{-m\epsilon^2/c}. \qquad (14)$$

for any vector $\mathbf{x} \in \mathbf{R}^n$ with $\|\mathbf{x}\|_0 = k$, where $c > 2$ is a constant and $\|\mathbf{x}\|_0$ denotes the number of nonzero entries of vector $\mathbf{x}$.

Once we choose $k < m$ such that $\binom{n}{k}(1+2/\epsilon)^k e^{-m\epsilon^2/c} < 1$ small enough, we will have a good probability to have a matrix satisfying the RIP. Indeed, since $\binom{n}{k} \le (n/e)^k$,

$$\binom{n}{k}(1+2/\epsilon)^k e^{-m\epsilon^2/c} \le e^{-m\epsilon^2/c + k\ln(n/e) + k\ln(1+2/\epsilon)}.$$

As long as $m > kc\ln(n(1+2/\epsilon)/e)/\epsilon^2$, we have

$$\mathcal{P}\left(\left|\|\Phi\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \le \epsilon\|\mathbf{x}\|_2^2\right) > 0.$$

That is, a matrix with RIP can be found with positive probability.

Theorem 3.10 can be simply proved based on the following theorem (cf. [Baranuik, Daveport, DeVore, Wakin'08]).

Theorem 3.11 *Suppose that $A = [a_{ij}]_{1\le i\le m, 1\le j\le n}$ be a matrix with entries a_{ij} being iid Gaussian random variables with mean zero and variance $1/\sqrt{m}$. Then for any $\epsilon > 0$, the probability*

$$\mathcal{P}(\big|\|A\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| < \epsilon\|\mathbf{x}\|_2^2) \ge 1 - 2\exp(-\frac{\epsilon^2 m}{c}), \tag{15}$$

where c is a positive constant independent of ϵ and $\|\mathbf{x}\|_2$ for any $\mathbf{x} \in \mathbf{R}^n$.

In general, there are many other random matrices satisfying the above probability estimate. Typically, matrices with sub-Gaussian random variables possess the RIP. See [Mendelson, Pajor, and Tomczak-Jaegermann'07, '08]. In addition to the measure concentration approach, there are several other ideas to prove the results in the above theorem. For example, [Pisier'86] and [Lai'08]. We refer to [Lai'08] for an elementary proof of Theorems 3.11 and 3.10 and similar theorems for sub-Gaussian random matrices. For convenience, we borrow the proof of Theorem 3.11 from [Lai'08] and present it in the Appendix for interested reader.

3.3 The re-weighted ℓ_1 Method

The re-weighted ℓ_1 minimization is the following iterations:

(1) for $k = 0$, solve the standard ℓ_1 problem:

$$\min\{\|\mathbf{x}\|_1, \quad \mathbf{x} \in \mathbb{R}^n, A\mathbf{x} = \mathbf{y}\}, \tag{16}$$

(2) for $k > 0$, find $\mathbf{x}^{(k)}$ which solves the following weighted ℓ_1 problem:

$$\min \sum_{i=1}^n \frac{|x_i|}{w_i}, \quad \mathbf{x} \in \mathbb{R}^n, A\mathbf{x} = \mathbf{y}\}, \tag{17}$$

with $w_i = |x_i^{(k-1)}| + \epsilon$ for $k = 1, 2, 3, \cdots, n$.

This method is introduced in [Candés, Watkin, and Boyd'07]. The researchers gave some heuristic reasons that the algorithm above converges much faster than the standard ℓ_1 method. It is still interesting why the method works better in theory.

4 The OGA Approach

There are many versions of the Optimal Greedy Algorithm(OGA) available in the literature. See [Temlyakov'00], [Temlyakov'03], [Tropp'04], and [Petukhov'06]. We mainly explain the optimal greedy algorithm (OGA) proposed by A. Petukhov in 2006 when Φ is obtained from a tight wavelet frame. That is, Φ is a matrix whose columns are frame components $\phi_i, i = 1, \cdots, n$ satisfying $\Phi\Phi^T = I_m$, where I_m is the identity matrix of size $m \times m$. It has two distinct advantages: (1) Iterative steps for the least squares solution and (2) more than one terms are chosen in each iteration.

Let Λ be an index set which is a subset of $\{1, 2, \cdots, n\}$ and $\widetilde{\Lambda}$ be the complement of Λ in $\{1, 2, \cdots, n\}$. Also let P_Λ be the diagonal matrix of size $n \times n$ with entries to be 1 if the index is in Λ and 0 otherwise.

Suppose that we have a fixed index set Λ. We first introduce a computationally efficient algorithm for finding coefficients of the linear combination $f_\Lambda = \sum_{i\in\Lambda} a_i\phi_i$ which is the least squares approximation of f, i.e., $\|f - f_\Lambda\|_2 = \min\{\|f - g\|_2, g \in S_\Lambda\}$ where $f \in \mathbf{R}^m$ is a given vector in $\mathbf{R}^m$ and S_Λ is the span of $\phi_i, i \in \Lambda$. In general, f_Λ can be computed directly by inverting a Gram matrix $[\langle\phi_i, \phi_j\rangle]_{i,j\in\Lambda}$. When m is large, it is more efficient to use the following algorithm to find an approximation of f_Λ.

Algorithm LSA (least squares approximation): Set $k = 0, g^0 = f, f^0 = 0$. For $k \geq 1$, let $g^k = g^{k-1} - \Phi P_\Lambda \Phi^T g^{k-1}$ and $f^k = f^{k-1} + \Phi P_\Lambda \Phi^T g^{k-1}$. Stop the iterations when $g^k - g^{k-1}$ is very small.

We have the following

Theorem 4.1 *The sequence f^k converges to f_Λ in the following sense:*

$$\|f^k - f_\Lambda\|_2 \leq (1 - \gamma^2)^{k/2}\|f_\Lambda\|_2,$$

where γ is the least non-zero singular value of the matrix Φ_Λ.

Proof. We rewrite g^k as $g^k = g^k_\Lambda + \widetilde{g}^k$, where g^k_Λ is the best approximation of g^k using the span of columns from Φ_Λ. Clearly, $g^0_\Lambda = f_\Lambda$. For $k \geq 1$, $g^k_\Lambda = g^{k-1}_\Lambda - \Phi P_\Lambda \Phi^T g^{k-1}_\Lambda$ since the best approximation operator in $\mathbf{R}^m$ is a linear operator. Similar for $f^k_\Lambda = f^{k-1}_\Lambda + \Phi P_\Lambda \Phi^T g^{k-1}_\Lambda$. Note that $f^k_\Lambda = f^k$ for all $k \geq 0$. We have

$$f^k + g^k_\Lambda = f^k_\Lambda + g^k_\Lambda = f^{k-1}_\Lambda + g^{k-1}_\Lambda = \cdots = f^0_\Lambda + g^0_\Lambda = f_\Lambda.$$

It follows that

$$\|f^k - f_\Lambda\|_2 = \|g_\Lambda^k\|_2 = \|(I - \Phi P_\Lambda \Phi^T) g_\Lambda^{k-1}\|_2.$$

Note that $I - \Phi P_\Lambda \Phi^T = \Phi(I - P_\Lambda)\Phi^T$ and hence

$$\|I - \Phi P_\Lambda \Phi^T\|_2 \leq \|\Phi(I - P_\Lambda)\Phi^T\|_2 \leq (1-\gamma^2)^{1/2}.$$

Therefore,

$$\begin{aligned} \|f^k - f_\Lambda\|_2 &= \|g_\Lambda^k\|_2 \leq (1-\gamma^2)^{1/2}\|g_\Lambda^{k-1}\|_2 \\ &\leq \cdots \leq (1-\gamma^2)^{k/2}\|g_\Lambda^0\|_2 = (1-\gamma^2)^{k/2}\|f_\Lambda\|_2. \end{aligned}$$

This completes the proof. ■

We are now ready to present the Petukhov version of orthogonal greedy algorithm (OGA).

