Hierarchical Reconstruction of Sparse Signals

Ming Zhong[*]

Advisor: Dr. Eitan Tadmor[†]
AMSC and *CSCAMM*
University of Maryland College Park
College Park, Maryland 20742 USA

May 15, 2013

Abstract

We consider a family of constrained ℓ_p minimizations: $\min_{\mathbf{x}\in\mathbf{R}^n}\{||\mathbf{x}||_{\ell_p} \mid A\mathbf{x} = A\mathbf{x}_*\}$ to recover $\mathbf{x}_*$ with knowledge of only A, $A\mathbf{x}_*$ and $||\mathbf{x}_*||_{\ell_0}$. Such family of problems has been extensively studied in the Compressed Sensing Community and are used to recover sparse signals. We then go through the reasoning on why $p = 1$ is the most suitable choice. Due to the possible ill-posedness of the constrained ℓ_1 minimization, we introduce the Tikhonov Regularization to arrive at an unconstrained version, $\min_{\mathbf{x}\in\mathbf{R}^n}\{||\mathbf{x}||_{\ell_1} + \frac{\lambda}{2}||A\mathbf{x}_* - A\mathbf{x}||^2_{\ell_2}\}$. Through this regularized version, we construct the approximation based on a given scale λ. We also show that the two problems are equivalent when $\lambda \to \infty$. Using the idea of multi-scale approximation, we adopt the method of Hierarchical Decomposition from Image Processing to reconstruct sparse signals on layers of dyadic scales. We proceed to show that this Hierarchical Decomposition approach alleviate the dependence on the regularization parameter λ and can be used to de-noise corrupted signals. We are also able to show that the difference between the approximation from the Hierarchical Reconstruction, call it $\mathbf{x}_{HRSS}$ and $\mathbf{x}_*$ is in the Null(A); and various numerical examples support the fact that this approach offer a better approximation to the original $\mathbf{x}_*$.

1 Background: A Constrained Minimal ℓ_1-norm Problem

The ingenious Nyquist-Shannon Sampling Theorem addresses the question about the possibility to recover any signal using finite number of sampling (measurements); however, according to the theorem, the number of sampling to take is two times the highest frequency in a signal. The pursuit of improvement on reducing the sampling rates have been non-stop since 1949; and significant progress has been made especially regarding the sparse signal recovery. In fact, hundreds of papers in the Compressed Sensing community have been published to reduce the sampling rates to a significant level. The theory regarding sparse signal recovery has been

[*] mzhong1@umd.edu
[†] tadmor@cscamm.umd.edu

proven and perfected in a series of papers starting in mid 2004. See [4, 9, 13, 10, 7, 6, 5] for complete details.

Let us consider a target signal $\mathbf{x}_* \in \mathbf{R}^n$, with sparsity l, which simply means l non-zero entries in $\mathbf{x}_*$. let m represent the number of linear non-adaptive[1] measurements one wants to take: Let $\mathbf{a}_i \in \mathbf{R}^n$ be a basis vector. Measurements are made by taking the usual Euclidean inner product between $\mathbf{a}_i$ and $\mathbf{x}$, that is $\langle \mathbf{a}_i, \mathbf{x} \rangle_{\ell_2}$. Let A be the concatenation of the m (randomly picked) basis vectors $\mathbf{a}_i \in \mathbf{R}^n$, for $i = 1, \ldots, m$; then $A = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \cdots \\ \mathbf{a}_m^T \end{bmatrix} \in \mathbf{R}^{m \times n}$. Two possible scenarios exist for choosing the matrix A: it is either prescribed by a collection of basis vectors from a specific transformation or constructed with certain properties, namely the Restricted Isometry Properties (or other equivalent properties). The problem which we mentioned is basically asking that with the knowledge of only $A\mathbf{x}_*$ and $m \ll n$, can $\mathbf{x}_*$ be recovered? And if so, how? The answer in general is no. However when l is small, equivalently saying that $\mathbf{x}_*$ is sparse, then full recovery is highly possible according to the compressive sensing principle. Moreover the sparse signal $\mathbf{x}_*$ can be likely recovered through the following constrained minimal ℓ_p-norm problem:

$$\min_{\mathbf{x} \in \mathbf{R}^n} \{ ||\mathbf{x}||_{\ell_p} \mid A\mathbf{x} = A\mathbf{x}_* \} \tag{1}$$

The question remains: what would be an appropriate p for the purpose of recovering a sparse signal? As stated in [13], the sparsity of a vector $\mathbf{x}$ is usually defined as (for some $0 < p < 2$):

$$||\mathbf{x}||_{\ell_p} \equiv (\sum_i |x_i|^p)^{\frac{1}{p}} \tag{2}$$

(We will also consider the border line cases $p = 0$ and $p = 2$). And according to [11, 12], ℓ_p norm with $0 \leq p \leq 1$ are intuitive ways to preserve sparsity in the mathematical setting. Let us first consider an extreme case when $p = 0$, when the ℓ_0-norm simply measures the number of non-zero elements in a vector $\mathbf{x}$, directly imposing sparsity. Letting $p = 0$ in (1), we obtain:

$$\min_{\mathbf{x} \in \mathbf{R}^n} \{ ||\mathbf{x}||_{\ell_0} \mid A\mathbf{x} = A\mathbf{x}_* \} \tag{3}$$

The minimizer $\mathbf{x}_{\ell_0}$ will the same as $\mathbf{x}_*$ if we take $m = l + 1$ measurements[2]. However, it is shown in [22] that (3) is NP hard and requires techniques from combinatorial optimization. Hence it is numerically appropriate to consider a convex (1) by using either the ℓ_1 or ℓ_2-norm[3]. With a convex (1), we are able to use linear programming to find the minimizer. From now on, we will denote the measurement results $A\mathbf{x}_*$ as $\mathbf{b}$ and turn to the following Constrained Least Square Problem:

$$\min_{\mathbf{x} \in \mathbf{R}^n} \{ ||\mathbf{x}||_{\ell_2} \mid A\mathbf{x} = \mathbf{b} \} \tag{4}$$

It is also called a constrained minimum norm problem. An analytic solution exists and is $\mathbf{x}_{\ell_2} = A^T(AA^T)^{-1}\mathbf{b}$, since A is a fat matrix, AA^T has an inverse when A has linearly independent rows.[4] Apparently $A\mathbf{x}_{\ell_2} = AA^T(AA^T)^{-1}\mathbf{b} = \mathbf{b}$. Let $\mathbf{x}$ be another solution of the linear

[1] Means of measurements does not depend on $\mathbf{x}$.

[2] Bresler; Wakin et al

[3] We do not consider $0 < p < 1$, since such ℓ_p-norm is not convex.

[4] Otherwise $(AA^T)^{-1}$ is replaced by its pseudo-inverse (also known as the Moore-Penrose inverse) and is computed using SVD.

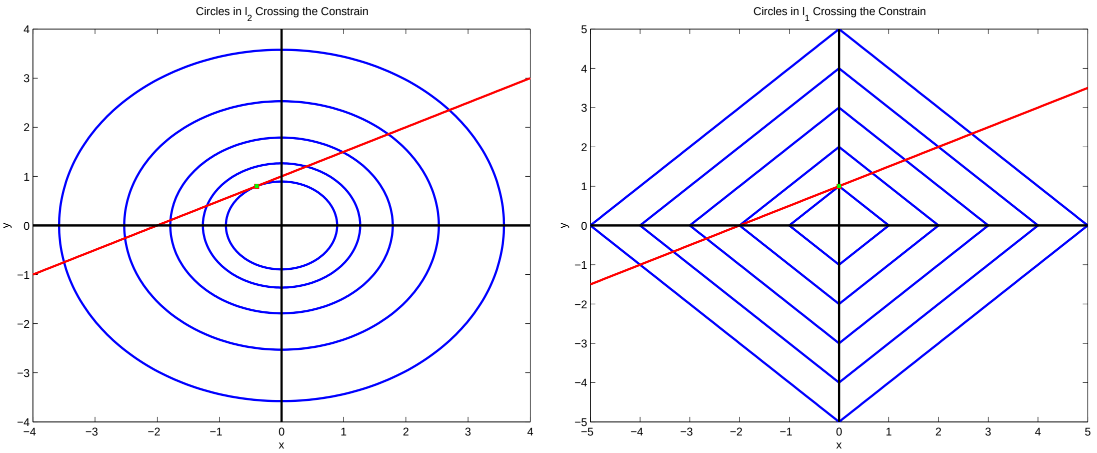
system, that is $A\mathbf{x} = \mathbf{b}$. Then we claim that $\mathbf{x} - \mathbf{x}_{\ell_2}$ is perpendicular to $\mathbf{x}_{\ell_2}$. To justify our claim, observe the following:

$$\begin{aligned}\langle \mathbf{x} - \mathbf{x}_{\ell_2}, \mathbf{x}_{\ell_2} \rangle_{\ell_2} &= \langle \mathbf{x} - \mathbf{x}_{\ell_2}, A^T(AA^T)^{-1}\mathbf{b} \rangle_{\ell_2} = \langle A(\mathbf{x} - \mathbf{x}_{\ell_2}), (AA^T)^{-1}\mathbf{b} \rangle_{\ell_2} \\ &= \langle \mathbf{0}, (AA^T)^{-1}\mathbf{b} \rangle_{\ell_2} = 0\end{aligned}$$

Hence by the Pythagoras' Theorem in $\mathbf{R}^n$:

$$||\mathbf{x}||_{\ell_2}^2 = ||\mathbf{x} - \mathbf{x}_{\ell_2} + \mathbf{x}_{\ell_2}||_{\ell_2}^2 = ||\mathbf{x} - \mathbf{x}_{\ell_2}||_{\ell_2}^2 + ||\mathbf{x}_{\ell_2}||_{\ell_2}^2 \geq ||\mathbf{x}_{\ell_2}||_{\ell_2}^2$$

Therefore $||\mathbf{x}||_{\ell_2} \geq ||\mathbf{x}_{\ell_2}||_{\ell_2}$, with equality realized only when $\mathbf{x} = \mathbf{x}_{\ell_2}$. $\mathbf{x}_{\ell_2}$ is indeed the optimizer. Unfortunately, in most cases, $\mathbf{x}_{\ell_2}$ is not going to be sparse, despite the fact that $\mathbf{x}_{\ell_2}$ is relatively easy to compute. Let us explain why $\mathbf{x}_{\ell_2}$ is dense in the two dimensional setting: the solution set of $A\mathbf{x} = \mathbf{b}$ would consist of a straight line, and the level sets $\mathcal{B}(r) = \{\mathbf{x} \in \mathbf{R}^n \mid ||\mathbf{x}||_{\ell_2} \leq r\}$ are circles of radius r centered at the origin. The constrained least square problem is asking that out of all the possible points on the said line $A\mathbf{x} = \mathbf{b}$, which one would be on the level set $\mathcal{B}(r)$ with the smallest r. As it is shown below, most of the time, the intersection happens in the interior of a quadrant, thus making $\mathbf{x}_{\ell_2}$ not sparse. However if the ℓ_1-norm is used, since the level sets $\mathcal{D}(r) = \{\mathbf{x} \in \mathbf{R}^n \mid ||\mathbf{x}||_{\ell_1} \leq r\}$ are diamonds, The intersection can happen at one of the axis, thus making the minimizer sparse.



