Locating Sparse Solutions of Underdetermined Linear Systems via the Reweighted ℓ_1-Method

Yunbin ZHAO and Duan LI

School of Matheamtics, University of Birmingham, United Kingdom
http://web.mat.bham.ac.uk/Y.Zhao
(E-mail: y.zhao.2@bham.ac.uk)

March, 2012

Outline

- Introduction
- Reweighted algorithm framework
- Convergence
- Numerical performance
- Conclusions

Introduction: ℓ_0-problem

- Many data types (e.g. signal and image processing) can be sparsely represented.
- Processing tasks handling such data
  - Compression
  - Reconstruction
  - Storing
  - Separation
  - Transmission
  - ...

often amount to the problem:

$$(\ell_0) \qquad \text{Minimize } \{\|x\|_0 : Ax = b\},$$

where A is an $m \times n$ matrix with $m < n$.

Introduction: ℓ_1-problem

- ℓ_0-problem is an NP-hard discrete optimization problem [Natarajan, 1995].
- ℓ_1-norm, i.e.,

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|,$$

is the convex envelope of $\|x\|_0$ over the region $\|x\|_\infty \leq 1$.

- Replacing $\|x\|_0$ by $\|x\|_1$ yields the ℓ_1-minimization:

$$(\ell_1) \qquad \text{Minimize } \{\|x\|_1 : Ax = b\},$$

which is identical to a linear programming (LP) problem.

When does ℓ_1- solves ℓ_0-minimization?

Conditions for the matrix A under which ℓ_0-problem is computationally tractable:

- Spark and Mutual Cohence [Donoho and Elad (2003)]
- Restricted Isometry Property (RIP) [Candès and Tao (2005)]
- Null Space Property (NSP) [Cohen et al (2009), Zhang (2008), etc.]
- Verifiable conditions (Juditski and Nemirovski (2011)]
- Range Space Property [Zhao (2012a, 2012b)]

Introduction: Reweighted ℓ_1-minimization

Reweighted ℓ_1-minimization)

S1. Choose $x^0 \in R^n$ be an initial point.

S2. Define the weight w^k which is determined by x^k. Then

$$x^{k+1} = \arg\min\{\|W^k x\|_1 : Ax = b\},$$

where $W^k = \mathrm{diag}(w^k)$.

S3. Use x^{k+1} to define W^{k+1} Repeat S2.

Numerical experiments indicate that the reweighted ℓ_1-minimization does outperform ℓ_1-minimization in many situations [Candès, et al (2008), ...]

Examples: Reweighted ℓ_1-minimization

- (Candès, Wakin and Boyd (2008)) The method (CWB) with

$$w_i^k = \frac{1}{|x_i^k| + \varepsilon}, \;\; i = 1, ..., n,$$

- (Foucart and Lai (2009), etc) Weight

$$w_i^k = \frac{1}{(|x_i^k| + \varepsilon)^{1-p}}, \;\; i = 1, ..., n, \tag{1}$$

  where $p \in (0, 1)$ is a given parameter,

The understanding of reweighted ℓ_1-minimization remains very incomplete so far. Even the convergence property of the CWB algorithm remains unclear at present.

A unified framework of reweighted ℓ_1-method

Definition. A function from R^n to R is said to be a *merit function for sparsity* if it is an approximation to $\|x\|_0$ in some sense.

Examples:

- $\|x\|_1$. (convex approximation of $\|x\|_0$ over $\|x\|_\infty \leq 1$)
- $\|x\|_p$, $p \in (0, 1)$. ($\|x\|_p^p \to \|x\|_0$ as $p \to 0$)

- There exist a vast number of merit functions for sparsity. Minimizing such a function may drive the variable x to become sparse when a sparse solution exists.
- From a computational point of view, a merit function for sparsity should admit certain desired properties such as **convexity** or **concavity**.