Algorithm OGA: Set $\Lambda_0 = \emptyset, g^0 = f, f^0 = 0$. Choose a threshold $r \in (0,1]$ and a precision $\epsilon > 0$;

Step 1. For $k \geq 1$, find $M_k = \max_{i \notin \Lambda_{k-1}} |\langle g^{k-1}, \phi_i/\|\phi_i\|\rangle|$; and Let $\Lambda_k = \Lambda_{k-1} \cup \{i, |\langle g^{k-1}, \phi_i/\|\phi_i\|\rangle \geq rM_k\}$;

Step 2. Apply Algorithm LSA above over Λ_k to approximate g^{k-1} to find f_{Λ_k} and g_{Λ_k}. Update $f^k = f^{k-1} + f_{\Lambda_k}$ and $g^k = g^{k-1} - f_{\Lambda_k}$.

Step 3. If $\|f - f^k\|_2 \leq \epsilon$, we stop the algorithm. Otherwise we advance k to $k+1$ and go to Step 1.

There is lack of theory to justify why the above OGA is convergent in the original paper [Petukhov'06] and in the literature so far. We now present an analysis of the convergence of the above OGA.

Theorem 4.2 *Suppose that Φ of size $n \times N$ has the RIP for order k with $1 \leq k \leq n$. Then the above OGA converges.*

Proof. Without loss of generality we may assume that $\Lambda_m = \{1, 2, \cdots, n_m\}$ for some $n_m < n$, where $m = 1, 2, \cdots$. Let

$$G_m = [\langle \phi_i, \phi_j \rangle]_{1 \leq i,j \leq n_m}$$

be the Grammian matrix. Define

$$a_m \leq \|G_m\|_2 \leq b_m$$

to be the smallest and largest eigenvalues of symmetric G_m. The RIP of Φ for integer n_m implies that $a_m > 0$ for $m = 1, 2, \cdots, m_0$ with $n_{m_0} = n$.

We first observe that the best approximation $f_{\Lambda_m} = \Phi_m^{-1} \left[\langle f, \phi_i \rangle \right]^T_{1 \le i \le n_m}$.

Then due to the result in Theorem 4.1, let us for simplicity, assume that f_{Λ_m} is the best approximation of $R_{m-1}(f)$. We next note that for $i \in \Lambda_m \backslash \Lambda_{m-1}$,

$$|\langle R_{m-1}(f), \phi_i/\|\phi_i\| \rangle| \ge r M_m$$

with

$$\begin{aligned} M_m &= \max_{i \notin \Lambda_{m-1}} |\langle R_{m-1}(f), \phi_i/\|\phi_i\| \rangle| = \max_{i=1,\cdots,n} |\langle R_{m-1}(f), \phi_i/\|\phi_i\| \rangle| \\ &\ge |\sum_{i=1}^n \alpha_i \langle R_{m-1}(f), \phi_i \rangle| \end{aligned}$$

for any α_i such that $\sum_{i=1}^n |\alpha_i| \le 1$. Assume that $f = \sum_{i=1}^n c_i \phi_i$ with $\sum_{i=1}^n |c_i| \le 1$ (with appropriate normalization). It follows that

$$M_m \ge |\langle R_{m-1}(f), f \rangle| = \|R_{m-1}(f)\|^2.$$

Hence we have

$$\|R_m(f)\|^2 = \langle R_{m-1}(f) - f_{\Lambda_m}, R_{m-1}(f) - f_{\Lambda_m} \rangle = \|R_{m-1}(f)\|^2 - \|f_{\Lambda_m}\|^2$$

and

$$\begin{aligned} \|f_{\Lambda_m}\|^2 &= \|\Phi_m^{-1} \left[\langle R_{m-1}(f), \phi_i \right]^T_{i=1,\cdots,n_m} \|^2 \\ &\ge \frac{1}{a_m^2} \| \left[\langle R_{m-1}(f), \phi_i \right]^T_{i=1,\cdots,n_m} \|^2 \\ &\ge \frac{1}{a_m^2} r^2 n_m^2 \|R_{m-1}(f)\|^2. \end{aligned}$$

That is,

$$\|R_m(f)\|^2 = \|R_{m-1}(f)\|^2 - \|f_{\Lambda_m}\|^2 \le \|R_{m-1}(f)\|^2 - \frac{1}{a_m^2} r^2 n_m^2 \|R_{m-1}(f)\|^2.$$

Summing the above inequality over $m = 1, \cdots, k$, we get

$$\|R_k(f)\|^2 \le \|R_0(f)\|^2 - r^2 \sum_{m=1}^k \frac{1}{a_m} n_m^2 \|R_{m-1}(f)\|^2 \le \|f\|^2 - r^2 \sum_{m=1}^k \frac{1}{a_m} n_m^2 \|R_k(f)\|^2.$$

because of the monotonicity of $\|R_m(f)\|$. In other words,

$$(r^2 \sum_{m=1}^{k} \frac{1}{a_m} n_m^2 + 1)\|R_k(f)\|^2 \leq \|f\|^2.$$

As $\sum_{m=1}^{k} n_m^2$ diverges and a_m nonincreases, $\|R_k(f)\|$ has to converge to zero. This completes a proof of the convergence of this OGA. ■

The OGA can be used to solve our research problem Eq. (1). For $\mathbf{y} \in \mathcal{R}_k$, the OGA algorithm uses the indices which are associated with the terms $|\langle \mathbf{y}, \phi_i \rangle|, i \in \{1, 2, \cdots, n\}$ which is $\geq r\%$ of the largest value. As the size of Λ_i increases, it finds an approximation $\mathbf{x}_{OGA,\epsilon}$ such that $\Phi \mathbf{x}_{OGA,\epsilon}$ is closed to $\mathbf{y}$ within the given ϵ. That is, $\|\Phi \mathbf{x}_{OGA,\epsilon} - \mathbf{y}\| \leq \epsilon$.