(a) ℓ_2 Minimizer

(b) ℓ_1 Minimizer

Figure 1: Comparison between the two minimizers

With $\mathbf{x}_{\ell_2}$ not satisfying the sparsity requirement, we would like to consider the ℓ_1-norm instead. In fact, the ℓ_1 technique is widely used to induce sparsity on solutions. Hence we have:

$$\min_{\mathbf{x} \in \mathbf{R}^n} \{||\mathbf{x}||_{\ell_1} \mid A\mathbf{x} = \mathbf{b}\} \tag{5}$$

Under certain conditions, the solution $\mathbf{x}_{\ell_1}$ from (5) recovers the original signal $\mathbf{x}_*$ and actually coincides with $\mathbf{x}_{\ell_0}$ from (3). The authors in [4] proved such equivalence with the matrix A being a Fourier Transform matrix. They later generalized the proof to any matrix A with the Restricted Isometry Property in [7, 6, 5]. Then separately the author in [13, 10] showed that when A has the CS1-CS3 properties, $\mathbf{x}_*$ can be recovered from solving (5). Moreover, all of the papers require that the recovery can be achieved only when $m = C * l * log(n) \ll n$.

Remark 1.1. *Specifically the ℓ_1 problem may not have a unique solution. Let us begin our argument by considering a family of more general problems:*

$$\min_{\mathbf{x}\in\mathbf{R}^n}\{J(\mathbf{x}) \mid A\mathbf{x}=\mathbf{b}\} \tag{6}$$

Where $J(\cdot)$ is a continuous and convex energy functional defined on $\mathbf{R}^n$. Let $J(\cdot)$ also be coercive, that is, $J(\mathbf{x}) \geq a||\mathbf{x}||_{\ell_2}$ for some $a > 0$. By convex analysis, the solution set of (6) *is nonempty and convex. And if $J(\cdot)$ is strictly or strongly convex, the solution set contains only one point (unique solution, such as $\mathbf{x}_{\ell_2}$ from* (4)*). However $J(\mathbf{x}) = ||\mathbf{x}||_{\ell_1}$ is not strictly convex, the solution of* (5) *might not be unique. In order to have a unique minimizer, the linear measurement matrix A has to have the Null Space Property stated in [19], that is (for the case $p = 1$) for any index set $\mathcal{I} \subset \{1, 2, \ldots, n\}$ with card($\mathcal{I}$) $\leq l = ||\mathbf{x}_*||_{\ell_0}$ and* $\displaystyle\max_{\mathbf{x}\in Null(A),\mathbf{x}\neq\mathbf{0}} \frac{\sum_{i\in\mathcal{I}}|\mathbf{x}_i|}{||\mathbf{x}||_{\ell_1}} < 1/2.$

1.1 Tikhonov Regularization

Not all of the measurement matrices A have the Null Space Property, thus the minimizer of (5) is not guaranteed to be unique. In order to tackle such ill-posedness, we add Tikhonov Regularization to (5) and bring the constraint into the minimization as a penalty term. Such regularization, which was proposed and developed in [28, 29], is specially designed to tackle ill-posed problems,

$$\min_{\mathbf{x}\in\mathbf{R}^n}\{||\mathbf{x}||_{\ell_1} + \frac{\lambda}{2}||\mathbf{b} - A\mathbf{x}||_{\ell_2}^2\} \tag{7}$$

$||\mathbf{x}||_{\ell_1} + \frac{\lambda}{2}||\mathbf{b} - A\mathbf{x}||_{\ell_2}^2$ is a convex functional defined on $\mathbf{R}^n$ and it is non-negative $\forall\mathbf{x}\in\mathbf{R}^n$, hence the minimizer exists and is unique. This regularized version produces a minimizer in a $\frac{1}{\lambda}$-neighborhood of the minimizer of (5), see [18] regarding an equation of $\mathbf{x}_{\ell_1}$ and $\mathbf{x}_{\ell_1,\lambda}$. We will give a rigorous proof in sec 1.2 to show that $\mathbf{x}_{\ell_1,\lambda} \to \mathbf{x}_{\ell_1}$ when $\lambda\to\infty$. Intuitively, when λ is small, we are basically minimizing $||\mathbf{x}||_{\ell_1}$, which might lead to $\mathbf{x}_{\ell_1,\lambda} = \mathbf{0}$; when λ is large, we are putting more emphasis on the ℓ_2 part, thus forcing $\mathbf{x}_{\ell_1,\lambda}$ to get closer to the affine space $\mathcal{H} = \{\mathbf{x}\in\mathbf{R}^n \mid A\mathbf{x} = \mathbf{b}\}$, hence getting closer to $\mathbf{x}_{\ell_1}$. An optimality condition on $\mathbf{x}_{\ell_1,\lambda}$ being a solution of (7) is that $\mathbf{x}_{\ell_1,\lambda}$ satisfies the following,

$$\mathbf{sgn}(\mathbf{x}) - \lambda A^T\mathbf{r}(\mathbf{x}) = \mathbf{0} \tag{8}$$

Where $\mathbf{r}(\mathbf{x}) = \mathbf{b} - A\mathbf{x}$, and the $\text{sgn}(a) = \begin{cases} 1, & a > 0 \\ -1, & a < 0 \end{cases}$; the vector version, $\mathbf{sgn}(\cdot)$, is defined component wise for vectors. However, sgn(a) is not defined at $a = 0$. Such delicacy makes (8) hard to solve directly; furthermore, the solution set of (8) is contained in the solution set of (7) due to non-differentiability property of sgn($\cdot$) at 0. Since $\mathbf{x}_{\ell_1,\lambda}$ solves (7), then $\mathbf{x}_{\ell_1,\lambda}$ and $\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})$ satisfy (8), and hence

$$\langle\mathbf{x}_{\ell_1,\lambda}, A^T\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})\rangle_{\ell_2} = \langle\mathbf{x}_{\ell_1,\lambda}, \frac{1}{\lambda}\mathbf{sgn}(\mathbf{x}_{\ell_1,\lambda})\rangle_{\ell_2} = \frac{1}{\lambda}||\mathbf{x}_{\ell_1,\lambda}||_{\ell_1}$$

Meanwhile, since $A^T\mathbf{r}(\mathbf{x}_{\ell_1,\lambda}) = \frac{1}{\lambda}\mathbf{sgn}(\mathbf{x}_{\ell_1,\lambda})$, $||A^T\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})||_{\ell_\infty} = \frac{1}{\lambda}$. Assuming $\lambda \geq \frac{1}{||A^T\mathbf{b}||_{\ell_\infty}}$, We show the following theorem,

Theorem 1.2 (Validation Principles)**.**

$$\begin{aligned} \langle\mathbf{x}_{\ell_1,\lambda}, A^T\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})\rangle_{\ell_2} &= ||\mathbf{x}_{\ell_1,\lambda}||_{\ell_1}\,||A^T\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})||_{\ell_\infty} && (9) \\ ||A^T\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})||_{\ell_\infty} &= \frac{1}{\lambda} && (10) \end{aligned}$$

The authors in [26, 27] show that the above principles still hold true for other normed vector spaces like (U, V) pair instead of (ℓ_1, ℓ_2). And we will use (9) and (10) to verify the implementation of various ℓ_1 solvers in sec 3.

Remark 1.3. *By Hölder's inequality, $\langle \mathbf{x}, A^T\mathbf{r}(\mathbf{x})\rangle_{\ell_2} \leq ||\mathbf{x}||_{\ell_1}\ ||A^T\mathbf{r}(\mathbf{x})||_{\ell_\infty}$. Equality is always realized whenever $\mathbf{x} = \mathbf{x}_{\ell_1,\lambda}$. Therefore we call this pair ($\mathbf{x}_{\ell_1,\lambda}$, $\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})$) an extremal pair.*

1.2 Relationship between the Constrained and Unconstrained Minimizations

We consider the relationship between the solution $\mathbf{x}_{\ell_1}$ of (5) and the solution $\mathbf{x}_{\ell_1,\lambda}$ of (7). Then $\mathbf{x}_{\ell_1,\lambda}$ satisfies (8), let $\mathbf{r}(\mathbf{x}_{\ell_1,\lambda}) = \mathbf{b} - A\mathbf{x}_{\ell_1,\lambda}$, we have

$$\begin{aligned} A^T\mathbf{r}(\mathbf{x}_{\ell_1,\lambda}) &= \frac{1}{\lambda}\mathbf{sgn}(\mathbf{x}_{\ell_1,\lambda}) \\ \mathbf{r}(\mathbf{x}_{\ell_1,\lambda}) &= \frac{(AA^T)^{-1}}{\lambda}A\mathbf{sgn}(\mathbf{x}_{\ell_1,\lambda}) \end{aligned}$$

And we have

$$\begin{aligned} ||\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})||_{\ell_p} &\leq \frac{n^{\frac{1}{p}}}{\lambda}||(AA^T)^{-1}||_{\ell_p}||A||_{\ell_p} \quad \text{for } 1 \leq p < \infty \\ ||\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})||_{\ell_\infty} &\leq \frac{1}{\lambda}||(AA^T)^{-1}||_{\ell_\infty}||A||_{\ell_\infty} \end{aligned}$$

We also have

$$\begin{aligned} ||\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})||_{\ell_p} &\geq \frac{(\text{supp}(\mathbf{x}_{\ell_1,\lambda}))^{\frac{1}{p}}}{\lambda||A^T||_{\ell_p}} \quad \text{for } 1 \leq p < \infty \\ ||\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})||_{\ell_\infty} &\geq \frac{1}{\lambda||A^T||_{\ell_\infty}} \end{aligned}$$

We just proved the following lemma

Lemma 1.4. *The residual $\mathbf{r}(\mathbf{x}_{\ell_1,\lambda}) = \mathbf{b} - A\mathbf{x}_{\ell_1,\lambda}$ of (7) satisfies the following bound:*

$$\frac{(\mathit{supp}(\mathbf{x}_{\ell_1,\lambda}))^{\frac{1}{p}}}{\lambda}||A^T||_{\ell_p}^{-1} \leq \ ||\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})||_{\ell_p} \ \leq \frac{n^{\frac{1}{p}}}{\lambda}||(AA^T)^{-1}||_{\ell_p}||A||_{\ell_p} \quad \mathit{for}\ 1 \leq p < \infty \tag{11}$$

$$\frac{1}{\lambda}||A^T||_{\ell_\infty}^{-1} \leq \ ||\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})||_{\ell_\infty} \ \leq \frac{1}{\lambda}||(AA^T)^{-1}||_{\ell_\infty}||A||_{\ell_\infty} \tag{12}$$

Then we can look at $||\mathbf{x}_{\ell_1}||_{\ell_1} - ||\mathbf{x}_{\ell_1,\lambda}||_{\ell_1}$. First, since $\mathbf{x}_{\ell_1,\lambda}$ solves (7),

$$||\mathbf{x}_{\ell_1,\lambda}||_{\ell_1} + \frac{\lambda}{2}||A\mathbf{x}_{\ell_1,\lambda} - \mathbf{b}||_{\ell_2}^2 \leq ||\mathbf{x}_{\ell_1}||_{\ell_1} + \frac{\lambda}{2}||A\mathbf{x}_{\ell_1} - \mathbf{b}||_{\ell_2}^2 = ||\mathbf{x}_{\ell_1}||_{\ell_1}$$