A unified framework of reweighted ℓ_1-method

Separable concave merit functions:

- $$F(x) = \sum_{i=1}^{n} \phi_i(|x_i|),$$
  where $\phi_i : R_+ \to R$ is called the kernel function.
- Replacing $|x_i|$ by $|x_i| + \varepsilon$ where $\varepsilon > 0$ yields
  $$\min \left\{ F_\varepsilon(x) = \sum_{i=1}^{n} \phi_i(|x_i| + \varepsilon) : Ax = b \right\}.$$

Notation:

- For a subset $S \subseteq \{1, ..., n\}$, $F(x_S)$ is the reduced function, $F(x_S) := \sum_{i \in S} \phi_i(|x_i|)$.
- $f : R^n \to R_+$ is said to be coercive in the region $D \subseteq R^n$ if $f(x) \to \infty$ as $\|x\| \to \infty$ and $x \in D$.

A unified framework

For any given $\varepsilon > 0$, consider the merit function $F_\varepsilon : R^n \to R$ satisfies all the following properties:

Assumption:

(a) *$F_\varepsilon(x) = F_\varepsilon(|x|)$ for any $x \in R^n$, and $F_\varepsilon(x)$ is separable in x, and twice continuously differentiable with respect to $x \in R^n_+$.*

(b) *In R^n_+, $F_\varepsilon(x)$ is strictly increasing with respect to every component x_i and ε, and for any given $\gamma > 0$ there exists a finite number $Q(\gamma)$ such that for any $S \subseteq \{1, ..., n\}$, $g(x_S) := \inf_{\varepsilon \downarrow 0} F_\varepsilon(x_S) \geq Q(\gamma)$ provided $x_S \geq \gamma e_S$, and $g(x_S)$ is coercive in the set $\{x_S : x_S \geq \gamma e_S\}$.*

Assumption (continued)

(c) *In* R^n_+, *the gradient satisfies that* $\nabla F_\varepsilon(x) \in R^n_{++}$ *for any* $(x, \varepsilon) \in R^n_+ \times R_{++}$, $[\nabla F_\varepsilon(x)]_i \to \infty$ *as* $(x_i, \varepsilon) \to 0$, *and for any given* $x_i > 0$ *the component* $[\nabla F_\varepsilon(x)]_i$ *is continuous in* ε *and tends to a finite positive number as* $\varepsilon \to 0$.

(d) *In* R^n_+, *the Hessian satisfies that* $y^T \nabla^2 F_\varepsilon(x) y \leq -C(\varepsilon, r)\|y\|^2$ *for any* $y \in R^n$ *and* $x \in R^n_+$ *with* $\|x\| \leq r$, *where* $r > 0$ *and* $C(\varepsilon, r) > 0$ *are constants, and* $C(\varepsilon, r)$ *is continuous in* ε *and bounded away from zero as* $\varepsilon \to 0$.

A unified framework

The class $\mathcal{M}$ of merit functions:

$$\mathcal{M} = \{F_\varepsilon : \ F_\varepsilon \text{ satisfies the assumption above for } \varepsilon > 0\}.$$

Problem:

$$\min \left\{ F_\varepsilon(x) = \sum_{i=1}^{n} \phi_i(|x_i| + \varepsilon) : Ax = b \right\}. \tag{2}$$

- For any $F_\varepsilon \in \mathcal{M}$, (2) can be rewritten as

$$\min\{F_\varepsilon(v) : Ax = b, |x| \leq v\} = \min_{(x,v)\in\mathcal{F}} F_\varepsilon(v), \tag{3}$$

where

$$\mathcal{F} = \{(x, v) : Ax = b, |x| \leq v\}.$$

How to handle (2)?