We now explain why $\mathbf{x}_{OGA,\epsilon}$ is a good approximation of $\mathbf{x}$. Due to the construction, the number of nonzero entries $\|\mathbf{x}_{OGA,\epsilon}\|_0 = k^* << n$. Similar to the RIP, let $\alpha_k, \beta_k \geq 0$ be the best constants in the inequalities

$$\alpha_k \|\mathbf{z}\|_2 \leq \|\Phi \mathbf{z}\|_2 \leq \beta_k \|\mathbf{z}\|_2, \qquad \text{for all } \|\mathbf{z}\|_0 \leq k.$$

Then since $\|\mathbf{x} - \mathbf{x}_{OGA,\epsilon}\|_0 \leq k + k^*$,

$$\|\mathbf{x} - \mathbf{x}_{OGA,\epsilon}\|_2 \alpha_{k+k^*} \leq \|\Phi(\mathbf{x} - \mathbf{x}_{OGA,\epsilon})\|_2 = \|\mathbf{y} - \Phi \mathbf{x}_{OGA,\epsilon}\|_2 \leq \epsilon.$$

That is, we have

$$\|\mathbf{x} - \mathbf{x}_{OGA,\epsilon}\|_2 \leq \frac{\epsilon}{\alpha_{k+k^*}}.$$

In particular, when $k^* \leq k$, all we need is to assume that $\alpha_{2k} > 0$ and hence $\mathbf{x}_{OGA,\epsilon}$ is away from $\mathbf{x}$ by ϵ/α_{2k}.

Next we need to show that $k^* \leq k$ may happen. Assume that each column of Φ is normalized. For $\mathbf{y} = \Phi \mathbf{x}$ with $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$, without loss of generality, we may assume that the support of $\mathbf{x}$ is $S = \{1, 2, \cdots, k\}$, $|x_k| = \min\{|x_j| \neq 0, i = 1, \cdots, n\}$, and $|x_1| = \|\mathbf{x}\|_\infty$.

Suppose that

$$k \leq \frac{1}{2M} + \frac{1}{2}, \tag{18}$$

where $M = M(\Phi)$ stands for the mutual coherence of Φ. Then we can claim that the support$(\mathbf{x}_{OGA,\epsilon}) \subset$ support$(\mathbf{x})$. Recall $\Phi = [\phi_1, \cdots, \phi_n]$ with ϕ_i

being the $i th$ column of Φ. Let us first compute the inner products of $\mathbf{y}$ with ϕ_i's.

$$|\langle y, \phi_i\rangle| = |\langle \Phi \mathbf{x}, \phi_i\rangle| = |\sum_{j=1}^{k} \langle x_j \phi_j, \phi_i\rangle|.$$

and

$$|\sum_{j=1}^{k} \langle (x_j) \phi_j, \phi_i\rangle| \geq |\langle x_1 \phi_1, \phi_i\rangle| - \sum_{j=2}^{k} |\langle x_j \phi_j, \phi_i\rangle|.$$

In particular, we have

$$|\langle \mathbf{y}, \phi_1\rangle| \geq |x_1| - M(k-1)|x_2|.$$

and

$$|\langle \mathbf{y}, \phi_i\rangle| \leq |\sum_{j=1}^{k} \langle x_j \phi_j, \phi_i\rangle| \leq |x_1| k M.$$

By our assumption in Eq. (18), we have

$$|x_1| - M(k-1)|x_2| \geq |x_1| - M(k-1)|x_1| \geq |x_1| k M.$$

it follows that

$$|\langle \mathbf{y}, \phi_1\rangle| \geq |\langle \mathbf{y}, \phi_i\rangle|, \quad \forall i \geq 2.$$

That is, the largest inner product is $|\langle \mathbf{y}, \phi_1\rangle|$.

Furthermore, let us assume

$$k \leq \frac{1}{(1+r)M}\left(\frac{|x_k|}{|x_1|} + M\right) \text{ or } rk|x_1|M \leq |x_k| - M(k-1)|x_1|, \tag{19}$$

where r is the positive constant $r < 1$ employed in the OGA.

Then for $2 \leq j \leq k$, $|\langle \mathbf{y}, \phi_j\rangle| \geq |x_j| - M(k-1)|x_1| \geq |x_k| - M(k-1)|x_1|$ and $r|\langle \mathbf{y}, \phi_1\rangle| \leq rk|x_1|M$. It follows that $|\langle \mathbf{y}, \phi_j\rangle| \geq r|\langle \mathbf{y}, \phi_1\rangle|$. That is, the first greedy step in the above OGA picks up all the indices of the nonzero entries of $\mathbf{x}$.

In particular, when the nonzero entries of $\mathbf{x}$ are 1 in absolute value, the condition in Eq. (19) is simplified to

$$k \leq \frac{1}{(1+r)M}(1+M).$$

That is, if k satisfies Eq. (18), then k satisfies Eq. (19). Under the condition in (18) or the conditions in (18) and (19), the OGA picks all the entries $\phi_1, \cdots, \phi_k$. Hence, the support($\mathbf{x}_{OGA,\epsilon}$) is the same as the support of $\mathbf{x}$.

Furthermore, $\|\mathbf{x}_{OGA,\epsilon} - \mathbf{x}\|_1 \leq \sqrt{k}\|\mathbf{x}^* - \mathbf{x}\|_2$. Since $\|\Phi(\mathbf{x}_{OGA,\epsilon} - \mathbf{x})\|_2 = \|\Phi\mathbf{x}_{OGA,\epsilon} - \mathbf{y}\|_2 \leq \epsilon$, we have

$$\begin{array}{rcl} \epsilon^2 & \geq & \|\Phi(\mathbf{x}_{OGA,\epsilon} - \mathbf{x})\|_2^2 \\ & = & \|\mathbf{x}_{OGA,\epsilon} - \mathbf{x}\|^2 + (\mathbf{x}_{OGA,\epsilon} - \mathbf{x})^T(G - I)(\mathbf{x}_{OGA,\epsilon} - \mathbf{x}) \\ & \geq & \|\mathbf{x}_{OGA,\epsilon} - \mathbf{x}\|_2^2 - M(\|\mathbf{x}_{OGA,\epsilon} - \mathbf{x}\|_1^2 - \|\mathbf{x}_{OGA,\epsilon} - \mathbf{x}\|_2^2) \\ & = & (1 + M)\|\mathbf{x}_{OGA,\epsilon} - \mathbf{x}\|_2^2 - Mk\|\mathbf{x}_{OGA,\epsilon} - \mathbf{x}\|_2^2. \end{array}$$

That is,

$$\|\mathbf{x}_{OGA,\epsilon} - \mathbf{x}\|_2^2 \leq \frac{\epsilon^2}{1 - M(k-1)}.$$

That is, under the assumption that the sparsity of $\mathbf{x}$ is small, i.e., Eq. (18), $\mathbf{x}_{OGA,\epsilon}$ approximates the sparse solution $\mathbf{x}$ very well.

