Postprint

This is the accepted version of a paper presented at 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015.

Citation for the original published paper:

Greedy minimization of l1-norm with high empirical success.
In:

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-163700
Greedy minimization of $\ell_1$-norm with high empirical success

© 2015 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

MARTIN SUNDIN, SAIKAT CHATTERJEE AND MAGNUS JANSSON

Stockholm 2015

Signal Processing
School of Electrical Engineering
Kungliga Tekniska Högskolan
GREEDY MINIMIZATION OF $\ell_1$-NORM WITH HIGH EMPIRICAL SUCCESS

Martin Sundin, Saikat Chatterjee and Magnus Jansson

ACCESS Linnaeus Center, School of Electrical Engineering
KTH Royal Institute of Technology, Stockholm, Sweden
masundi@kth.se, sach@kth.se, magnus.jansson@ee.kth.se

ABSTRACT

We develop a greedy algorithm for the basis-pursuit problem. The algorithm is empirically found to provide the same solution as convex optimization based solvers. The method uses only a subset of the optimization variables in each iteration and iterates until an optimality condition is satisfied. In simulations, the algorithm converges faster than standard methods when the number of measurements is small and the number of variables large.

Index Terms—Convex optimization, basis-pursuit, greedy algorithms.

1. INTRODUCTION

Convex optimization based methods are efficient for finding sparse solutions to underdetermined linear systems of equations. This has received much attention [1–4] and has several applications in signal processing [5–9]. One convex method for finding sparse representations is basis-pursuit [10] (or $\ell_1$-minimization) which finds a sparse solution as

$$\hat{x}_{BP} = \min_{x} \|x\|_1,$$

s.t. $y = Ax$.

where $A \in \mathbb{R}^{m \times n}$ is the dictionary, $y \in \mathbb{R}^m$ is the data/observations and $x \in \mathbb{R}^n$ is the optimization variable. Typically $m \ll n$.

1.1. Prior work

Many efficient methods have been constructed to solve (1) and related noisy problems [11–16]. Methods have also been developed for large scale problems [17–20] using specialized and distributed optimization methods. Another approach to increasing the speed is to start with the full set of variables and then gradually eliminate variables based on screening principles [21, 22].

While solvers for (1) are efficient, they are often slower than greedy search methods [23–26] which in a greedy manner solve the problem

$$\min_{x} \|y - Ax\|_2,$$

s.t. $\|x\|_0 \leq K$,

where the sparsity $\|x\|_0$ is assumed to be known.

To develop fast methods for efficiently finding sparse approximations, it is desirable to combine the effectiveness of convex optimization with the speed of greedy algorithms.

This work was partially supported by the Swedish Research Council under contract 621-2011-5847.

1.2. Our contribution

By using that the basis-pursuit solution is at most $m$-sparse [4] we construct an iterative algorithm, GL1, which performs basis-pursuit (1) in a greedy fashion using only $m$ variables in each iteration. The algorithm replaces vectors in an active set for vectors which lower the $\ell_1$-norm of the solution. This is done until no replacement lowers the $\ell_1$-norm or an optimality condition is met. If the optimality condition is satisfied, then the algorithm recovers the basis-pursuit solution. Even though we were not able to prove convergence, the algorithm has been found to empirically converge to the basis-pursuit solution in all problem realizations.

1.3. Notation

We denote the active set by $I = \{i_1, i_2, \ldots, i_m\} \subset [n] = \{1, 2, \ldots, n\}$ and use $|J|$ to denote the number of elements in a set $J$. We set $A_I = [a_{i_1}, a_{i_2}, \ldots, a_{i_m}]$, $A = A_{[n]}$ and use $x_I$ to denote the vector consisting of the elements of $x$ with indices in $I$.

The support set of a vector $x$ is denoted by $\text{supp}(x) = \{i| x_i \neq 0\}$ and the $l_0$-norm is the size of the support set, i.e. $\|x\|_0 = |\text{supp}(x)|$.

We use $\circ$ to denote the Hadamard (elementwise) product of two vectors and assume that $A$ has full rank and column vectors of unit $l_2$-norm.

2. THE GEOMETRY OF BASIS PURSUIT

Given an at most $m$-sparse vector $x$ such that $\text{supp}(x) = J$, $|J| \leq m$ and $A_Jx_J = y$, we can always enlarge the set $J$ to a set $I$ such that $J \subset I$, $|I| = m$ and $A_I$ is full rank (since $A$ is full rank).

Thus $x_I = A_I^{-1}y$. Since the basis-pursuit solution has at most $m$ non-zero components, this means that (1) is equivalent to

$$I = \arg\min_{|I|=m} \|A_I^{-1}y\|_1, \quad \text{s.t. } A_{I'} \text{ is invertible}$$

(\hat{x}_{BP})_I = A_I^{-1}y, \quad (\hat{x}_{BP})_{I^C} = 0.