Linearization:

- Given the current point v^k,
  $F_\varepsilon(v) = F_\varepsilon(v^k) + \nabla F_\varepsilon(v^k)^T(v - v^k) + o(\|v - v^k\|)$.
- Thus it makes sense to solve the following problem to generate the next point (x^{k+1}, v^{k+1}) :

$$\begin{aligned}(x^{k+1}, v^{k+1}) &= \arg\min_{(x,v)\in\mathcal{F}} \left\{F_\varepsilon(v^k) + \nabla F_\varepsilon(v^k)^T(v - v^k)\right\} \\ &= \arg\min_{(x,v)\in\mathcal{F}} \nabla F_\varepsilon(v^k)^T v \qquad (4)\end{aligned}$$

which is a linear programming (LP) problem.

A unified framework

- The optimal solution (x^{k+1}, v^{k+1}) of (4) satisfies

$$v^{k+1} = |x^{k+1}| \text{ for all } k \geq 0.$$

- Hence

$$\nabla F_\varepsilon(v^k)^T v^{k+1} = \nabla F_\varepsilon(|x^k|)^T |x^{k+1}| = \left\| \mathrm{diag}\left(\nabla F_\varepsilon(|x^k|)\right) x^{k+1} \right\|_1 .$$

- Therefore, the iterative scheme (4) is nothing but

$$x^{k+1} = \arg\min \left\{ \left\| \mathrm{diag}\left(\nabla F_\varepsilon(|x^k|) \right) x \right\|_1 : \; Ax = b \right\}$$

  which is the **reweighted ℓ_1-minimization** with weight $w^k = \nabla F_\varepsilon(|x^k|) \in R^n_{++}$.

Reweighted ℓ_1-Algorithm (see Zhao and Li (2012)):

S1. Choose α, $\varepsilon_0 \in (0,1)$, and let $(x^0, v^0) \in R^n \times R^n_+$ be an initial point.

S2. At the current iterate (x^k, v^k) with $\varepsilon_k > 0$, compute

$$(x^{k+1}, v^{k+1}) = \arg \min_{(x,v)\in\mathcal{F}} \left(\nabla F_{\varepsilon_k}(v^k)\right)^T v$$

i.e.,

$$x^{k+1} = \arg\min\{\|\mathrm{diag}(\nabla F_{\varepsilon_k}(|x^k|))x\|_1 : Ax = b\}.$$

S3. Set $\varepsilon_{k+1} = \alpha\varepsilon_k$. Replace $(x^k, v^k, \varepsilon_k)$ by $(x^{k+1}, v^{k+1}, \varepsilon_{k+1})$ and repeat S2.

Specific Examples

Example 1 Notice that

$$\left(n + \frac{\sum_{i=1}^{n} \log(|x_i| + \varepsilon)}{-\log \varepsilon}\right) \to \|x\|_0$$

as $\varepsilon \to 0$. This motivates the merit function in $\mathcal{M}$:

$$F_\varepsilon(x) = \sum_{i=1}^{n} \log(|x_i| + \varepsilon).$$

At $x \in R_+^n$, the gradient is given by $\nabla F_\varepsilon(x) = \left(\frac{1}{x_1+\varepsilon}, \ldots, \frac{1}{x_n+\varepsilon}\right)^T \in R_{++}^n$, yielding the well-known Candés, Walkin and Boyd (CWB) reweighted ℓ_1-method (2008) with $w_i = \frac{1}{|x_i|+\varepsilon}, i = 1, \ldots, n$

Examples

Example 2 Note that $\|x\|_p^p \to \|x\|_0$ as $p \to 0$. For a given $p \in (0,1)$, we define

$$F_\varepsilon(x) = \frac{1}{p}\sum_{i=1}^{n}(|x_i| + \varepsilon)^p.$$

At $x \in R_+^n$, the gradient is given by

$$\nabla F_\varepsilon(x) = \left(\frac{1}{(x_1+\varepsilon)^{1-p}}, \dots, \frac{1}{(x_n+\varepsilon)^{1-p}}\right)^T \in R_{++}^n.$$

By this merit function, (4) is exactly the reweighted ℓ_1-method with weight

$$w_i = \frac{1}{(|x_i| + \varepsilon)^{1-p}}$$

which was recently studied by many researchers (e.g. Foucart and Lai (2009), ...).