Thus

$$||\mathbf{x}_{\ell_1}||_{\ell_1} - ||\mathbf{x}_{\ell_1,\lambda}||_{\ell_1} \geq \frac{\lambda}{2}||\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})||_{\ell_2}^2 \geq \frac{\lambda}{2}\frac{\text{supp}(\mathbf{x}_{\ell_1,\lambda})}{\lambda^2}||A^T||_{\ell_2}^{-2} = \frac{\text{supp}(\mathbf{x}_{\ell_1,\lambda})}{2\lambda}||A||_{\ell_2}^{-2}$$

Second, pick an $\Delta\mathbf{x}$ be such that $A(\mathbf{x}_{\ell_1,\lambda} + \Delta\mathbf{x}) = \mathbf{b}$, then from the (5), we know

$$||\mathbf{x}_{\ell_1}||_{\ell_1} \leq ||\mathbf{x}_{\ell_1,\lambda} + \Delta\mathbf{x}||_{\ell_1} \leq ||\mathbf{x}_{\ell_1,\lambda}||_{\ell_1} + ||\Delta\mathbf{x}||_{\ell_1}$$

Hence $||\mathbf{x}||_{\ell_1} - ||\mathbf{x}_*||_{\ell_1} \leq ||\Delta\mathbf{x}||_{\ell_1}$; since $\Delta\mathbf{x} = \mathbf{b} - A\mathbf{x}_{\ell_1,\lambda} = \mathbf{r}(\mathbf{x}_{\ell_1,\lambda})$, then

$$||\Delta\mathbf{x}||_{\ell_1} = ||\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})||_{\ell_1} \leq \frac{n}{\lambda}||(AA^T)^{-1}||_{\ell_1}||A||_{\ell_1}$$

Therefore

$$||\mathbf{x}_{\ell_1}||_{\ell_1} - ||\mathbf{x}_{\ell_1,\lambda}||_{\ell_1} \leq \frac{n}{\lambda}||(AA^T)^{-1}||_{\ell_1}||A||_{\ell_1}$$

We conclude the following lemma

Lemma 1.5. *The difference between the* $||\mathbf{x}_{\ell_1}||_{\ell_1}$ *and* $||\mathbf{x}_{\ell_1,\lambda}||_{\ell_1}$ *is:*

$$\frac{supp(\mathbf{x}_{\ell_1,\lambda})}{2\lambda}||A||_{\ell_2}^{-2} \leq ||\mathbf{x}_{\ell_1}||_{\ell_1} - ||\mathbf{x}_{\ell_1,\lambda}||_{\ell_1} \leq \frac{n}{\lambda}||(AA^T)^{-1}||_{\ell_1}||A||_{\ell_1} \tag{13}$$

With the two lemmas established, we are able to present the following theorem

Theorem 1.6. *Given that* A *has the Null Space Property, then* $\mathbf{x}_{\ell_1,\lambda} \to \mathbf{x}_{\ell_1}$ *as* $\lambda \to \infty$.

Proof. By lemma 1.4, we know that $||\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})||_{\ell_\infty} \to 0$ as $\lambda \to \infty$, hence $\mathbf{x}_{\ell_1,\lambda} \in \mathcal{H}$. And by lemma 1.5, we know $\mathbf{x}_{\ell_1,\lambda}$ and $\mathbf{x}_{\ell_1}$ have the same ℓ_1-norm. since A has the Null Space Property, (5) has unique minimizer, then the two must be the same. □

2 Hierarchical Reconstruction

Following similar arguments by the authors in [26], we present the following motivation for the Hierarchical Reconstruction algorithm: We want to recover the support of $\mathbf{x}_*$ through a series of iterative minimization by extracting useful signal from the previous "residual".

2.1 Motivation

Let $\mathbf{x}_\lambda$ and $\mathbf{r}_\lambda$ be an extremal pair such that they solve (7), that is:

$$[\mathbf{x}_\lambda, \mathbf{r}_\lambda] = \underset{A\mathbf{x}+\mathbf{r}=\mathbf{b}}{\arg\min}\ \{||\mathbf{x}||_{\ell_1} + \frac{\lambda}{2}||\mathbf{r}||_{\ell_2}^2\} \tag{14}$$

The pair $[\mathbf{x}_\lambda, \mathbf{r}_\lambda]$ also decompose $\mathbf{b}$ into two parts: the recovered sparse signal $\mathbf{x}_\lambda$ and residual $\mathbf{r}_\lambda$ under a given scale λ. However, the residual term $\mathbf{r}_\lambda$ still contains useful information (possible other basis vectors within the support of $\mathbf{x}_*$) when viewed under a refined scale, say 2λ:

$$[\mathbf{x}_{2\lambda}, \mathbf{r}_{2\lambda}] = \underset{A\mathbf{x}+\mathbf{r}=\mathbf{r}_\lambda}{\arg\min}\ \{||\mathbf{x}||_{\ell_1} + \lambda||\mathbf{r}||_{\ell_2}^2\} \tag{15}$$

Since $\mathbf{b} = A(\mathbf{x}_\lambda + \mathbf{x}_{2\lambda}) + \mathbf{r}_{2\lambda} \approx A(\mathbf{x}_\lambda + \mathbf{x}_{2\lambda})$, we obtain a better two-scale approximation to $\mathbf{b}$; signal below scale $\frac{1}{2\lambda}$ remains unresolved in $\mathbf{r}_{2\lambda}$. This process of (14) and (15) can continue. Beginning with an initial scale λ_0, and initial residual $\mathbf{r}_0 = \mathbf{b}$,

$$[\mathbf{x}_1, \mathbf{r}_1] = \underset{A\mathbf{x}+\mathbf{r}=\mathbf{r}_0}{\arg\min}\ \{||\mathbf{x}||_{\ell_1} + \frac{\lambda_0}{2}||\mathbf{r}||_{\ell_2}^2\}$$

we continue in this iterative manner for the decomposition of the dyadic refinement step of (15),

$$[\mathbf{x}_{j+1}, \mathbf{r}_{j+1}] = \underset{A\mathbf{x}+\mathbf{r}=\mathbf{r}_j}{\arg\min}\ \{||\mathbf{x}||_{\ell_1} + \frac{\lambda_j}{2}||\mathbf{r}||_{\ell_2}^2\}, \quad j = 1, 2, \ldots \tag{16}$$

generating, after k such steps, the following **Hierarchical Decomposition** of b:

$$\begin{aligned} \mathbf{b} &= A\mathbf{x}_1 + \mathbf{r}_1 \\ &= A\mathbf{x}_1 + A\mathbf{x}_2 + \mathbf{r}_2 \\ &= \ldots\ldots \\ &= A\mathbf{x}_1 + A\mathbf{x}_2 + \ldots + A\mathbf{x}_k + \mathbf{r}_k. \end{aligned} \tag{17}$$

We arrive at a new multi-scale decomposition, $\mathbf{b} \approx A\mathbf{x}_1 + A\mathbf{x}_2 + \ldots + A\mathbf{x}_k$, with a residual $\mathbf{r}_k$. And such iterative approach gives rise to our algorithm:

Algorithm 1 Hierarchical Reconstruction of Sparse Signals

```
Require: $A$ and $\mathbf{b}$, pick a $\lambda_0 > \frac{1}{||A^T\mathbf{b}||_{\ell_\infty}}$
  Initialize: $\mathbf{x}_{HRSS} = 0$, $\mathbf{r}_0 = \mathbf{b}$, and $j = 1$
  while $j \leq J$ or other convergence criteria not satisfied do
    $\mathbf{x}_j := \arg\min_{\mathbf{x}\in\mathbf{R}^n} \{||\mathbf{x}||_{\ell_1} + \frac{\lambda_{j-1}}{2}||A\mathbf{x} - \mathbf{r}_{j-1}||^2_{\ell_2}\}$
    $\mathbf{r}_j = \mathbf{r}_{j-1} - A\mathbf{x}_j$
    $\mathbf{x}_{HRSS} = \mathbf{x}_{HRSS} + \mathbf{x}_j$
    $\lambda_{j+1} = 2\lambda_j;$
    $j = j + 1;$
  end while
Ensure: $\mathbf{x}_{HRSS} = \sum_{j=0}^{J} \mathbf{x}_j;$
```

As the final stopping J increases, we construct the signal $\mathbf{x}_{HRSS}$ with increasing support. Since only $l = ||\mathbf{x}_*||_{\ell_0}$ is known, not the actually support of $\mathbf{x}_*$, this iterative process might recover something else rather then the original signal $\mathbf{x}_*$; however the difference, $\mathbf{x}_{HRSS} - \mathbf{x}_*$, is going to be in Null$(A)$. We still need to find a way to stop the algorithm. Meanwhile such a dyadic approach is not necessary, one can use $\lambda_j = 3\lambda_{j-1}$, and run the algorithm, and the analysis will still go through similarly.

2.2 Convergence Analysis of the Hierarchical Reconstruction Approach

We consider the minimizer, $\mathbf{x}_k$, at the k^{th} step ($1 \leq k \leq J-1$), it satisfies the signum equation, namely, $\mathbf{sgn}(\mathbf{x}_k) + \lambda_k A^T(A\mathbf{x}_k - \mathbf{r}_{k-1}) = \mathbf{0}$. And since$\mathbf{r}_k = \mathbf{r}_{k-1} - A\mathbf{x}_k$, we can have

$$A^T\mathbf{r}_k = \frac{1}{\lambda_k}\mathbf{sgn}(\mathbf{x}_k) \tag{18}$$

Combined with the signum equation at the $(k+1)^{th}$ step, we derive the following

$$A^TA\mathbf{x}_{k+1} = \frac{1}{\lambda_k}\mathbf{sgn}(\mathbf{x}_k) - \frac{1}{\lambda_{k+1}}\mathbf{sgn}(\mathbf{x}_{k+1}) \tag{19}$$

Since $\lambda_{k+1} = 2\lambda_k$, $||A^TA\mathbf{x}_{k+1}||_\infty \leq \frac{3}{2\lambda_k}$. As $\lambda_k \to \infty$, $||A^TA\mathbf{x}_{k+1}||_\infty \to 0$. And A^T has linearly independent rows, $A\mathbf{x}_{k+1} \to \mathbf{0}$. As λ_k is sufficiently large, we are roughly adding more and more correction from Null(A). Meanwhile $\mathbf{b} = A\mathbf{x}_{HRSS} + \mathbf{r}_J$ combined together with (18), we show that

$$A^TA\mathbf{x}_{HRSS} = A^T\mathbf{b} - \frac{1}{\lambda_J}\mathbf{sgn}(\mathbf{x}_J) \tag{20}$$

If there is no noise in the measurements, then $\mathbf{b} = A\mathbf{x}_*$, hence

$$A^T A(\mathbf{x}_{HRSS} - \mathbf{x}_*) = -\frac{1}{\lambda_J}\mathbf{sgn}(\mathbf{x}_J) \tag{21}$$