We find that (2) is an exhaustive search over all subsets $I \subset [n]$ of size $|I| = m$. However, due to the geometry of the basis-pursuit problem, many subsets can be eliminated from the search.

Let $C(I,s)$ denote the convex cone

$$C(I,s) = \{A_{I}r, r \in \mathbb{R}^{|I|}, r_is_i \geq 0 \forall i \in I\}$$

of $A$, where $|I| \leq m$ and $s = (s_1, s_2, \ldots, s_{|I|})$ with $s_i = \pm 1$ for $1 \leq i \leq m$. We say that a cone $C(I,s)$ is minimal (in $A$) if there is no $j \notin I$ such that $a_j = A_Ix_J$ and $\text{sign}(x'_J) = \pm s_i$. An important property of basis-pursuit is that the solution is always contained in a minimal cone. We formulate this as a proposition.
Proposition 1. The support set of \( \hat{x}_{BP} \) is contained in a minimal cone \( C(I, s) \), where \( \text{supp}(\hat{x}_{BP}) = J \subset I \) and \( \text{sign}(\hat{x}_{jk}) = s_k \) for \( k = 1, 2, \ldots, |J| \).

The proof is given in the Appendix. An illustration in two dimensions is given in Figure 1. In two dimensions the basis-pursuit solution is given by the unique minimal cone. Note that the minimal cone containing \( y \) need not be unique in general. When \( y \) has a sparse representation, then \( y \) lies on the boundary of several minimal cones. By proposition 1 it is sufficient to search over all minimal cones in \( (2) \).

If basis-pursuit recovers an \( m \)-sparse vector \( x \) from measurements \( Ax \), then basis-pursuit recovers any other vector \( x' \) with \( \text{supp}(x') = \text{supp}(x) \) and \( \text{sign}(x') = \text{sign}(x) \) \( (27) \). (i) and (2) gives a solution with \( \text{supp}(\hat{x}_{BP}) = I \) and \( \text{sign}(\hat{x}_{BP})_I = s \) for all \( y \) in the interior of \( C(I, s) \).

3. Greedy \( \ell_1 \)-Minimization

Since basis-pursuit gives a solution with the same support set and sign-pattern for all measurements inside the cone of the basis-pursuit solution, we cannot interchange a vector in the support set for a vector in the complement to lower the \( \ell_1 \)-norm. This can be verified without explicitly computing the new solutions, as the following theorem explains.

Theorem 1. Let \( x \) be an \( m \)-sparse solution to \( y = Ax \), \( I = \text{supp}(x) \) and \( s = \text{sign}(x_j) \). Then the \( \ell_1 \)-norm of the solution cannot be lowered by replacing the \( k \)-th column vector of \( A_I \) for a vector \( a_j \) \((j \notin I)\) if

$$1 \geq \text{sign}(s_k z_k)(s^\top z),$$

where \( z = A_I^{-1} a_j \). Furthermore, if (3) is satisfied with strict inequality for all \( k \) and \( a_j \) \((j \in I)\), then \( |s^\top z| < 1 \) for all \( a_j \) \((j \in I)\) and \( x = \hat{x}_{BP} \).

Noting that we can write

$$s^\top z = s^\top A_I^{-1} a_j = h^\top a_j,$$

where \( h = (A_I^\top)^{-1} s \), we find that if \( |h^\top a_j| \leq 1 \), then no column vector in \( A_I \) can be replaced by \( a_j \) to lower the \( \ell_1 \)-norm. By the theorem, if \( |h^\top a_j| < 1 \) for all \( j \in I \), then \( x = \hat{x}_{BP} \).

3.1. The GL1 Algorithm

Using (3) we construct the greedy algorithm for \( \ell_1 \)-minimization. GL1. The algorithm starts with an initial active set \( I \) and use \( I \) to construct a certificate \( h \). The algorithm searches for candidate vectors in \( I^c \) that satisfy \( |h^\top a_j| > 1 \). The candidate vectors are tested (for all \( k \) such that (3) is violated) if replacing \( a_{jk} \) by the candidate vector lowers the \( \ell_1 \)-norm, i.e. if

$$|\langle x' \rangle_1 | = |\langle A_{I^c} a_j \rangle \hat{x} \rangle |_1 = \left| \hat{x} + (e_k - z) \frac{\hat{x}}{z} \right|_1 < |\langle \hat{x} \rangle |_1, \quad (4)$$

where \( e_k \) denotes the \( k \)-th basis vector in the coordinate basis. This is done until (4) is fulfilled.

The algorithm can get stuck in a local optima (of the algorithm) if it encounters a sparse solution. If this happens, we slightly perturb \( y \) to ensure that it lies in the interior of a cone. When the algorithm terminates, \( y \) is restored and the estimate is computed. The GL1 algorithm can be summarized as follows.