Examples

Example 3 Let $p \in (0,1)$. It is easy to verify that the following function is in $\mathcal{M}$:

$$F_\varepsilon(x) = \sum_{i=1}^{n} \log\left(|x_i| + \varepsilon + (|x_i| + \varepsilon)^p\right).$$

This merit function yields a reweighted ℓ_1-algorithm with the following weight:

$$w_i = [\nabla F_\varepsilon(|x|)]_i = \frac{p + (|x_i| + \varepsilon)^{1-p}}{(|x_i| + \varepsilon)^{1-p}\left[|x_i| + \varepsilon + (|x_i| + \varepsilon)^p\right]},\ i = 1, \ldots, n,$$

which has not been studied in the literature.

Examples

Example 4

Let p and $q \in (0,1)$ be given. Define

$$F_\varepsilon(x) = \frac{1}{p} \sum_{i=1}^{n} \left(|x_i| + \varepsilon + (|x_i| + \varepsilon)^q \right)^p ,$$

which is in $\mathcal{M}$. The gradient of this function at $x \in R_+^n$ is given by

$$[\nabla F_\varepsilon(x)]_i = \frac{q + (x_i + \varepsilon)^{1-q}}{(x_i + \varepsilon)^{1-q} \left[x_i + \varepsilon + (x_i + \varepsilon)^q \right]^{1-p}}, \ \ i = 1, \ldots, n.$$

Examples

Example 5

Then the following function remains in $\mathcal{M}$:

$$F_\varepsilon(x) = \frac{1}{p}\sum_{i=1}^{n}\left(|x_i| + \varepsilon + (|x_i| + \varepsilon)^2\right)^p .$$

The associated reweighted ℓ_1-minimization with

$$w_i = [\nabla F_\varepsilon(|x|)]_i = \frac{1 + 2(|x_i| + \varepsilon)}{(|x_i| + \varepsilon + (|x_i| + \varepsilon)^2)^{1-p}}, i = 1, \dots, n$$

is also a new algorithm.

Examples

Example 6

$$F_\varepsilon(x) = \sum_{i=1}^{n} \left[\log(|x_i| + \varepsilon) - \frac{1}{(|x_i| + \varepsilon)^p} \right] \tag{5}$$

The associated reweighted ℓ_1-minimization uses the weight:

$$w_i = \frac{1 + (|x_i| + \varepsilon)^p}{(|x_i| + \varepsilon)^{p+1}}, \;\; i = 1, ..., n.$$

Convergence of Algorithm (see Zhao and Li (2012))

Definition (Rang Space Property (RSP)).
Let A be an $m \times n$ matrix with $m \leq m$. A^T is said to satisfy the range space property of order K with a constant $\rho > 0$ if

$$\|\xi_{S_c}\|_1 \leq \rho \|\xi_S\|_1$$

for all sets $S \subseteq \{1, ..., n\}$ with $|S| \geq K$, and for all $\xi \in \mathcal{R}(A^T)$, the range space of A^T.

Related to existing conditions:

- Restricted Isometry Property (RIP) [Candés and Tao(2005]: A has the RIP of order k if there exists a constant $\delta \in (0,1)$ such that

$$(1-\delta)\|z\|_2 \leq \|Az\|_2 \leq (1+\delta)\|z\|_2$$

  for any k-sparse vector $z \in R^n$.

- Null Space Property (NSP) [Cohen et al (2009), Zhang (2008),..]: A has the NSP of order k if there exists a constant $\tau \in (0,1)$ such that

$$\|\eta_S\|_1 \leq \tau\|\eta_{S_c}\|_1$$

  for all the sets $S \subseteq \{1,...,n\}$ with $|S| \leq k$, and any $\eta \in \mathcal{N}(A)$, the null space of A.