As $J \to \infty$, $\frac{1}{\lambda_J} \to 0$, and $||A^T A(\mathbf{x}_{HRSS} - \mathbf{x}_*)||_{\ell_\infty} \to 0$, then $\mathbf{x}_{HRSS} - \mathbf{x}_*$ will be eventually in Null(A). In order to avoid recovery of a different minimizer, we will stop the algorithm using criteria based on relative update and small residual. However most of the time, there is noise in the measurement, that is, $\mathbf{b} = A\mathbf{x}_* + \epsilon$, then (20) becomes

$$A^T A(\mathbf{x}_{HRSS} - \mathbf{x}_*) = A^T\epsilon - \frac{1}{\lambda_J}\mathbf{sgn}(\mathbf{x}_J) \tag{22}$$

In order to de-noise, we will pick a λ_J suth that $\frac{1}{\lambda_J}\mathbf{sgn}(\mathbf{x}_J)$ is comparable to $A^T\epsilon$ in ℓ_2-norm; when ϵ is a white noise, that is, $\epsilon \sim \mathrm{N}(\mathbf{0}, I)$, and together a priori information of the bounds on ϵ, i.e. $||A^T\epsilon||_{\ell_2} \leq \mathrm{tol}_\epsilon$. We will pick $\lambda_J > \frac{1}{\mathrm{tol}_\epsilon}$. And if ϵ is correlated, that is $\epsilon \sim \mathrm{N}(\mathbf{0}, R)$ where R is a correlation matrix (non-diagonal). Then we will just work on similar problem normalized by $R^{\frac{1}{2}}$, that is, $\min_{\mathbf{x}\in\mathbf{R}^n}\{||R^{\frac{1}{2}}\mathbf{x}||_{\ell_1} + \frac{\lambda}{2}||R^{\frac{1}{2}}(\mathbf{b} - A\mathbf{x})||^2_{\ell_2}\}$. Let us also consider the solution, $\mathbf{x}_{\ell_1,\lambda}$, of (7) with $\lambda = \lambda_J$, and then $\mathbf{sgn}(\mathbf{x}_{\ell_1,\lambda}) + \lambda_J A^T(A\mathbf{x}_{\ell_1,\lambda} - \mathbf{b}) = \mathbf{0}$, together with (20), we get

$$A^T A(\mathbf{x}_{HRSS} - \mathbf{x}_{\ell_1,\lambda}) = \frac{1}{\lambda_J}(\mathbf{sgn}(\mathbf{x}_{\ell_1,\lambda}) - \mathbf{sgn}(\mathbf{x}_J)) \tag{23}$$

If the $\mathbf{x}_J$ recovered from the J^{th} minimization process contains mostly zero and λ_J is large, then $\mathbf{x}_{HRSS} - \mathbf{x}_{\ell_1,\lambda} \in$ Null(A).

3 Implementation

We choose the Gradient Projection for Sparse Reconstruction method and A Fixed-Point Continuation method as built-in single scale solvers for (7). We have picked these two methods out for their robustness and efficiency over IST [8, 16], l1_ls package [21], ℓ_1-magic toolbox[3], and homotopy method [15].

3.1 Gradient Projection for Sparse Reconstruction

The **Gradient Projection for Sparse Reconstruction** algorithm is proposed in [17] solve the following:

$$\min_{\mathbf{x}\in\mathbf{R}^n}\{\tau||\mathbf{x}||_{\ell_1} + \frac{1}{2}||\mathbf{b} - A\mathbf{x}||^2_{\ell_2}\} \tag{24}$$

Compared to (7), (24) puts the regularization on the ℓ_1 term instead. Since $\tau > 0$ and $\lambda > 0$, apparently $\tau = \frac{1}{\lambda}$ for finding the same minimizer. Let $(a)_+ = \begin{cases} a, & a \geq 0 \\ 0, & \text{otherwise} \end{cases}$, and let $\mathbf{u} = (\mathbf{x})_+$, $\mathbf{v} = (-\mathbf{x})_+$. With $\mathbf{u}$ and $\mathbf{v}$ substituted back into (24), eqrefeq:anotherl1 can be transformed into a linear problem. First let

$$\mathbf{z} = \begin{bmatrix}\mathbf{u}\\ \mathbf{v}\end{bmatrix}, \quad \mathbf{y} = A^T\mathbf{b}, \quad \mathbf{c} = \tau\mathbf{1}_{2n} + \begin{bmatrix}-\mathbf{y}\\ \mathbf{y}\end{bmatrix}, \quad B = \begin{bmatrix}A^TA & -A^TA\\ -A^TA & A^TA\end{bmatrix}$$

then,

$$\min_{\mathbf{z}\in\mathbf{R}^{2n}}\{F(\mathbf{z}) \equiv \mathbf{c}^T\mathbf{z} + \frac{1}{2}\mathbf{z}^TB\mathbf{z} \mid \mathbf{z} \geq 0\} \tag{25}$$

Two algorithms are presented, both use gradient projection techniques to find the minimizer of (24). They both pick the descent direction $\mathbf{d}^{(k)}$ as (with $\alpha^{(k)} > 0$),

$$\mathbf{d}^{(k)} = (\mathbf{z}^{(k)} - \alpha^{(k)}\nabla F(\mathbf{z}^{(k)}))_+ - \mathbf{z}^{(k)} \tag{26}$$

and both update the next iterate $\mathbf{z}^{(k+1)}$ as (with $\nu^{(k)}$),

$$\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + \nu^{(k)}\mathbf{d}^{(k)} \tag{27}$$

The two approaches differ by choosing different $\alpha^{(k)}$ and $\lambda^{(k)}$. Let us consider the basic approach,

Algorithm 2 GPSR Basic

```
Require: $A$, $\mathbf{b}$, $\tau$, and $\mathbf{z}^{(0)}$, pick $\beta \in (0,1)$ and $\mu \in (0,1/2)$
  Initialize: $k = 0$;
  while A stopping criteria is not satisfied do
    Compute $\alpha_0 = \arg\min_{\alpha\in\mathbf{R}^1} \{F(\mathbf{z}^{(k)} - \alpha\mathbf{G}^{(k)})\}$
    Let $\alpha^{(k)}$ be the first in the sequence: $\alpha_0$, $\beta\alpha_0$, $\beta^2\alpha_0, \ldots$, such that
  $F((\mathbf{z}^{(k)} - \alpha^{(k)}\nabla F(\mathbf{z}^{(k)}))_+) \le F(\mathbf{z}^{(k)}) - \mu\nabla F(\mathbf{z}^{(k)})^T(\mathbf{z}^{(k)} - (\mathbf{z}^{(k)} - \alpha^{(k)}\nabla F(\mathbf{z}^{(k)}))_+)$
    Set $\mathbf{z}^{(k+1)} = (\mathbf{z}^{(k)} - \alpha^{(k)}\nabla F(\mathbf{z}^{(k)}))_+$
    Update $k = k + 1$
  end while
Ensure: $\mathbf{z}^{(K)}(\tau) := \arg\min_{\mathbf{z}\in\mathbf{R}^{2n}} \{F(\mathbf{z}) \equiv \mathbf{c}^*\mathbf{z} + \frac{1}{2}\mathbf{z}^*B\mathbf{z} \mid \mathbf{z} \ge 0\}$;
```

The basic approach chooses $\alpha^{(k)}$ from a variable line search such that $F((\mathbf{z}^{(k)}-\alpha^{(k)}\nabla F(\mathbf{z}^{(k)}))_+)$ is minimal. An initial guess $\mathbf{z}^{(0)} = \frac{1}{2n}\mathbf{1}_{2n}$ is chosen because $||\mathbf{z}^{(0)}||_{\ell_1} = 1$. Since $F(\cdot)$ is quadratic, there is formula for calculating α_0, namely $\alpha_0 = \frac{(\mathbf{G}^{(k)})^T\nabla F(\mathbf{z}^{(k)})}{(\mathbf{G}^{(k)})^T B\mathbf{G}^{(k)}}$; and α_0 is restricted that that $\alpha_0 \in [\alpha_{\min}, \alpha_{\max}]$. β is a back-tracking parameter so that the step size for the gradient descent would be optimal. μ is used that that $F(\cdot)$ is decreased sufficiently from the "Armijo rule along the projection arc" by [2, p. 226] and $\mathbf{G}^{(k)}$ is a projected gradient, defined component wise:

$$G_i^{(k)} = \begin{cases} (\nabla F(\mathbf{z}^{(k)}))_i, & \text{if} \quad \mathbf{z}_i^{(k)} > 0 \quad \text{or} \quad (\nabla F(\mathbf{z}^{(k)}))_i < 0 \\ 0, & \text{otherwise} \end{cases} \tag{28}$$

Thus $(\mathbf{G}^{(k)})^T\nabla F(\mathbf{z}^{(k)}) = (\mathbf{G}^{(k)})^T\mathbf{G}^{(k)}$. The second approach is based on Barzilai and Broweign's paper [1] and BCQP approach from [25] to avoid the decaying convergence rate property of some steepest descent methods. Each time an update is calculated as $\mathbf{d}^{(k)} = -H_k^{-1}\nabla F(\mathbf{z}^{(k)})$, where H_k is the Hessian of $F(\mathbf{z}^{(k)})$. H_k is approximated by a multiple of the identity $H_k \approx \rho^{(k)}I$, where $\rho^{(k)}$ is chosen as:

$$\nabla F(\mathbf{z}^{(k)}) - \nabla F(\mathbf{z}^{(k-1)}) \approx \rho^{(k)}\big(\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\big)$$

Then $\alpha^{(k)} = (\rho^{(k)})^{-1}$ and is restricted to the interval $[\alpha_{\min}, \alpha_{\max}]$; and let $\nu^{(k)} \in [0,1]$ be the exact minimizer of $F(\mathbf{z}^{(k)} + \nu\mathbf{d}^{(k)})$,

The same initial guess $\mathbf{z}^{(0)}$ as it is used in algorithm 2 is chosen. As it is stated in [17], several schemes for picking suitable stopping criteria are considered,

- The approximation $\mathbf{z}$ is close to a solution $\mathbf{z}_*$.

Algorithm 3 GPSR Barzilai Browein

```
Require: $A$, $\mathbf{b}$, $\tau$, $\mathbf{z}^{(0)}$, $\alpha_{\min}$, $\alpha_{\max}$, and pick $\alpha^{(0)} \in [\alpha_{\min}, \alpha_{\max}]$
  Initialize: $k = 0$;
  while A stopping criteria is not satisfied do
      Compute $\mathbf{d}^{(k)} = (\mathbf{z}^{(k)} - \alpha^{(k)}\nabla F(\mathbf{z}^{(k)}))_+ - \mathbf{z}^{(k)}$
      Compute $\omega^{(k)} = (\mathbf{d}^{(k)})^T B\mathbf{d}^{(k)}$
      if $\omega^{(k)} = 0$ then
          Set $\alpha^{(k+1)} = \alpha_{\max}$ and $\nu^{(k)} = 1$
      else
          Compute $\alpha^{(k)} = \text{mid}(\alpha_{\min}, \frac{||\mathbf{d}^{(k)}||_{\ell_2}^2}{\omega^{(k)}}, \alpha_{\max})$ and $\nu^{(k)} = \text{mid}(0, \frac{-(\mathbf{d}^{(k)})^T\nabla F(\mathbf{z}^{(k)})}{(\mathbf{d}^{(k)})^T B\mathbf{d}^{(k)}}, 1)$
      end if
      set $\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + \nu^{(k)}\mathbf{d}^{(k)}$
      Update $k = k + 1$
  end while
Ensure: $\mathbf{z}^{(K)} := \arg\min_{\mathbf{z}\in\mathbf{R}^{2n}} \{F(\mathbf{z}) \equiv \mathbf{c}^T\mathbf{z} + \frac{1}{2}\mathbf{z}^T B\mathbf{z} \mid \mathbf{z} \geq 0\}$;
```

- The function value $F(\mathbf{z})$ is close to $F(\mathbf{z}_*)$.
- The non-zero components of the approximation $\mathbf{z}$ are close to the non-zero components of $\mathbf{z}_*$.