4.1 L_1 Greedy Algorithm

Recently, Kozlov and Petukhov proposed a new greedy algorithm (cf. [Kozlov and Petukhov'08]). It is called L_1 Greedy Algorithm. The algorithm starts with the solution of the ℓ_1 minimization under the constraint $A\mathbf{z}_0 = \mathbf{y}$.

[1] Let $\mathbf{z}_0$ be the solution of the ℓ_1 minimization under the constraint $A\mathbf{z} = \mathbf{y}$ among $\mathbf{z} \in \mathbf{R}^n$.

[2] Let $M = \|\mathbf{z}^0\|_\infty$.

[3] For $i = 1, \cdots, N$, let $W \in \mathbf{R}^n$ be a weighted vector with 1 in the all entries except for those entries which are 1/10000 when $|\mathbf{z}_j^{i-1}| \geq 0.8M$, $1 \leq j \leq n$.

[4] Solve the weighted ℓ_1 minimization problem

$$\min\{\sum_{j=1}^{n} |z_i|/w_i, A\mathbf{z} = \mathbf{y}, \mathbf{z} \in \mathbf{R}^n\}$$

and let $\mathbf{z}^i$ be the solution.

[5] If $\mathbf{z}^i$ is not yet a sparse solution, let $M = 0.8M$ and return to Step 3.

The algorithm works well for random matrix A of size 512×1024. It is interesting to give an analysis of the convergence or reasons why the algorithm works.

5 The ℓ_q Approach

Let

$$\|\mathbf{x}\|_q = (\sum_{i=1}^{n} |x_i|^q)^{1/q}$$

be the standard ℓ^q quasi-norm for $0 < q < 1$. It is easy to see that $\lim_{q \to 0+} \|\mathbf{x}\|_q^q = \|\mathbf{x}\|_0$. We can use $\|\mathbf{x}\|_q^q$ to approximate $\|\mathbf{x}\|_0$. Thus, we consider the following minimization

$$\min\{\|\mathbf{x}\|_q^q, \quad \mathbf{x} \in \mathbb{R}^n, \Phi\mathbf{x} = \mathbf{y}\}. \tag{20}$$

for $0 < q \leq 1$ as an approximation of the original research problem Eq. (1). A solution of the above minimization is denoted by $\Delta_q \Phi \mathbf{x}$.

5.1 Recent Results on the ℓ_q Approach

The several ℓ_q methods were studied recently in [Chartrand'07], [Foucart and Lai'08], [Davies and Gribonval'08] and [R. Saab and Ö. Yilmaz'08]. The first piece of results is shown in [Chartrand'07]

Theorem 5.1 *Let $q \in (0, 1]$. Suppose that there exists a $k > 1$ such that the matrix Φ has RIP constant such that*

$$\delta_{ks} + k^{2/q-1}\delta_{(k+1)s} < k^{2/q-1} - 1.$$

Then the solution of Eq. (20) is the sparest solution.

One can see that this result is a generalization of Theorem 3.10. When $q = 1$ and $k = 3$, the above condition is the condition in Theorem 3.10. In fact, the proof is a generalization of the proof in [Candés, Romberg, and Tao'06] for ℓ_1 norm to ℓ_q quasi-norm. In [Foucart and Lai'08], we felt that the non-homogeneity of the Restricted Isometry Property (13) contradicted the consistency of the problem with respect to measurement amplification, or in other words, that it was in conflict with the equivalence of all the linear systems $(c\,A)\mathbf{z} = c\,\mathbf{y}$, $c \in \mathbb{R}$. Instead, we introduce $\alpha_k, \beta_k \geq 0$ to be the best constants in the inequalities

$$\alpha_k \|\mathbf{z}\|_2 \leq \|A\mathbf{z}\|_2 \leq \beta_k \|\mathbf{z}\|_2, \qquad \|\mathbf{z}\|_0 \leq k.$$

Our results are to be stated in terms of a quantity invariant under the change $A \leftarrow c\,A$, namely

$$\gamma_{2s} := \frac{{\beta_{2s}}^2}{{\alpha_{2s}}^2} \geq 1.$$

In fact, $\alpha_k^2 = 1 - \delta_k$ and $\beta_k^2 = 1 + \delta_k$. We use this slightly modified version of RIP and work through the arguments of [Candes, Romberg, Tao'06] in terms of quasi-norm ℓ_q to get the following theorem.

Our main result in this section is the following (see [Foucart and Lai'08] for a proof)

Theorem 5.2 *Given $0 < q \leq 1$, if*

$$\gamma_{2t} - 1 \; < \; 4(\sqrt{2} - 1) \left(\frac{t}{s}\right)^{1/q - 1/2} \qquad \textit{for some integer } t \geq s, \tag{21}$$

then every s-sparse vector is exactly recovered by solving Eq. (20).

Corollary 5.3 *Under the assumption that*

$$\gamma_{2s} < 4\sqrt{2} - 3 \approx 2.6569, \tag{22}$$

every s-sparse vector is exactly recovered by solving (7).

When $q = 1$, this result slightly improves Candès' condition in Theorem 3.9, since the constant γ_{2s} is expressed in terms of the Restricted Isometry Constant δ_{2s} as

$$\gamma_{2s} = \frac{1 + \delta_{2s}}{1 - \delta_{2s}},$$

hence the condition (22) becomes $\delta_{2s} < 2(3 - \sqrt{2})/7 \approx 0.4531$.

The second special instance we are pointing out corresponds to the choice $t = s + 1$. In this case, Condition (21) reads

$$\gamma_{2s+2} < 1 + 4(\sqrt{2} - 1) \left(1 + \frac{1}{s}\right)^{1/q - 1/2}.$$

The right-hand side of this inequality tends to infinity as q approaches zero. The following result is then straightforward.

Corollary 5.4 *Under the assumption that*

$$\gamma_{2s+2} < +\infty,$$

every s-sparse vector is exactly recovered by solving (20) for some $q > 0$ *small enough.*

The key point is to show for any $\mathbf{v}$ which is in the null space of Φ, i.e, $\Phi\mathbf{v} = 0$, $\|\mathbf{v}_S\|_q < \|\mathbf{v}_{\bar{S}}\|_q$ unless $\mathbf{v} = 0$, where S stands for the index set of the nonzero entries of the solution $\mathbf{x} \in \mathcal{R}_k$, $\mathbf{v}_S$ denotes the vector $\mathbf{v}$ restricted in S with other entries being zero and $\bar{S}$ is the complement indices of S.

This is indeed the case since for $\mathbf{v} = \mathbf{x} - \mathbf{x}^*$ in the null space of Φ, $\|\mathbf{v}_{\bar{S}}\|_q \leq \|\mathbf{v}_S\|_q$, where $\mathbf{x}^*$ is the solution of Eq. (20) and $\mathbf{x}$ is the sparse vector supported on S satisfying $\Phi\mathbf{x} = \mathbf{y}$. Combining the above inequality, we have a contradiction that $\|\mathbf{v}_S\|_q < \|\mathbf{v}_S\|_q$ unless $\mathbf{v} = 0$. Thus, the solution of the minimization is the exact solution if we can show $\|\mathbf{v}_S\|_q < \|\mathbf{v}_{\bar{S}}\|_q$. This inequality was recognized in [Grinoval and Nielson'03]. The condition (21) in Theorem 5.2 implies this inequality.