Relationship of RSP, RIP and NSP

Proposition Let $m < n$, and let $A \in R^{m\times n}$ and $M \in R^{(n-m)\times n}$ be full-rank matrices satisfying $AM^T = 0$. Then the following holds:

- M has the NSP of order k with constant $\tau \in (0,1)$ if and only if A^T has the RSP of order $(n-k)$ with the same constant $\rho = \tau$.
- If M has the RIP of order k with constant $\delta \in (0,1)$, then A^T has the RSP of order $\left(n - \lfloor \frac{\varrho k}{1+\varrho} \rfloor\right)$ with the constant $\rho = \left(\frac{\lfloor \frac{\varrho k}{1+\varrho} \rfloor}{k - \lfloor \frac{\varrho k}{1+\varrho} \rfloor}\right)^{1/2} \left(\frac{1}{\varrho}\right) < 1$ where $\varrho = (1-\delta)/(1+\delta)$.

Theorem 4.8. *Let $A \in R^{m\times n}$ with $m \leq n$, and assume that A^T has the RSP of order K with constant $\rho > 0$ satisfying $1 + \rho < \frac{n}{K}$. Let $F_\varepsilon \in \mathcal{M}$ and $\{(x^k, v^k)\}$ be generated by Algorithm 3.2. Then*

$$[\sigma(x^k)]_n = \min_{1\leq i\leq n} |x_i^k| \to 0 \ as \ k \to \infty. \tag{6}$$

Theorem 4.10. *Let $A \in R^{m\times n}$ with $m \leq n$. Assume that A^T has the RSP of order K with constant $\rho > 0$ satisfying $1 + \rho < \frac{n}{K}$. Let $F_\varepsilon \in \mathcal{M}$ and the sequence $\{(x^k, v^k)\}$ be generated by Algorithm 3.2. If $\|v^{k+1} - v^k\| \to 0$ as $k \to \infty$, then there is a subsequence $\{x^{k_j}\}$ that converges to a $\lfloor(1 + \rho)K\rfloor$-sparse solution of $Ax = b$ in the sense that $[\sigma(x^{k_j})]_{\lfloor(1+\rho)K+1\rfloor} \to 0$ as $j \to \infty$.*

Theorem 4.12. *Assume that A^T has the RSP of order K with constant $\rho > 0$ satisfying that $1 + \rho < \frac{n}{K}$. Let $F_\varepsilon \in \mathcal{M}$ and $F_\varepsilon(v)$ be bounded below in $(x, v) \in R^n_+ \times R_+$ and $g(x) = \inf_{\varepsilon \downarrow 0} F_\varepsilon(x)$ be coercive in R^n_+. Let $\{(x^k, v^k)\}$ be generated by Algorithm 3.2. Then there is a subsequence $\{x^{k_j}\}$ that converges to a $\lfloor (1 + \rho)K \rfloor$-sparse solution of $Ax = b$ in the sense that $[\sigma(x^{k_j})]_{\lfloor (1+\rho)K+1 \rfloor} \to 0$ as $j \to \infty$.*

Numerical Experiments

The following reweighted algorithms were compared:

(a) Candès-Wakin-Boyd (**CWB**) method:

$$x^{k+1} = \arg\ \min \left\{ \sum_{i=1}^{n} \left(\frac{1}{|x_i^k| + \varepsilon_k} \right) |x_i| : Ax = b \right\}.$$

(b) '**Wl**$_p$' method :

$$x^{k+1} = \arg\min \left\{ \sum_{i=1}^{n} \left(\frac{1}{(|x_i^k| + \varepsilon_k)^{1-p}} \right) |x_i| : Ax = b \right\}, \quad p \in (0,1).$$

(c) '**NW1**' algorithm derived from Example 3.3(ii):

$$x^{k+1} = \arg\min \left\{ \sum_{i=1}^{n} \left(\frac{p + (|x_i^k| + \varepsilon_k)^{1-p}}{(|x_i^k| + \varepsilon_k)^{1-p} \left[|x_i^k| + \varepsilon_k + (|x_i^k| + \varepsilon_k)^p \right]} \right) |x_i| : Ax = b \right\}$$

where $p \in (0,1)$.