Suggested in [17], the following stopping criterion are implemented,

- $||\mathbf{z} - (\mathbf{z} - \bar{\alpha}\nabla F(\mathbf{z}))_+||_2 \leq$ tolP, where $\bar{\alpha}$ is a positive constant, and tolP is a small parameter. Note that when $\mathbf{z}$ is optimal, the left hand side is zero.
- $||\mathbf{min}(\mathbf{z}, \nabla F(\mathbf{z}))||_2 \leq$ tolP, the $\mathbf{min}(\cdot)$ is taken component wise. It is motivated by perturbation results from linear complementarity problems (LCP). LCP results show that there is a constant C_{LCP} such that dist$(\mathbf{z}, \mathcal{S}) \leq C_{CLP}||\mathbf{min}(\mathbf{z}, \nabla F(\mathbf{z}))||_2$, where $\mathcal{S}$ is a solution set of (25). When $||\mathbf{min}(\mathbf{z}, \nabla F(\mathbf{z}))||_2 \leq$ tolP, $\mathbf{z}$ is forced to be within a certain distance of the solution set $\mathcal{S}$.
- $\left|\frac{1}{2}||\mathbf{b} - A\mathbf{x}||_{\ell_2}^2 + \tau||\mathbf{x}||_{\ell_1} + \frac{1}{2}\mathbf{s}^T\mathbf{s} + \mathbf{b}^T\mathbf{s}\right| \leq$ tolP, where $\mathbf{s}$ is a minimizer of the dual problem of (24),

  $$\max_{\mathbf{s}}\{-\frac{1}{2}\mathbf{s}^T\mathbf{s} - \mathbf{b}^T\mathbf{s} \mid -\tau\mathbf{1}_n \leq A^T\mathbf{s} \leq \tau\mathbf{1}_n\} \tag{29}$$

  Therefore on each iterate we resemble $\mathbf{x} = \mathbf{u} - \mathbf{v}$ and compute $\mathbf{s} = \tau\frac{A\mathbf{x}-\mathbf{b}}{||A^T(A\mathbf{x}-\mathbf{b})||_{\ell_\infty}}$, and substitute the pair back into the left hand side of the inequality and to check to see if the left hand side is less than tolP.
- $|\mathcal{C}_k|/|\mathcal{I}_k| \leq$ tolP, where $\mathcal{I}_k = \{i \mid \mathbf{z}_i^{(k)} \neq 0\}$ and $\mathcal{C}_k = \{i \mid i \in \mathcal{I}_k \oplus \mathcal{I}_{k-1}\}$[5]. Such ratio takes into account that changes in the non-zero entries in the approximation $\mathbf{z}$ are going to be small when $\mathbf{z}$ is close to $\mathbf{z}_*$.

[5] $\oplus$ stands for the exclusive union of two sets

Remark 3.1. *Although GPSR algorithms have increased the system size from n to $2n$, the matrix-vector multiplication can still be done at the $\mathcal{O}(n)$. We can simply the matrix-vector multiplication by:*

$$B\mathbf{z} = \begin{bmatrix} A^TA(\mathbf{u}-\mathbf{v}) \\ -A^TA(\mathbf{u}-\mathbf{v}) \end{bmatrix}, \quad \mathbf{c}^T\mathbf{z} = \tau\mathbf{1}_n^T(\mathbf{u}+\mathbf{v}) - \mathbf{y}^T(\mathbf{u}-\mathbf{v}), \quad \mathbf{z}^TB\mathbf{z} = ||A(\mathbf{u}-\mathbf{v})||_2$$

Hence,

$$F(\mathbf{z}) = \tau\mathbf{1}_n^T(\mathbf{u}+\mathbf{v}) - \mathbf{y}^T(\mathbf{u}-\mathbf{v}) + \frac{1}{2}||A(\mathbf{u}-\mathbf{v})||_2, \quad \nabla F(\mathbf{z}) = \begin{bmatrix} \tau\mathbf{1}_n - \mathbf{y} + A^TA(\mathbf{u}-\mathbf{v}) \\ \tau\mathbf{1}_n + \mathbf{y} - A^TA(\mathbf{u}-\mathbf{v}) \end{bmatrix}$$

$\mathbf{1}_n^T(\mathbf{u}+\mathbf{v})$ *can be done using the MATLAB^{TM} built-in summation function for vectors. We do not compute the matrix-matrix product A^TA directly; instead we will do $A(\mathbf{u}-\mathbf{v})$ first then $A^T(\cdot)$.*

Theorem 1 in [17] states that the sequence of $\{\mathbf{z}^{(k)}\}$ generated by the GPSR algorithms either terminate at a solution of (25) or converge to a solution of (25) at an R-linear rate.

3.2 Fixed-Point Continuation Method

The authors in [20] develops the **Fixed-Point Continuation** algorithm to solve a signum equation associated with the following minimization problem,

$$\min_{\mathbf{x}\in\mathbf{R}^n}\{||\mathbf{x}||_1 + \frac{\lambda}{2}||\mathbf{b} - A\mathbf{x}||_D^2\} \tag{30}$$

Where $||\mathbf{x}||_D := \sqrt{\mathbf{x}^TD\mathbf{x}}$ with D being a Symmetric Positive Definite matrix ($||\cdot||_D$ is a scaled ℓ_2 norm). In this project, we will take $D = I$. And they also define $f : \mathbf{R}^n \to \mathbf{R}$ and $\mathbf{g} : \mathbf{R}^n \to \mathbf{R}^n$ as,

$$\begin{aligned} f(\mathbf{x}) &:= \frac{1}{2}||\mathbf{b} - A\mathbf{x}||_{\ell_2}^2 \\ \mathbf{g}(\mathbf{x}) &:= \nabla f(\mathbf{x}) = A^T(A\mathbf{x} - \mathbf{b}) \end{aligned}$$

And $\mathbf{s}_\xi$ and $\mathbf{h}$, both from $\mathbf{R}^n$ to $\mathbf{R}^n$, are defined as (for any $\eta > 0$),

$$\mathbf{h}(\mathbf{x}) := \mathbf{x} - \eta\mathbf{g}(\mathbf{x}) \tag{31}$$

$$\mathbf{s}_\xi(\mathbf{x}) := \mathbf{sgn}(\mathbf{x}) \odot \max\{|\mathbf{x}| - \xi, 0\} \tag{32}$$

where $\xi = \frac{\eta}{\lambda}$ and $\odot$ is a component wise multiplication for vectors. Let us consider $T(\mathbf{x}) = \frac{1}{\lambda}\mathbf{sgn}(\mathbf{x}) + A^T(A\mathbf{x} - \mathbf{b})$. T is shown to be a maximal monotone operator in [24]. T can be split linearly into two parts, i.e. $T = T_1 + T_2$. A parameter $\eta > 0$ is picked that $I + \eta T_1$ is invertible. Then $\mathbf{0} \in T(\mathbf{x}_{\ell_1,\lambda}) \leftrightarrow \mathbf{x}_{\ell_1,\lambda} = (I + \eta T_1)^{-1}(I - \eta T_2)\mathbf{x}_{\ell_1,\lambda}$. For this project, $T_2(\mathbf{x}) = \mathbf{g}(\mathbf{x})$ and $T_1(\mathbf{x}) = \frac{1}{\lambda}\mathbf{sgn}(\mathbf{x})$. Then $(I + \eta T_1)^{-1} = \mathbf{s}_\xi$. It gives rise to the following fixed point equation,

$$\mathbf{x}^{(k+1)} = \mathbf{s}_\xi \circ \mathbf{h}(\mathbf{x}^{(k)}) \tag{33}$$

Remark 3.2. *As it is shown in [20], if $\mathbf{x}_{\ell_1,\lambda}$ solves* (7)*, then $\mathbf{0} \in \boldsymbol{sgn}(\mathbf{x}_{\ell_1,\lambda}) + \lambda\mathbf{g}(\mathbf{x}_{\ell_1,\lambda})$ and vice versa. And if $\mathbf{x}_{sgn,\lambda}$ is a fixed point of* (33)*, then $\mathbf{x}_{sgn,\lambda} \in \boldsymbol{sgn}(\mathbf{x}_{sgn,\lambda}) + \lambda\mathbf{g}(\mathbf{x}_{sgn,\lambda})$ and vice versa. Therefore $\mathbf{x}_{\ell_1,\lambda} = \mathbf{x}_{sgn,\lambda}$.*

Let $\varrho_{\max}$ be the maximum eigenvalue of the Hessian of $f(\mathbf{x})$, namely $H(\mathbf{x}) = A^T A$. As it is shown in [20], η is picked in $(0, \frac{2}{\varrho_{\max}})$[6] in order to have convergence results. Meanwhile, we know that for small λ, the minimizer $\mathbf{x}_{\ell_1,\lambda}$ is close to $\mathbf{0}$, making the program faster to find solution at such λ. Hence for a sequence of $\lambda_1 < \lambda_2 < \ldots < \lambda_J = \lambda$, a sequence of minimizers $\mathbf{x}_{\ell_1,\lambda_j}$, for $1 \le j \le J$, is generated using the minimizer from λ_{j-1} as an initial guess. Assuming that this solution path $\mathbf{x}_1 \to \mathbf{x}_2 \to \ldots \mathbf{x}_J$ is continuous, and A has the Null Space Property, then the final $\mathbf{x}_J$ is easier to find then just use the final λ at the beginning. Under this setting, [20] proposes the following algorithm:

Algorithm 4 Fixed Point Continuation Method

```
Require: $A$, $\mathbf{b}$, $\lambda$, pick $\mathbf{x}^{(0)}$, set $\bar{\lambda} = \lambda$
  Select: $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_L = \bar{u}$
  for $\lambda_* = \lambda_1, \lambda_2, \cdots, \lambda_L$ do
      while A convergence test is not satisfied do
          Select $\eta$ and set $\xi = \frac{\eta}{\lambda_*}$
          $\mathbf{x}^{(k+1)} = \mathbf{s}_\xi \circ \mathbf{h}(\mathbf{x}^{(k)})$
      end while
  end for
```

With 2 extra tolerances, xTol and gTol, the following convergence tests is used,

$$\frac{||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}||_{\ell_2}}{\max(||\mathbf{x}^{(k)}||_{\ell_2}, 1)} < \text{xtol} \quad \text{and} \quad \mu_i ||\mathbf{g}(\mathbf{x}^{(k)})||_{\ell_\infty} - 1 < \text{gtol} \tag{34}$$

(34) uses the idea of small relative update; and when $\mathbf{x}_{FPC}$ is closed to $\mathbf{x}_{\ell_1,\lambda}$, then $\mathbf{x}_{FPC}$ should satisfy (8), which leads to the second half of the stopping criterion. Although FPC is straightforward to implement, the algorithm depends on a suitable choice of η, an appropriate sequence of λ_i's and a proper initial guess $\mathbf{x}^{(0)}$. All of these requirement make the parameter toning more delicate when applied to other problems rather than Compressed Sensing.