5.2 More about the ℓ_q Approach

We first show that the minimization problem Eq. (20) has a solution for $q > 0$. That is, the existence of the solution is independent of the RIP of Φ. See [Foucart and Lai'08] for a proof.

Theorem 5.5 *Fix* $0 < q < 1$. *There exists a solution* $\Delta_q A\mathbf{x}$ *solving Eq. (20).*

We next consider the situation that the measurements $\mathbf{y}$ are imperfect. That is, $\mathbf{y} = \Phi\mathbf{x}_0 + \mathbf{e}$ with unknown perturbation $\mathbf{e}$ which is bounded by a known amount $\|\mathbf{e}\|_2 \leq \theta$. In this case we consider the following

$$\min\{\|\mathbf{x}\|_q^q, \quad \mathbf{x} \in \mathbb{R}^n, \|\Phi\mathbf{x} - \mathbf{y}\|_2 \leq \theta\}. \tag{23}$$

A solution of the above minimization is denoted by $\Delta_{q,\theta}\Phi\mathbf{x}$. As in the previous section, we have

Theorem 5.6 *Fix* $0 < q < 1$ *and* $\theta > 0$. *There exists a solution* $\Delta_{q,\theta}\Phi\mathbf{x}$ *solving Eq. (23).*

In [Saab and Yilmaz'08], they extended the proof in [Candes'08] in the ℓ_q setting. They have

Theorem 5.7 *Let $q \in (0,1]$. Suppose that $\delta_{ks} + k^{2/q-1}\delta_{(k+1)s} < k^{2/q-1} - 1$ for some $k > 1$ with $kS \in \mathbf{Z}_+$. Let $\mathbf{x}^*$ be the solution of Eq. (23). Then*

$$\|\mathbf{x} - \mathbf{x}^*\|_2^q \leq C_1 \eta^p + \frac{C_2}{s^{1-q/2}} \Delta_s(\mathbf{x})_q^q$$

for two positive constants C_1 and C_2.

Here, the quantity $\Delta_k(\mathbf{x})_q$ denotes the error of best k-term approximation to $\mathbf{x}$ with respect to the ℓ_q-quasinorm, that is

$$\Delta_k(\mathbf{x})_q := \inf_{\|\mathbf{z}\|_0 \leq k} \|\mathbf{x} - \mathbf{z}\|_q.$$

The above theorem is an extension of Chartrand's result (cf. Theorem 5.1). Next we state another main theoretical result of this survey. We refer to [Foucart and Lai'08] for a proof.

Theorem 5.8 *Given $0 < q \leq 1$, if Condition (21) holds, i.e. if*

$$\gamma_{2t} - 1 \; < \; 4(\sqrt{2} - 1) \left(\frac{t}{s}\right)^{1/q-1/2} \qquad \text{for some integer } t \geq s, \tag{24}$$

then a solution $\mathbf{x}^*$ of $(P_{q,\theta})$ approximate the original vector $\mathbf{x}$ with errors

$$\|\mathbf{x} - \mathbf{x}^*\|_q \;\leq\; C_1 \cdot \sigma_s(\mathbf{x})_q \; + D_1 \cdot s^{1/q-1/2} \cdot \theta, \tag{25}$$

$$\|\mathbf{x} - \mathbf{x}^*\|_2 \;\leq\; C_2 \cdot \frac{\sigma_s(\mathbf{x})_q}{t^{1/q-1/2}} + D_2 \cdot \theta. \tag{26}$$

The constants C_1, C_2, D_1, and D_2 depend only on q, γ_{2t}, and the ratio s/t.

Comparison of the results in Theorems 5.8 and 5.7 is given in [Saab and Yilmaz'08]. It concludes that when k is around 2, the sufficient condition (24) is weaker while the condition in Theorem 5.7 is weaker when $k > 2$. Numerical experiemental results in [Foucart and Lai'08] show that the ℓ_q method is able to 100% recovery all the sparse vectors with sparsity is about $s = m/2$. That is, in order to have $ks \leq m$, k is about 2.

Next we consider that negative results discussed in [Davies and Gribnoval'08]. That is, when the ℓ_q method may fail.

Theorem 5.9 *For any $\epsilon > 0$, there exists an integer s and dictionary Φ with a restricted isometry constant $\delta_{2s} \leq 1/\sqrt{2} + \epsilon$ for which ℓ_1 method fails on some k sparse vector.*

Now the gap between the positive result $\delta_{2s} = 2(3 - \sqrt{2})/7 = 0.4531$ and the negative result $\delta_{2s} = 1/\sqrt{2} + \epsilon = 0.7071$ is about 0.2540. In general, Davies and Gribnoval consider a special matrix Φ which has a unit spectral norm, i.e.,

$$\|\Phi\|_2 = \sup_{\mathbf{y} \neq 0} \frac{\|\Phi \mathbf{y}\|_2}{\|\mathbf{y}\|_2} = 1.$$

Then they define

$$\sigma_k^2(\Phi) := \min_{\substack{y \in \\ \|\mathbf{y}\|_0 \leq k}} \frac{\|\Phi \mathbf{y}\|_2}{\|\mathbf{y}\|_2}$$

which is equal to α_k^2 in [Foucart and Lai'08].

Theorem 5.10 *Fix $0 < q \leq 1$ and let $0 < \eta_q < 1$ be the unique positive solution to*

$$\eta_q^{2/q} + 1 = \frac{2}{p}(1 - \eta_p).$$

For any $\epsilon > 0$, there exist integers $s \geq 1$, $N \geq 2s + 1$ and a minimally redundant unit spectral norm tight frame $\Phi_{N-1 \times N}$ with

$$\sigma_{2s}^2(\Phi) \geq 1 - \frac{2}{2-q}\eta_q - \epsilon$$

for which there exists an s-sparse vector which cannot be uniquely recovered by the ℓ_q method.

5.3 Numerical Computation of the ℓ_q Approach

The minimization problem (P$_q$) suggested to recover $\mathbf{x}$ is nonconvex, Following [Foucart and Lai'08], we introduce an algorithm to compute a minimizer of the approximated problem, for which we give an informal but detailed justification.