(d) ‘**NW2**’ algorithm derived from Example 3.5:

$$x^{k+1} = \arg\min \left\{ \sum_{i=1}^{n} \left(\frac{q + (|x_i^k| + \varepsilon_k)^{1-q}}{(|x_i^k| + \varepsilon_k)^{1-q} \left[|x_i^k| + \varepsilon_k + (|x_i^k| + \varepsilon_k)^q \right]^{1-p}} \right) |x_i| : Ax = b \right.$$

where $p, q \in (0, 1)$.

(e) ‘**NW3**’ algorithm derived from Example 3.5 ($p \in (0, 1/2]$):

$$x^{k+1} = \arg\min \left\{ \sum_{i=1}^{n} \left(\frac{1 + 2(|x_i^k| + \varepsilon_k)}{(|x_i^k| + \varepsilon_k + (|x_i^k| + \varepsilon_k)^2)^{1-p}} \right) |x_i| : Ax = b \right\}.$$

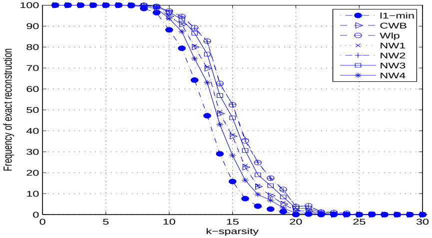
(f) ‘**NW4**’ algorithm based on Example 3.6($p \in (0, \infty)$):

$$x^{k+1} = \arg\min \left\{ \sum_{i=1}^{n} \left(\frac{1 + (|x_i^k| + \varepsilon_k)^p}{(|x_i^k| + \varepsilon_k)^{1+p}} \right) |x_i| : Ax = b \right\}.$$

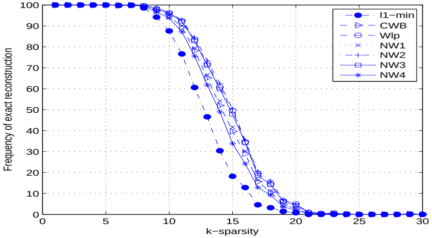
Numerical Experiments

- Randomly generate (A, x) where $A \in R^{50 \times 250}$, and x is a k-sparse vector in R^{250}, and $k = 1, 2, ..., 30$.
- Based on the following assumption: The entries of A and x on its support are i.i.d Gaussian random variables with zero mean and unit variances.
- For every given sparsity k, 500 pairs of (A, x) were generated.
- Every reweighted algorithm was executed only 4 iterations, and $\alpha = 0.5, \varepsilon_0 = 0.01$ and $x^0 = e \in R^{250}$ were used in Algorithm 3.2.
- Given a k-sparse solution x, the algorithm claims to be successful in finding the k-sparse solution x if the found solution x^k satisfies that $\|x^k\|_{\tilde{0}} \leq k$ and $\|x^k - x\| \leq 10^{-5}$. $\|x^k\|_{\tilde{0}}$ is the number of components of x satisfying $|x_i^k| \geq 10^{-5}$.