1. Input: \( y, A, I \).
2. Initialization: \( y' = y \).
3. Repeat:
4. \( \hat{x} = A_I^{-1} y', s = \text{sign}(\hat{x}), l_{min} = |\langle \hat{x} \rangle |_1, h = (A_I^\top)^{-1} s \).
5. If \( |\langle \hat{x} \rangle |_1 < m \) perturb \( y' = y + \Delta \cdot A_I 1 \), go to 4.
6. For all \( j \in I^c \) such that \( h^\top a_j > 1 \):
   a. Compute \( z = A_I^{-1} a_j \) and \( t = \text{sign}(s \circ z)(s^\top z) \).
   b. For all \( k \) such that \( t_k > 1 \) and \( |z_k| > \epsilon \):
      i. If \( |\langle \hat{x} + (e_k - z) \frac{\hat{x}}{z} \rangle |_1 < l_{min} \):
         A. \( I \rightarrow (I \cup \{j\}) \{i_k\} \).
         B. \( I^c \rightarrow (I^c \cup \{i_k\}) \{j\} \), go to 4.
    7. If \( |\langle \hat{x} \rangle |_1 = l_{min} \) break.
8. Output: \( I, \hat{x}_l = A_I^\top y' \).

Notes on the algorithm:

- \( \Delta, \epsilon > 0 \) are small constants (e.g. \( 10^{-5} \)).
- Usually, \( \hat{x} \) is not exactly sparse due to numerical errors. Then it is better to perturb as
  $$y' = y + \Delta \cdot As,$$
  where \( s = \text{sign}(\hat{x}) \).
In implementation, matrix inverses are replaced by solving
the set of linear equations, to increase numerical accuracy and
speed.

The runtime can be decreased by selecting the initial set $I$ to
be the $m$ vectors which have largest inner products $|a_i^T y|$.

When determining if the candidate vectors can lower the
$d_1$-norm, it is beneficial to start with the vector with largest inner
product $|h^T a_j|$ and then proceed to the one with next largest
inner product.

We note that the algorithm consists of two parts, determining if
$|h^T a_j| > 1$ and if replacing a column vector with $a_j$ lowers the
d_1-minimization, see Figure 2. Both of these operations can be parallelized,
making the algorithm suitable for large scale problems.

4. NUMERICAL COMPARISON

To evaluate the performance of the algorithm, we compared it with
three other solvers for (1). We used the simplex method [28], $d_1$
magic [12] and the Iterative Reweighted Least Squares (IRLS) [29]
for $d_1$-minimization.

For the Simplex method we reformulate (1) as
\[
\begin{align*}
\min & \ 1^T (x_+ + x_-), \\
\text{s.t.} & \ y = A(x_+ - x_-) \\
& \ x_+, x_- \geq 0
\end{align*}
\]

where $1 \in \mathbb{R}^n$ is a vector of ones and we used MATLAB’s linprog
to run the simplex algorithm. For $d_1$-magic, (1) is reformulated as
\[
\begin{align*}
\min & \ 1^T t \\
\text{s.t.} & \ y = Ax \\
& \ -t_i \leq x_i \leq t_i, \text{ for } i = 1, 2, \ldots, n
\end{align*}
\]

and solved using a primal-dual interior point method [12]. The IRLS
algorithm approximates the $d_1$-norm by a weighted $d_2$-norm which
is updated iteratively [29].

In the simulation we used $n = 8000$ and varied $m$. The
measurements were generated by drawing the elements of $A$ from
$\mathcal{N}(0, 1)$ and normalizing the column vectors, the vector $x$ was
generated as a $[0.25m]$-sparse vector with its non-zero elements
drawn from $\mathcal{N}(0, 1)$. Finally we calculated $y = Ax$. Because $A$ is
Gaussian, $\hat{x}_{BP}$ in (1) is unique (with probability 1).

Because the algorithms have different complexity per iteration,
we measure the total cputime and the $d_1$-norm rather than the num-
ber of iterations. We were not able to access the intermediate times
and $d_1$-norms of linprog and therefore only display the final norm
and cputime. We show the convergence in individual realizations for
$m = 50$ and $m = 100$ in Figure 3 and 4. We find that the sim-
plex method converged faster than $d_1$-magic for $m = 50$, but slower
for $m = 100$. In both realizations, GL1 was the fastest algorithm.

In Figure 5 we show the average cputime for different values of $m$
averaged over 10 realizations of $A$ and $x$. We see that $d_1$-magic is
slower than the simplex method for $m \leq 80$ and slower than GL1
for $m \leq 140$. The simplex method was about 3 times slower than
GL1 for all values of $m$.  