We shall proceed iteratively, starting from a vector $\mathbf{z_0}$ satisfying $A\mathbf{z_0} = \mathbf{y}$, which is a reasonable guess for $\mathbf{x}$, and constructing a sequence $(\mathbf{z}_n)$ recursively by defining $\mathbf{z}_{n+1}$ as a solution of the minimization problem

$$\underset{\mathbf{z} \in \mathbb{R}^N}{\text{minimize}} \sum_{i=1}^{N} \frac{|z_i|}{(|z_{n,i}| + \epsilon_n)^{1-q}} \qquad \text{subject to} \quad A\mathbf{z} = \mathbf{y}. \tag{27}$$

Here, the sequence (ϵ_n) is a nonincreasing sequence of positive numbers. It might be prescribed from the start or defined during the iterative process. In practice, we will take $\lim_{n\to\infty} \epsilon_n = 0$. We shall now concentrate on convergence issues. We start with the following

Proposition 5.11 *For any nonincreasing sequence (ϵ_n) of positive numbers and for any initial vector $\mathbf{z}_0$ satisfying $A\mathbf{z}_0 = \mathbf{y}$, the sequence $(\mathbf{z}_n)$ defined by (27) admits a convergent subsequence.*

Similar to the proof of Theorem 5.5, we can see that the solution of the above minimization exists. We further show that the solution $\mathbf{x}_\epsilon$ of Eq.(27) will converge to the solution Eq. (20). For convenience, let $\widehat{\mathbf{x}}$ be a solution of Eq. (20).

Theorem 5.12 *Fix $0 < q \le 1$. Let $\mathbf{x}_\epsilon$ be the solution of Eq. (27). Then $\mathbf{x}_\epsilon$ converges to $\widehat{\mathbf{x}}$ as $\epsilon \to 0_+$.*

The new minimization problem Eq. (27) can be solved using ℓ_1 method since $F_{q,\epsilon}(\mathbf{x})$ is a weighted ℓ_1 norm.

Proposition 5.13 *Given $0 < q < 1$ and the original s-sparse vector $\mathbf{x}$, there exists $\eta > 0$ such that, if*

$$\epsilon_n < \eta \qquad \textit{and} \qquad \|\mathbf{z}_n - \mathbf{x}\|_\infty < \eta \qquad \textit{for some } n, \tag{28}$$

then the algorithm 27 produces the exact solution. That is,

$$\mathbf{z}_k = \mathbf{x} \qquad \textit{for all } k > n.$$

The constant η depends only on q, $\mathbf{x}$, and γ_{2s}.

Lemma 5.14 *Given $0 < q \le 1$ and an s-sparse vector $\mathbf{x}$, if Condition (21) holds, i.e. if*

$$\gamma_{2t} - 1 \;<\; 4(\sqrt{2}-1)\left(\frac{t}{s}\right)^{1/q-1/2} \qquad \textit{for some integer } t \ge s,$$

then for any vector $\mathbf{z}$ satisfying $A\mathbf{z} = \mathbf{y}$, one has

$$\|\mathbf{z}-\mathbf{x}\|_q^q \le C\left[\|\mathbf{z}\|_q^q - \|\mathbf{x}\|_q^q\right],$$

for some constant C depending only on q, γ_{2t}, and the ratio s/t.

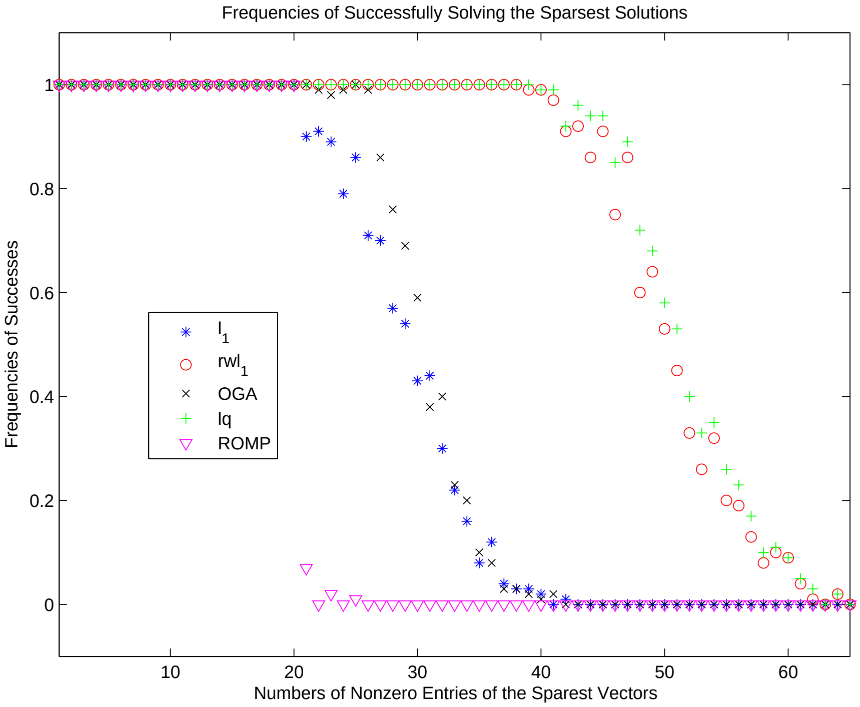

Figure 1: Comparison of ℓ_1, ℓ_q, and OGA methods for sparest solutions

Finally we present the following

Proposition 5.15 *Given $0 < q < 1$ and the original s-sparse vector $\mathbf{x}$, if Condition (21) holds, i.e. if*

$$\gamma_{2t} - 1 \ < \ 4(\sqrt{2} - 1)\left(\frac{t}{s}\right)^{1/q-1/2} \qquad \textit{for some integer } t \geq s,$$

then there exists $\zeta > 0$ such that, for any nonnegative ϵ less than ζ, the vector $\mathbf{x}$ is exactly recovered by solving (27). The constant ζ depends only on N, q, $\mathbf{x}$, γ_{2t}, and the ratio s/t.

Numerical results show that our ℓ_q approximation method works well. In Figure 1, we present the frequencies of the exact recovery using various methods for Gaussian random matrix of size 128×512 for various sparse vectors. For each sparsity, we randomly generate the Gaussian random matrix Φ and a vector **x** with the given sparsity and tested various methods to solve the **x** for 100 times. The number of exact recovery by each method is divided by 100 to obtain the frequency for the method.

References

[1] Baraniuk, R., M. Davenport, R. DeVore, and M. B. Wakin, A simple proof of the restricted isometry property for random matrices, Constructive Approximation, to appear, 2008.

[2] Buldygin, V. V. and Yu, V. Kozachenko, Metric Characterization of Random Variables and Random Processes, AMS Publication, Providence, 2000.

[3] Candés, E. J. Compressive sampling. International Congress of Mathematicians. Vol. III, 1433–1452, Eur. Math. Soc., Z"urich, 2006.

[4] Candés, E. J., J. K. Romberg, Quantitative robust uncertainty principles and optimally sparse decompositions. Found. Comput. Math. 6 (2006), 2, 227–254.

[5] Candés, E. J., J. K. Romberg, J. K. and T. Tao, Stable signal recovery from incomplete and inaccurate measurements. Comm. Pure Appl. Math. 59 (2006), 1207–1223.