Numerical Experiments

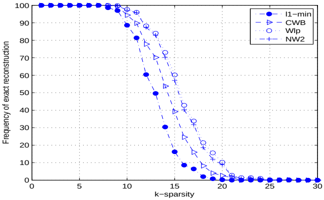


(a) $p = 0.5$

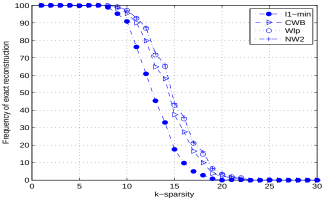


(b) $p = 0.3$

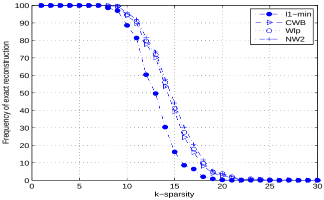
Numerical Experiments



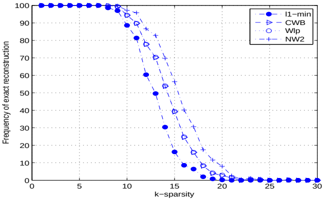
(a) $p = 0.7$



(b) $p = 0.3$



(c) $p = 0.1$



(d) $p = 0.01$

Numerical Experiments

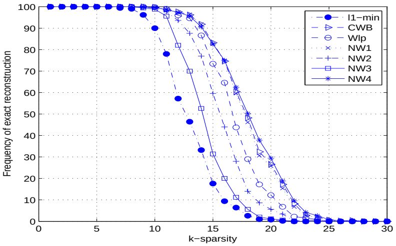
It is interesting to test algorithms using a different parameter updating rule.

Candès, Wakin and Boyd (2008) proposed the following rule:

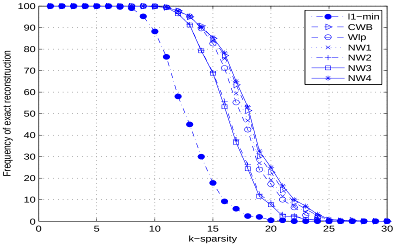
$$\varepsilon_k = \max\left\{[\sigma(x^k)]_{i_0}, 10^{-3}\right\},$$

where i_0 denotes the nearest integer to $m/[4\log(n/m)]$.

Numerical Experiments



(i) $p = 0.5$, ε_k is updated by (32)



(ii) $p = 0.3$, ε_k is updated by (32)

Conclusions

- Through the linearization technique, minimizing the **concave merit functions** for sparsity yields a unified approach for the reweighted ℓ_1-minimization algorithms.
- By this unified approach, we can construct various new specific reweighted ℓ_1-algorithms for the sparse solution of linear systems, and develop **a new and unified convergence theory** for a large family of these algorithms.
- Our convergence analysis is based on the **Range Space Property** (RSP), which is different from the existing RIP/NSP-based analysis.
- As special cases of our general framework, a convergence result for the well-known ℓ_p-**quasi-norm**-based algorithm and **Candès-Wakin-Boyd** method can be obtained.

Conclusions (continued)

- We have proved that, under suitable conditions, a large family of reweighted ℓ_1-algorithms can generate a solution with certain level of sparsity to the linear system.
- Although the simulation shows that reweighted ℓ_1-algorithms outperform the standard ℓ_1-method in many situations, a rigorous mathematical proof for this phenomena has not been carried out so far. This remains an open question in this field.

Key References

- E.J. Candès, M.B. Wakin and S.P. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization, *J. Fourier Anal. Appl.*, 14 (2008), pp. 877-905.
- S. Foucart and M.J. Lai, Sparsest solutions of underdetermined linear systems via ℓ_p-minimization for $0 < q \leq 1$, *Applied and compational Harmonic Analysis*, 26 (2009), pp. 395-407.
- Y.B. Zhao and D. Li, Reweighted ℓ_1-minimization for sparse solutions of underdetermined linear systems, *SIAM J. Optim*, 22 (2012), no.3, pp. 1065-1088.
- Y.B. Zhao, RSP-Based Analysis for Sparsest and Least ℓ_1-Norm Solutions to Underdetermined Linear Systems, Technical report, University of Birmingham, 2012.
- Y.B. Zhao, Equivalence and Strong Equivalence between the Sparsest and Least l1-Norm Nonnegative Solutions of Linear Systems and Their Applications, Technical report, University of Birmingham, 2012.