3.3 The De-biasing Step Vs. Hierarchical Reconstruction

After a minimizer is obtained either through one of the two GPSR algorithms or the FPC algorithm, call it $\mathbf{x}_{\min}$, we will perform a de-biasing step when necessary, that is, to minimize the residual $||\mathbf{b} - A\mathbf{x}||_{\ell_2}$. The general procedure goes like: we first find the zero entries of $\mathbf{x}_{\min}$, (an entry of $\mathbf{x}_{\min}$ is considered zero if its absolute value is either below 10^{-12} or a preset tolerance given by the user), and keep the zero entries of $\mathbf{x}_{\min}$ unchanged while performing CG steps on $||\mathbf{b} - A\mathbf{x}||_{\ell_2}^2$ so that either $||\mathbf{b} - A\mathbf{x}||_{\ell_2} < \text{tol}_D * ||\mathbf{b} - A\mathbf{x}_{\min}||_{\ell_2}$ or all the possible CG iterations are exhausted. De-biasing might not give satisfactory results, since $\mathbf{x}_{\min}$ recovered from one of the single scale solvers might not have the same support as the original $\mathbf{x}_*$. The aforementioned CG process might not be able to find a minimizer so that $||\mathbf{b} - A\mathbf{x}||_{\ell_2} < \text{tol}_D * ||\mathbf{b} - A\mathbf{x}_{\min}||_{\ell_2}$. The whole de-biasing is based up on the idea that the single scale solver would recover most of support of $\mathbf{x}_*$ ($\text{supp}(\mathbf{x}_{\min}) \subset \text{supp}(\mathbf{x}_*)$), so that $||\mathbf{b} - A\mathbf{x}_{\min}||_{\ell_2}$ could be toned down by only changing the non-zero entries of $\mathbf{x}_{\min}$. Compared to normal de-biasing steps, the Hierarchical Reconstruction algorithm offers a similar approach (decreasing residual in ℓ_2-norm), while keeping the $\mathbf{x}_k$ in ℓ_1-norm small at each iterate. What remains open is whether $\mathbf{x}_{HRSS}$, the sum of all J iterates, is also comparably small in the ℓ_1 norm.

[6] $\eta \in [\frac{1}{\varrho_{max}}, \frac{2}{\varrho_{max}})$ for faster convergence.

3.4 Implementation Platform and Memory Allocation

Codes will be written in MATLAB$^{\text{TM}}$for GPSR Basic, GPSR Barzilai Borwein, FPC, and the whole HRoSS algorithm. When time permits, parallel codes will be written in C. The version of the MATLAB$^{\text{TM}}$running on my personal computer is: $7.12.0.635(R2011a)$. It is installed on a copy of the Windows7$^{\text{TM}}$ Home Premium operating system (64 bit). Validations and testing of GPSR Basic, GPSR Barzilai Borwein and FPC will be run on my personal computer with AMD Phenom$^{\text{TM}}N950$ Quad-Core processor (clocked at 2.10 GHZ) and 4.00 GB (DDR3) memory. If the test problems are big enough, clusters at the CSCAMM will be used. Ax and A^Tx can be defined as function calls instead of direct matrix-vector multiplication in order to save memory allocation.

4 Selection of Databases

Databases are not needed for this stage of testing yet.

5 Validation Principles

If $\mathbf{x}_{\ell_1,\lambda}$ is a solution of (7), then $\mathbf{x}_{\ell_1,\lambda}$ and $\mathbf{r}(\mathbf{x}_{\ell_1,\lambda})$ form an extremal pair by theorem 1.2. They satisfy (9) and (10) by theorem 1.2. As it is suggested in [3] and [21], we are to validate the codes using $m = 1024$, $n = 4096$. The original signal $\mathbf{x}_*$ has 160 non-zero entries, and these entries are randomly filled with ± 1's, with locations are also unknown. Meanwhile the matrix A is generated first by filling the entries a_{ij} with independent samples of a standard Gaussian distribution and then orthonormalizing the rows (so that A would satisfy the Restricted Isometry Property). Therefore the solution $\mathbf{x}_{\ell_1}$ from (5) would be the same as $\mathbf{x}_*$ (and as well as $\mathbf{x}_{\ell_0}$). We will take an $\lambda = \frac{1.5}{||A^T\mathbf{b}||_{\ell_\infty}}$, and the measurement vector $\mathbf{b}$ is corrupted with noise, hence $\mathbf{b} = A\mathbf{x}_* + \zeta$, where ζ is a white Gaussian noise of a variance σ^2. σ various from 0, 10^{-2}, 10^{-1} upto 1.

5.1 Validation Results

In order to validate the codes, we investigate the following quantities,

$$\begin{aligned}
\text{diff}_1 &= \langle \mathbf{x}, A^T\mathbf{r}(\mathbf{x})\rangle_{\ell_2} - ||\mathbf{x}||_{\ell_1}||A^T\mathbf{r}(\mathbf{x})||_{\ell_\infty} \\
\text{diff}_2 &= ||A^T\mathbf{r}(\mathbf{x})||_{\ell_\infty} \\
J(\mathbf{x}) &= \tau||\mathbf{x}||_{\ell_1} + \frac{1}{2}||\mathbf{r}(\mathbf{x})||^2_{\ell_2} \\
\mathbf{r}(\mathbf{x}) &= \mathbf{b} - A\mathbf{x}
\end{aligned}$$

We hope to show that diff$_1$ and diff$_2$ are decreasing when we decrease the the tolerance tolP for the GPSR algorithms and xTol and gTol for the FPC algorithm (regardless of the noise level); meanwhile for the HRSS algorithm, we want to show that the residual $\mathbf{r}(\mathbf{x}_{HRSS})$ is going down like $\mathcal{O}(\frac{1}{\lambda})$. We also show the convergence rate of $\mathbf{x}_{\ell_1,\lambda} \to \mathbf{x}_{\ell_1}$ should behave like $\mathcal{O}(\frac{1}{\lambda})$.

tolP	diff$_1$	diff$_2$	Num. of Iter.	$J(\mathbf{x})$
10^{-4}	$-2.4644e-003$	$2.0224e-005$	27	6.7937
10^{-5}	$-2.2962e-004$	$1.8994e-006$	33	6.793661
10^{-6}	$-2.1692e-005$	$1.8028e-007$	39	6.7937
10^{-7}	$-2.0270e-006$	$1.6894e-008$	45	6.7937

Table 1: Result with $\sigma = 0$ for GPSR Basic

tolP	diff$_1$	diff$_2$	Num. of Iter.	$J(\mathbf{x})$
10^{-4}	$-2.3277e-003$	$1.9723e-005$	31	6.7937
10^{-5}	$-2.0638e-004$	$1.8571e-006$	37	6.793661
10^{-6}	$-2.5418e-005$	$2.1726e-007$	43	6.7937
10^{-7}	$-2.2379e-006$	$2.0232e-008$	49	6.7937

Table 2: Result with $\sigma = 0$ for GPSR Barzilai Borwein

xTol	gTol	diff$_1$	diff$_2$	Num. of Iter.	$J(\mathbf{x})$
10^{-4}	10^{-2}	$-3.4831e-002$	$3.4609e-004$	50	$9.9499e-001$
10^{-5}	10^{-3}	$-3.5118e-003$	$3.3977e-005$	79	$9.9311e-001$
10^{-6}	10^{-4}	$-3.7327e-004$	$3.5609e-006$	107	$9.9298e-001$
10^{-7}	10^{-5}	$-4.0284e-005$	$3.8085e-007$	134	$9.9297e-001$

Table 3: Result with $\sigma = 0$ for FPC

tolP	diff$_1$	diff$_2$	Num. of Iter.	$J(\mathbf{x})$
10^{-4}	$-2.0793e-003$	$2.2583e-005$	28	$6.8717e+000$
10^{-5}	$-1.9433e-004$	$1.6418e-006$	35	$6.871703e+000$
10^{-6}	$-2.1573e-005$	$1.8249e-007$	41	$6.871703e+000$
10^{-7}	$-2.3920e-006$	$2.0249e-008$	47	$6.871703e+000$

Table 4: Result with $\sigma = 10^{-2}$ for GPSR Basic

tolP	diff$_1$	diff$_2$	Num. of Iter.	$J(\mathbf{x})$
10^{-4}	$-1.8890e-003$	$2.0618e-005$	32	$6.871703e+000$
10^{-5}	$-2.1573e-004$	$2.3396e-006$	38	$6.871703e+000$
10^{-6}	$-2.4711e-005$	$2.6693e-007$	44	$6.871703e+000$
10^{-7}	$-2.3235e-006$	$1.9801e-008$	51	$6.871703e+000$

Table 5: Result with $\sigma = 10^{-2}$ for GPSR Barzilai Borwein

xTol	gTol	diff$_1$	diff$_2$	Num. of Iter.	$J(\mathbf{x})$
10^{-4}	10^{-2}	$-3.4324e-002$	$3.3979e-004$	53	$9.9717e-001$
10^{-5}	10^{-3}	$-3.3737e-003$	$3.2354e-005$	81	$9.9835e-001$
10^{-6}	10^{-4}	$-3.4371e-004$	$3.2333e-006$	109	$9.9823e-001$
10^{-7}	10^{-5}	$-4.1654e-005$	$3.8828e-007$	134	$9.9822e-001$

Table 6: Result with $\sigma = 10^{-2}$ for FPC

6 Testing

6.1 Compressed Sensing Cases

In the case of Compressed Sensing testing, we can have some pre-processing information about the parameters as suggested in [20] for the FPC algorithm. Let $\delta = \frac{m}{n}$ and $\gamma = \frac{l}{m}$. Then we

tolP	diff_1	diff_2	Num. of Iter.	$J(\mathbf{x})$
10^{-4}	$-2.5181e-003$	$1.3017e-005$	65	$1.1393e+001$
10^{-5}	$-2.4369e-004$	$1.2570e-006$	93	$1.1393e+001$
10^{-6}	$-2.6812e-005$	$1.3817e-007$	121	$1.1393e+001$
10^{-7}	$-2.6666e-006$	$1.3707e-008$	151	$1.1393e+001$

Table 7: Result with $\sigma = 10^{-1}$ for GPSR Basic

tolP	diff_1	diff_2	Num. of Iter.	$J(\mathbf{x})$
10^{-4}	$-1.6146e-003$	$1.3495e-005$	92	$1.1393e+001$
10^{-5}	$-1.3726e-004$	$1.1382e-006$	136	$1.1393e+001$
10^{-6}	$-1.5475e-005$	$1.2528e-007$	180	$1.1393e+001$
10^{-7}	$-1.7440e-006$	$1.3996e-008$	224	$1.1393e+001$