[6] Candés, E. J. and T. Tao, Decoding by linear programming, IEEE Trans. Inform. Theory 51 (2005), no. 12, 4203–4215.

[7] Candés, E. J. and T. Tao, Near-optimal signal recovery from random projections: universal encoding strategies, IEEE Trans. Inform. Theory 52 (2006), no. 12, 5406–5425.

[8] Candés, E. J., M. Watkin, and S. Boyd, Enhancing Sparsity by Reweighted l_1 Minimization, manuscript, 2007.

[9] R. Chartrand, Nonconvex compressed sensing and error correction, in International Conference on Acoustics, Speech, and Signal Processing, IEEE, 2007.

[10] R. Chartrand, Exact reconstruction of sparse signals via nonconvex minimization, IEEE Signal Process. Letters, 14(2007), 707–710.

[11] R. Chartrand and V. Staneva, Restricted isometry properties and nonconvex compressive sensing, Inverse Problem, to appear, 2008.

[12] M. Davies and Rémi Gribonval, Restricted Isometry constants where ℓ_q sparse recovery can fail for $0 < q \leq 1$, manuscript, 2008.

[13] Donoho, D. L., Compressed sensing, IEEE Trans. Inform. Theory 52 (2006), 1289–1306.

[14] Donoho, D. L., Sparse components of images and optimal atomic decompositions. Constr. Approx. 17 (2001), 353–382.

[15] D. L. Donoho, Unconditional bases are optimal bases for data compression and for statistical estimation, Appl. Comput. Harmonic Anal., 1(1993), 100–115.

[16] D. L. Donoho, Unconditional bases and bit-level compression, Appl. Comput. Harmonic Anal., 3(1996), pp. 388–92.

[17] Donoho, D. L. and M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization, Proc. Natl. Acad. Sci. USA 100 (2003), no. 5, 2197–2202.

[18] Donoho, D. L., M. Elad, and V. N. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inform. Theory, 52 (2006), 6–18.

[19] Donoho, D. L. and J. Tanner, Sparse nonnegative solution of underdetermined linear equations by linear programming. Proc. Natl. Acad. Sci. USA 102 (2005), no. 27, 9446–9451.

[20] Elad, M. and A. M. Bruckstein, IEEE Trans. Inf. Theory 48(2002), 2558–2567.

[21] S. Foucart and M. J. Lai, Sparsest Solutions of Underdetermined Linear Systems via ℓ_q minimization for $0 < q \leq 1$, to appear in Applied Comput. Harmonic Analysis, 2009.

[22] Geman, S., A limit theorem for the norm of random matrices, Ann. Prob., 8(1980), 252–261.

[23] Gribnoval, R and M. Nielsen, Sparse decompositions in unions of bases, IEEE Trans. Info. Theory, 49(2003), 3320–3325.

[24] I. Kozlov and A. Petukhov, ℓ_1 greedy algorithm for sparse solutions of underdetermined linear system, manuscript, 2008.

[25] M. J. Lai, Restricted isometry property for sub-Gaussian random matrices, unpublished manuscript, 2008.

[26] M. J. Lai and P. Wenston, L1 Spline Methods for Scattered Data Interpolation and Approximation , Advances in Computational Mathematics 21(2004), 293–315.

[27] Ledoux, M., *The concentration of measure phenomenon*, AMS Publication, Providence, 2001.

[28] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, Uniform uncertainty principle for Bernoulli and subgaussian ensembles, Constructive Approx. 28(2008) 277–289.

[29] S. Mendelson, A. Pajor, N. Tomczack-Jaegermann, Reconstruction and subgaussian operators in asymptotic geometric analysis, Geometric and Functional Analysis 17(2007), 1248-1272.

[30] Natarajan, B. K., Sparse approximate solutions to linear systems, SIAM J. Comput., vol. 24, pp. 227234, 1995.

[31] Needell, D. and R. Vershynin, Uniform uncertainty principal and signal recovery via regularized orthogonal matching pursuit, manuscript, 2007.

[32] Petukhov, A., Fast implementation of orthogonal greedy algorithm for tight wavelet frames, Signal Processing, 86(2006), 471–479.

[33] G. Pisier, Probabilistic Methods in the Geometry of Banach Spaces, Springer Verlag, Lecture Notes in Mathematics, No. 1206, 1986.

[34] R. Saab and Ö. Yilmaz, Sparse recovery by nonconvex optimization – instant optimality, manuscript, 2008.

[35] Temlyakov, V. N., Weak greedy algorithms, Adv. Comput. Math. 12 (2000), 213–227.

[36] Temlyakov, V. N., Nonlinear methods of approximation, Foundations of Comp. Math., 3 (2003), 33–107.

[37] Tropp, J. A., Greed is good: Algorithmic results for sparse approximation, IEEE Trans. Inf. Theory, 50 (2004), 2231–2242.

[38] Wachter, K. W., The strong limits of random matrix spectra for sample matrices of independent elements, Ann. Prob., 6(1978), 1–18.

6 Appendix 1: Gaussian Random Matrices

Let $A = [a_{ij}]_{1\le i\le m, 1\le j\le n}$ be a rectangular matrix with a_{ij} being iid Gaussian random variables with mean zero and variance σ^2. Let $\mathbf{x} = (x_1, \cdots, x_n)^T \in \mathbf{R}^n$ be a vector. We use $\|\mathbf{x}\|_2$ denotes the norm of $\mathbf{x}$. Consider a random variable $X = (X_1, \cdots, X_m)^T$ with $X_i = (\sum_{j=1}^n a_{ij}x_j)^2, i = 1, \cdots, m$. Since $\mathbf{E}(a_{ij}) = 0$, we have $\mathbf{E}(X_i) = \sigma^2\|\mathbf{x}\|_2^2$ for all i. Let $\xi_i = X_i - \mathbf{E}(X_i)$ be a new random variable and let

$$S_m = \sum_{i=1}^m \xi_i$$

be the sum of these new independent random variables. It is easy to see that

$$S_m = \|A\mathbf{x}\|_2^2 - m\sigma^2\|\mathbf{x}\|_2^2.$$

In this section, we are interested in proving the following inequality.

Theorem 6.1 *For any $\epsilon > 0$, the probability*

$$\mathcal{P}(|\|A\mathbf{x}\|_2^2 - m\sigma^2\|\mathbf{x}\|_2^2| < \epsilon\|\mathbf{x}\|_2^2) \ge 1 - 2\exp(-\frac{\epsilon^2 m}{(m\sigma^2)(c\epsilon + 2m\sigma^2)}), \quad (29)$$

where c is a positive constant independent of ϵ and $\|\mathbf{x}\|_2$.

We plan to use the Bernstein inequality (cf. [Buldygin andKozachenko'00, p.27]) to prove this result. For convenience, we state the inequality below.