Table 8: Result with $\sigma = 10^{-1}$ for GPSR Barzilai Borwein

xTol	gTol	diff_1	diff_2	Num. of Iter.	$J(\mathbf{x})$
10^{-4}	10^{-2}	$-6.5959e-002$	$4.4425e-004$	74	$9.9878e-001$
10^{-5}	10^{-3}	$-6.9510e-003$	$4.5291e-005$	165	$9.9982e-001$
10^{-6}	10^{-4}	$-6.9640e-004$	$4.4395e-006$	276	$9.9978e-001$
10^{-7}	10^{-5}	$-7.2050e-005$	$4.5526e-007$	387	$9.9977e-001$

Table 9: Result with $\sigma = 10^{-1}$ for FPC

Num of λ Iter.	$\|\|\mathbf{r}(\mathbf{x}_{HRSS})\|\|_{\ell_2}$	ratio
1	$6.2953e+000$	
2	$5.3806e+000$	$1.1700e+000$
3	$3.1649e+000$	$1.7001e+000$
4	$1.6012e+000$	$1.9766e+000$
5	$8.2591e-001$	$1.9387e+000$

Table 10: Result with $\sigma = 0$ for HRSS

Num of λ Iter.	$\|\|\mathbf{r}(\mathbf{x}_{HRSS})\|\|_{\ell_2}$	ratio
1	$6.3351e+000$	
2	$5.4130e+000$	$1.1704e+000$
3	$3.2022e+000$	$1.9535e+000$
4	$1.6393e+000$	$1.8641e+000$
5	$8.7938e-001$	$1.7647e+000$

Table 11: Result with $\sigma = 10^{-2}$ for HRSS

can pick $\eta = \min(1 + 1.665(1-\delta), 1.999)$, $\mathbf{x}^{(0)} = \eta A^T\mathbf{b}$, $\lambda_1 = \theta||\mathbf{x}^{(0)}||_\infty$ (where $0 < \theta < 1$ is a user-defined constant), set $\lambda_i = \min(\lambda_1\omega^{i-1}, \bar{\lambda})$ (where $\omega > 1$ and L is the first integer i such that $\lambda_L \geq \bar{\lambda})$[7]. As it was done in [20], we will take xtol $= 10^{-4}$ and gtol $= 10^{-2}$. We will also try the algorithm with A being a DCT transform matrix.

[7] [20] also suggests $\theta = 0.99$ and $\omega = 4$ or $\theta = 0.9$ and $\omega = 2$

Num of λ Iter.	$\|\|\mathbf{r}(\mathbf{x}_{HRSS})\|\|_{\ell_2}$	ratio
1	$7.3770e+000$	
2	$6.4127e+000$	$1.1504e+000$
3	$4.2269e+000$	$1.5171e+000$
4	$2.4559e+000$	$1.7211e+000$
5	$1.3333e+000$	$1.8420e+000$

Table 12: Result with $\sigma = 10^{-2}$ for HRSS

λ	$\|\|\mathbf{x}_{\ell_1}\|\|_{\ell_1} - \|\|\mathbf{x}_{\ell_1,\lambda}\|\|_{\ell_1}\|$	ratio
$2.0869e+000$	$1.5700e+002$	
$4.1738e+000$	$1.3911e+002$	$1.1286e+000$
$8.3476e+000$	$8.3440e+001$	$1.6672e+000$
$1.6695e+001$	$4.1722e+001$	$1.9999e+000$
$3.3390e+001$	$2.0861e+001$	$2.0000e+000$

Table 13: Convergence of $\mathbf{x}_{\ell_1,\lambda} \to \mathbf{x}_{\ell_1}$ with GPSR Basic

6.1.1 Compressed Sensing Test Results

The setting for Compressed Sensing are the same as in sec 5. And the Hierarchical Reconstruction code would do only 4 steps of λ-loop.

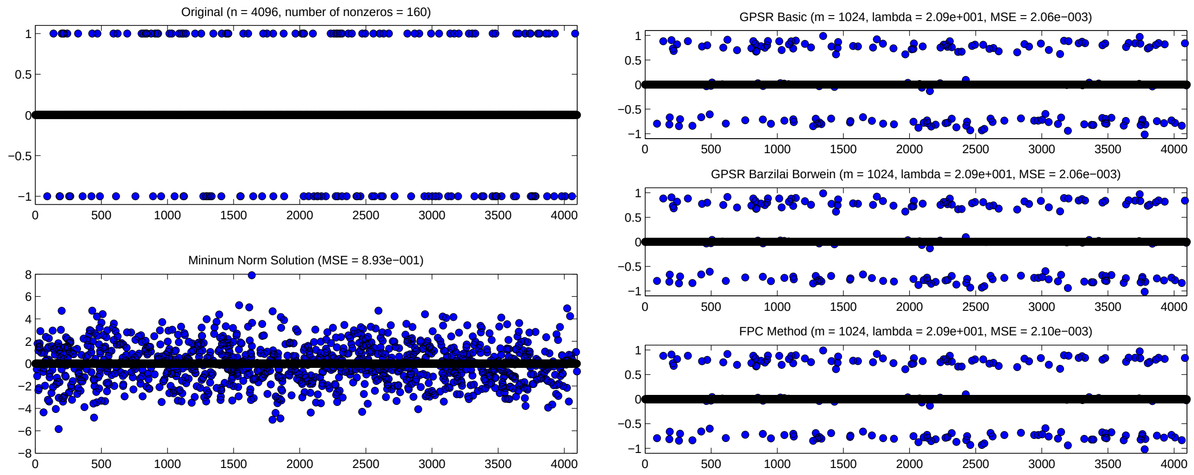


Figure 2: True Signal, Minimum Norm Soln., One Scale Approx.

Stopping at the same λ as the other one scale solvers, HRSS is able to recover the signal better, the error is roughly 0.25 of those from the single scale solvers. However, since the solutions recovered from the single scale solver contain enough support of $\mathbf{x}_*$, the de-biasing step is able to recover exactly the original signal. We will look into possible refinement of the HRSS algorithm in order to compensate the difficulties of eliminating points near 0. Moreover for the DCT transformation case, HRSS would do slightly better than the single scale solvers, and the de-biasing step would not produce better result, since the recovered solutions are not within the support of $\mathbf{x}_*$.

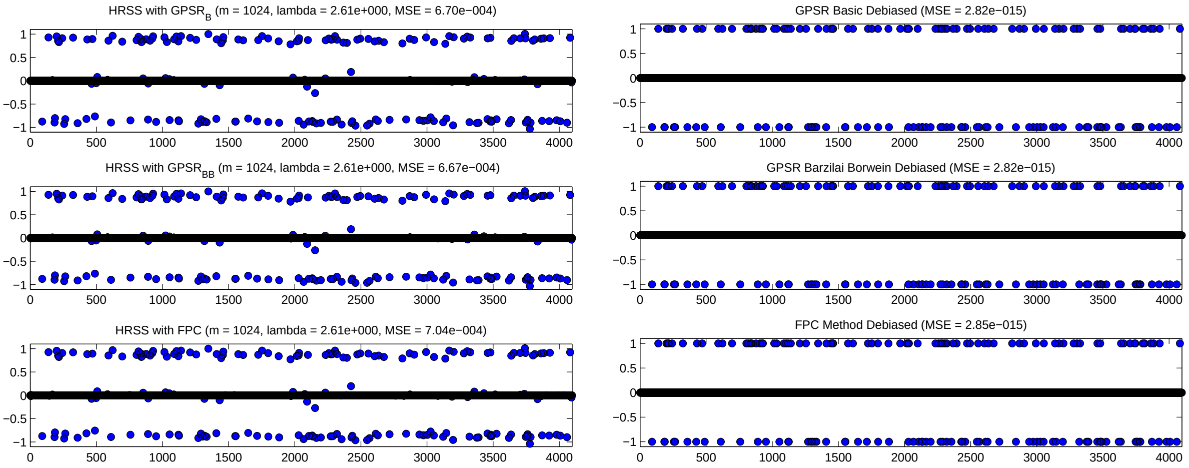


Figure 3: HRSS Approx. and Debaised Soln's

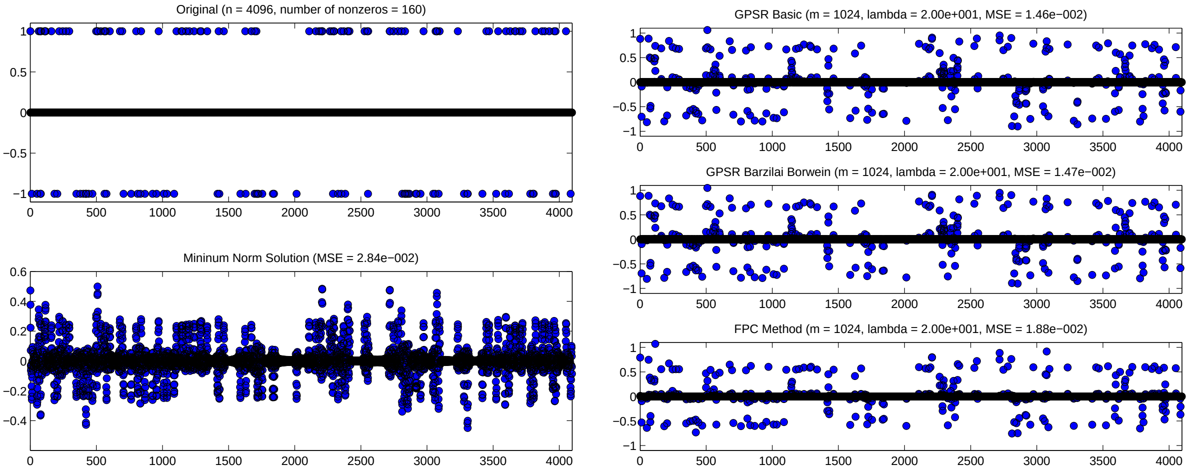


Figure 4: True Signal, Minimum Norm Soln., One Scale Approx. with DCT

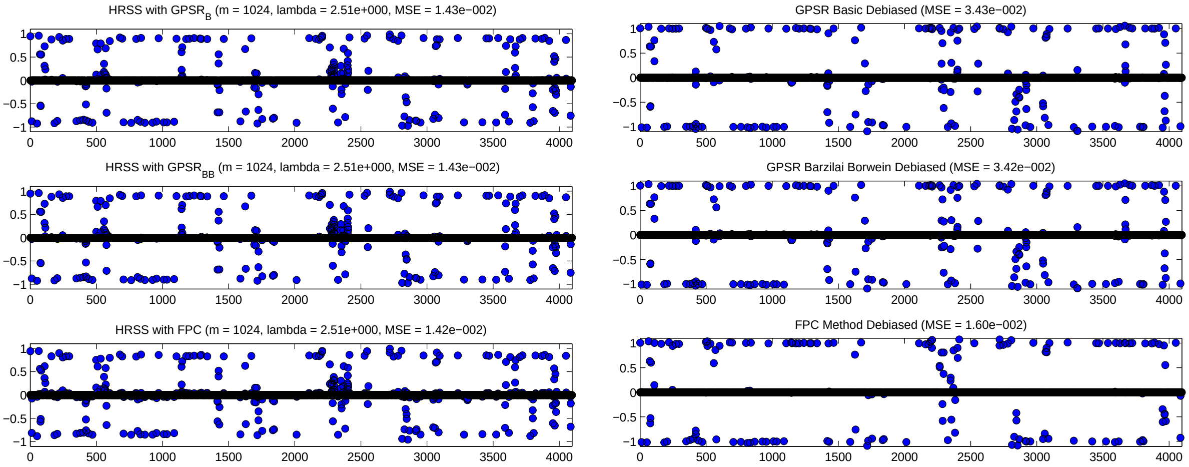


Figure 5: HRSS Approx. and Debaised Soln's with DCT

6.2 The Reconstruction Process of HRSS

Let us also look at what HRSS would generate after each iterate:

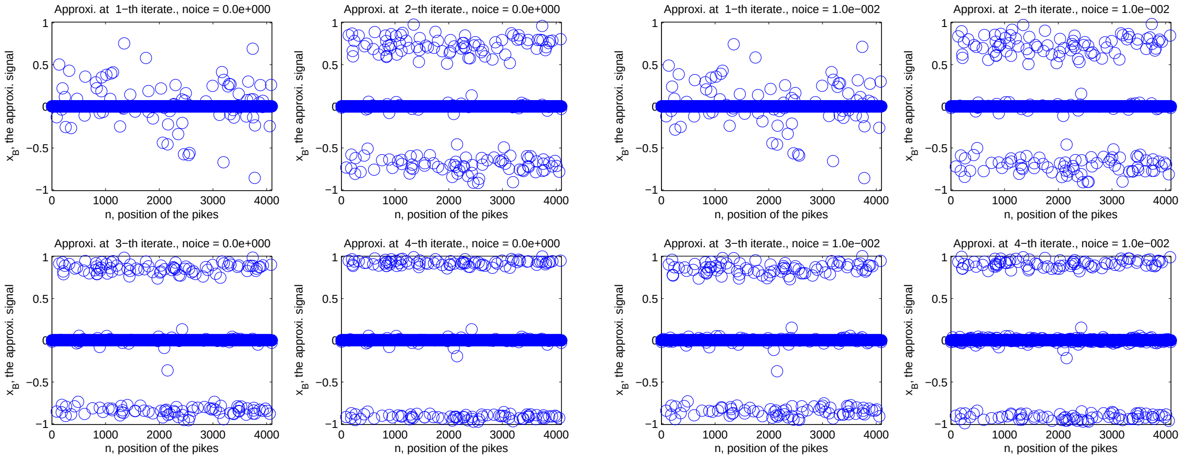


Figure 6: HRSS Approx. with 2 Different Noise Levels

As seen from these plots, the HRSS code gradually recover the approximations with more and more support, and push those points towards ± 1. At $\sigma = 0.01$, the noise has not made significant impact on the approximation. We want to look at the effect of the total number of λ refinements on the noise,

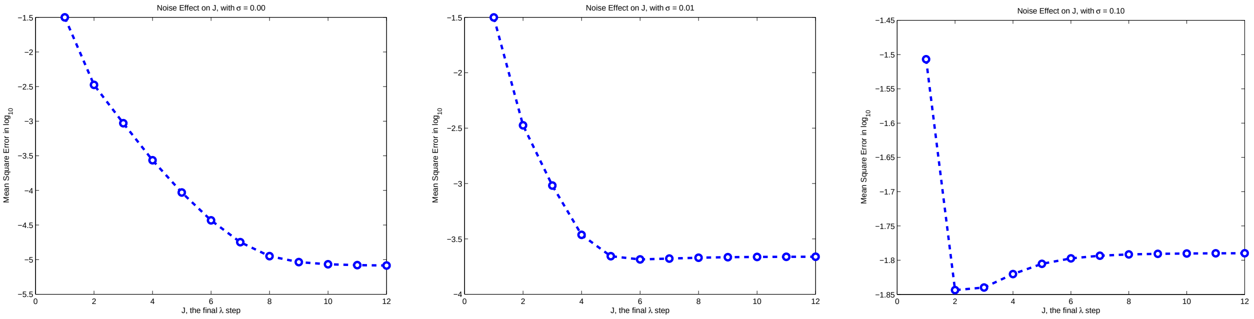


Figure 7: HRSS Approx. with 2 Different Noise Levels

The curve would level off after certain J's, especially when the noise level is large, around $\mathcal{O}(0.1)$, then we should stop the algorithm after 2 steps. We hope to develop automatic check on stopping J based on the noise level in future research.

7 Project Phases and Time lines

We will follow the following time lines for my project:

1. $08/29/2012$ to $10/05/2012$, Project Background Research, Project Proposal Presentation and Report.

2. $10/06/2012$ to $11/21/2012$, Implementation of the GPSR algorithm and validation by theorem 1.2.

3. $11/22/2012$ to $12/20/2012$, Implementation of the FPC algorithm and preparation for mid-year presentation and report.

4. $12/21/2012$ to $01/22/2013$, Validation of the FPC algorithm by theorem 1.2 and Convergence Rate study of GPSR.
5. $01/23/2013$ to $02/22/2013$, Implementation of the whole HRSS algorithm.
6. $02/23/2013$ to $03/16/2013$, Validation of the HRSS algorithm by theorem 1.2.
7. $03/17/2013$ to $05/14/2013$, Final Testing phase, more theories developed and preparation for end-of-year presentation and report.

8 Milestones

Here are major milestones about the project:

1. Presentation given on $10/02/2012$ and Project Proposal due on $10/05/2012$.
2. Implementation of the GPSR algorithm finished and debugged on $11/05/2012$, validation finished on $11/21/2012$.
3. Preparation given on $12/11/2012$, report due on $12/14/2012$, FPC implementation progress started.
4. Implementation of FPC done by $12/21/2012$, debugged and validated by $01/22/2013$.
5. Implementation of HRSS done by $02/22/2013$, Near-Completion Presentation done on $03/07/2013$.
6. HRSS validated by $03/22/2013$, theoretical results obtained by $04/22/2013$.
7. More test results obtained by $04/30/2013$, End-of-year presentation done on $05/07/2013$.
8. Final Projection Documentation and HRSS package due by $05/14/2013$.

9 Deliverables

MATLAB$^{\text{TM}}$codes of HRSS, presentation slides (proposal presentation, mid-year presentation, near-completion, end-of-year presentation), the complete project document, test databases (if any), and test results (both in text file and/or figures) will be delivered at the end of this year long sequence.

References

[1] J. Barzilai and J. Borwein, *Two-point step size gradient methods*, IMA Journal of Numerical Analysis, 8 (1988), pp. 141 – 148.

[2] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Boston, 2^{nd} ed., 1999.

[3] E. Candes and J. Romberg, *A collection of matlab routines for solving the convex optimization programs central to compressive sensing*.

[4] E. Candes, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Transactions on Information Theory, 52 (2004), pp. 489 – 509.

[5] ——, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics, 59 (2006), pp. 1207 – 1223.

[6] E. Candes and T. Tao, *Decoding by linear programming*, IEEE Transactions on Information Theory, 51 (2005), pp. 4203 – 4215.

[7] ——, *Near optimal signal recovery from random projections: Universal encoding strategies?*, IEEE Transactions on Information Theory, 52 (2006), pp. 5406 – 5425.

[8] I. Daubechies, M. Defrise, and C. D. Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Communications in Pure and Applied Mathematics, 57 (2004), pp. 1413 – 1457.

[9] D. L. Donoho, *For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparest near-solution*, technical report, Institute for Computational and Mathematical Engineering, Stanford University.

[10] ——, *For most large underdetermined systems of equations the minimal ℓ_1-norm solution is also the sparest solution*, technical report, Institute for Computational and Mathematical Engineering, Stanford University.

[11] ——, *Unconditional bases are optimal bases for data compression and for statistical estimation*, Applied Computational Harmonic Analysis, 1 (1993), pp. 100 – 115.

[12] ——, *Sparse components of images and optimal atomic decomposition*, Constructive Approximation, 17 (2001), pp. 353 – 382.

[13] ——, *Compressed sensing*, IEEE Transactions on Information Theory, 52 (2004), pp. 1289 – 1306.

[14] D. L. Donoho and M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization*, PNAS, 100 (2003), pp. 2197 – 2202.

[15] D. L. Donoho and Y. Tsaig, *Fast solution of ℓ_1-norm minimization problems when the solution may be sparse*, technical report, Institute for Computational and Mathematical Engineering, Stanford University.

[16] M. A. Figueiredo and R. D. Nowak, *An em algorithm for wavelet-based image restoration*, IEEE Transactions on Image Processing, 12 (2003), pp. 906 – 916.

[17] M. A. T. FIGUEIREDO, R. D. NOWAK, AND S. J. WRIGHT, *Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems*, IEEE Journal of Selected Topics in Signal Processing, 1 (2007), pp. 586 – 597.

[18] J. J. FUCHS, *More on sparse representations in arbitrary bases*, IEEE Transactions on Information Theory, (2004), pp. 1341 – 1344.

[19] R. GRIBONVAL AND M. NIELSEN, *Sparse representations in unions of bases*, IEEE Transactions on Information Theory, 49 (2003), pp. 3320 – 3325.

[20] E. T. HALE, W. YIN, AND Y. ZHANG, *A fixed-point continuation method for ℓ_1-regularized minization with applications to compressed sensing*, Technical Report TR07 - 07, CAAM.

[21] S.-J. KIM, K. KOH, M. LUSTIG, S. BOYD, AND D. GORINEVSKY, *An interior-point method for large-scale ℓ_1-regularized least squares*, IEEE Journal of Selected Topics in Signal Processing, 1 (2007), pp. 606 – 617.

[22] B. K. NATARAJAN, *Sparse approximate solutions to linear systems*, SIAM Journal on Computing, 24 (1995), pp. 227 – 234.

[23] S. OSHER, M. BURGER, D. GOLDFARB, J. XU, AND W. YIN, *An iterative regularization method for total variation-based image restoration*, Multiscale Modeling & Simulation, 4 (2005), pp. 460 – 489.

[24] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, 1970.

[25] T. SERAFINI, G. ZANGHIRATI, AND L. ZANNI, *Gradient projection methods for quadratic programs and applications in training support vector machines*, Optimization Methods and Software, 20 (2004), pp. 353 – 378.

[26] E. TADMOR, S. NEZZAR, AND L. VESE, *A multiscale image representation using hierarchical (bv, $\mathcal{L}^2$) decomposition*, Multiscale Modeling & Simulation, 2 (2004), pp. 554 – 579.

[27] ——, *Multiscale hierarchical decomposition of images with applications to deblurring, denoising and segmentation*, Communications in Mathematical Sciences, 6 (2008), pp. 281 – 307.

[28] A. N. TKHONOV, *On the stability of inverse problems*, Doklady Akademii Nauk SSSR, 39 (1943), pp. 195 – 198.

[29] ——, *Solution of incorrectly formulated problems and the regularization method*, Doklady Akademii Nauk SSSR, 151 (1963), pp. 501 – 504.