Theorem 6.2 *Suppose that $\xi_i, 1 \le i \le m$ are independent random variables with $\mathbf{E}(\xi_i) = 0$ and $\mathbf{E}(\xi_i^2) = \nu_i^2 < \infty, 1 \le i \le m$. Let $S_m = \sum_{i=1}^m \xi_i$. Moreover, suppose that there exists a constant $H > 0$ such that*

$$|\mathbf{E}(\xi_i^k)| \le \frac{m!}{2}\nu_i^2 H^{k-2} \tag{30}$$

for all integer $k > 1$ and all $i = 1, \cdots, m$. Then the following inequality holds for all $t > 0$: the probability

$$\mathcal{P}(|S_m| > t) \le \exp\left\{-\frac{t^2}{2(tH + \sum_{i=1}^m \nu_i^2)}\right\}.$$

Proof. (The proof of Theorem 6.1.) We need to study Eq. (30) for $\xi_i = X_i - \mathbf{E}(X_i)$ for $k \ge 3$ since for $k = 2$, Eq. (30) is satisfied trivially.

For convenience, let $\mu = \mathbf{E}(X_i) = \sigma^2\|\mathbf{x}\|_2^2$. It is easy to see $\mathbf{E}(|\xi_i|^2) = 2\mu^2$. Thus, $\nu_i^2 = 2\mu^2$. For $k \ge 3$, we have

$$\mathbf{E}(|\xi_i|^k) = \mathbf{E}((X_i - \mu)^k) = \sum_{j=0}^k \binom{k}{j}\mathbf{E}(X_i^j)(-1)^{k-j}\mu^{k-j}.$$

Let us spend some effort to compute $\mathbf{E}(X_i^j)$. We have

$$\mathbf{E}(X_i^j) = \mathbf{E}(\sum_{j=1}^n a_{ij}x_j)^{2j} = \sum_{j_1+\cdots+j_n=2j} \frac{(2j)!}{j_1!\cdots j_n!}\mathbf{E}(a_{i,1}^{j_1}a_{i,2}^{j_2}\cdots a_{i,n}^{j_n})x_1^{j_1}\cdots x_n^{j_n}.$$

Note that $\mathbf{E}(a_{ij}^\ell) = 0$ for all odd integers ℓ and it is known (using integration by parts) that $\mathbf{E}(a_{ij}^\ell) = \dfrac{\ell!}{2^{\ell/2}(\ell/2)!}\sigma^\ell$ for even integers ℓ. Since a_{ij} are iid random variables, we have

$$\begin{aligned}
\mathbf{E}(X_i^j) &= \sum_{2j_1+\cdots+2j_n=2j} \frac{(2j)!}{(2j_1)!\cdots(2j_n)!}\mathbf{E}(a_{i,1}^{2j_1}a_{i,2}^{2j_2}\cdots a_{i,n}^{2j_n})(x_1)^{2j_1}\cdots(x_n)^{2j_n} \\
&= \frac{(2j)!}{j!}\sum_{j_1+\cdots+j_n=j} \frac{j!}{(2j_1)!\cdots(2j_n)!}\frac{(2j_1)!\cdots(2j_n)!}{2^j j_1!\cdots j_n!}\sigma^{2j}(x_1)^{2j_1}\cdots x_n^{2j_n} \\
&= \frac{(2j)!\sigma^{2j}}{2^j j!}(\sum_{j=1}^n x_j^2)^j
\end{aligned}$$

$$= \frac{(2j)!}{2^j j!}\sigma^{2j}\|\mathbf{x}\|_2^{2j} = \frac{(2j)!}{2^j j!}\mu^j.$$

Thus,

$$\mathbf{E}(|\xi_i|^k) \le \sum_{j=0}^{k}\binom{k}{j}\frac{(2j)!}{2^j j!}\mu^j\mu^{k-j}.$$

By using Stirling's formula, we have $\frac{(2j)!}{2^j j!} \le 2^j j!/2 \le 2^j k!/2$ and hence,

$$\begin{aligned}
|\mathbf{E}(|\xi_i|^k)| &\le \sum_{j=0}^{k}\binom{k}{j}\frac{2^j k!}{2}\mu^k \\
&= \frac{k!}{2}3^k\mu^k = \frac{k!}{2}2\sigma^4\|\mathbf{x}\|_2^4\frac{9}{2}3^{k-2}(\sigma^2\|\mathbf{x}\|_2^2)^{k-2} \\
&\le \frac{k!}{2}2\sigma^4\|\mathbf{x}\|_2^4 H^{k-2}
\end{aligned}$$

with $H = 13.5\sigma^2\|\mathbf{x}\|_2^2$. That is, Eq. (30) is satisfied for $k \ge 3$. By Theorem6.2, we have

$$P(|\|A\mathbf{x}\|_2^2 - m\sigma^2\|\mathbf{x}\|_2^2| > t) \le 2\exp\left\{-\frac{t^2}{2(t13.5\mu + 2m\mu^2)}\right\}. \tag{31}$$

Choosing $t = \epsilon\|\mathbf{x}\|_2^2$, we have $t13.5\mu + 2m\mu^2 = \sigma^2\|\mathbf{x}\|_2^4(13.5\epsilon + 2m\sigma^2)$ and the above probability yields:

$$P(|\|A\mathbf{x}\|_2^2 - m\sigma^2\|\mathbf{x}\|_2^2| > \epsilon\|\mathbf{x}\|_2^2) \le 2\exp\left\{-\frac{\epsilon^2 m}{2(m\sigma^2)(13.5\epsilon + 2m\sigma^2)}\right\}. \tag{32}$$

In other word, the desirable result of Theorem 6.1 is proved. ■

We remark that when $\sigma = 1/\sqrt{m}$, the estimate Eq. (29) gives a proof of Theorem 3.11. For this special case, we have

Theorem 6.3 *Suppose that ξ is a Gaussian random variable with mean zero and variance σ^2. Let A be an $m \times n$ matrix whose entries are iid copies of ξ. For any $\epsilon > 0$, the probability*

$$\mathcal{P}(|\|A\mathbf{x}\|_2^2 - m\sigma^2\|\mathbf{x}\|_2^2| < \epsilon n\sigma^2\|\mathbf{x}\|_2^2) \ge 1 - 2\exp\left\{-\frac{\epsilon^2 m}{c}\right\}, \tag{33}$$

where c is a positive constant independent of ϵ and $\|\mathbf{x}\|_2$.

Proof. (The Proof of Theorem 6.3.) For a random matrix A of size $m \times n$ with entries a_{ij} being iid Gaussian random variables with zero mean and variance σ^2, then we use $\epsilon m \sigma^2$ for ϵ in Eq. (32). Then we have

$$\mathcal{P}\big(\big|\|A\mathbf{x}\|_2^2 - m\sigma^2\|\mathbf{x}\|_2^2\big| > \epsilon m \sigma^2 \|\mathbf{x}\|_2^2\big) \leq 2\exp\left\{-\frac{\epsilon^2 m}{2(\epsilon 13.5 + 2)}\right\}. \qquad (34)$$

This completes a proof of Theorem 6.3. ■