Fig. 3. $d_1$-norm vs. cputime for $n = 8000$ and $m = 50$.

Fig. 4. $d_1$-norm vs. cputime for $n = 8000$ and $m = 100$.

Fig. 5. Mean cputime vs. $m$ for $n = 8000$. 

5. CONCLUSION

In this paper we constructed a greedy algorithm for $\ell_1$-minimization which empirically converges to the basis-pursuit solution. The algorithm interchanges column vectors in an active set until an optimality condition is reached or the $\ell_1$-norm no longer decreases. In simulations, the algorithm converged faster than the simplex method, $\ell_1$-magic and IRLS when the number of measurements is small. Questions remain whether the algorithm always converges to the basis-pursuit solution and if it can be extended to other versions of basis-pursuit.

6. APPENDIX

Proof of proposition 1. We need to show that if the cone $C(I, s)$ contains a vector $a_j$, then we can replace a column vector in $A_I$ for $a_j$ to lower the $\ell_1$-norm of the solution.

Let $I' = (I \cup \{j\}) \setminus \{k\}$ and set

$$y = \sum_{i \in I} x_i a_i, \quad a_j = \sum_{i \in I} z_i a_i, \quad y = x'_j a_j + \sum_{i \in I, i \neq k} x'_i a_i.$$ (5)

Without loss of generality we can assume that $x_i, z_i, x'_i \geq 0$ for all $i \in I \cup \{j\}$. By inserting (6) into (7) we find that

$$x'_j a_j + \sum_{i \in I, i \neq k} x'_i a_i = x'_j z_k a_k + \sum_{i \in I, i \neq k} (x'_i + x'_j z_i) a_i$$

By comparing to (5) we get that when $A_I$ and $A_{I'}$ are full rank, then $x_k = x'_j z_k$ and $x_i = x'_i + x'_j z_i$ for $i \neq k$. Using that

$$1 = |a_j|_2 < \sum_{i \in I} |z_i| \cdot |a_i|_2 = ||z||_1,$$

where we have strong inequality because the column vectors in $A_I$ are not parallel, we find that

$$||x||_1 = \sum_{i \in I} x_i = x'_j z_k + \sum_{i \in I, i \neq k} (x'_i + x'_j z_i)$$

$$= x'_j \sum_{i \in I} z_i + \sum_{i \in I, i \neq k} x'_i$$

$$= x'_j (||z||_1 - 1) + ||x'||_1 > ||x'||_1,$$

provided that $x'_j > 0$.

Proof of Theorem 1. Assume that we cannot interchange the $k$'th column vector of $A_I$ for a vector $a_j$ to get a solution with lower $\ell_1$-norm. Then all measurements in the cone $C(I, s)$ give basis-pursuit solutions with the same support set and sign-pattern, i.e. for all $w \in \mathbb{R}^m$, $w \geq 0$

$$||w||_1 \leq ||A_{I'}^{-1} A_I S w||_1,$$ (8)

where $S = \text{diag}(s)$, $I' = (I \cup \{j\}) \setminus \{i_k\}$ and we assumed that $A_{I'}$ is invertible.

Replacing $a_{i_k}$ for $a_j$ corresponds to making a rank-1 update of $A_I$, i.e.

$$A_{I'} = A_I + (a_j - a_{i_k}) e_k^T.$$

Setting $z = A_{I'}^{-1} a_j$, we get that

$$A_{I'}^{-1} A_I S w = \left( A_I^{-1} - \frac{A_I^{-1} (a_j - a_{i_k}) e_k^T A_{I'}^{-1}}{1 + \sum_{i \in I} \cdot (a_j - a_{i_k}) e_k^T} \right) A_I S w$$

$$= S w - \frac{(a_j - e_k) e_k^T S w}{e_k^T A_I^{-1} a_j} = S w + (e_k - z) \frac{s_k w_k}{z_k}.$$

So if (8) holds, then

$$1 \leq \min_{w \geq 0, \|y\|_1} \|S w + (e_k - z) \frac{s_k w_k}{z_k}\|_1$$

$$= \min_{w \geq 0, \|y\|_1} \frac{w_k}{z_k} + \sum_{i \in I \setminus \{s_k\}} \frac{s_i w_i}{z_i} = 1 + \sum_{i \in I \setminus \{s_k\}} |z_i|$$

$$= 1 + \sum_{i \in J_+} |z_i| \geq 1,$$ (9)

where

$$J_+ = \{ i \in I \setminus \{i_k\}, z_i s_i z_k > 0 \}, \quad J_- = \{ i \in I \setminus \{i_k\}, z_i s_i z_k \leq 0 \}.$$

Rewriting (9) as

$$\sum_{l: z_l s_l z_k > 0} |z_l| \leq 1 + \sum_{l: z_l s_l z_k \leq 0} |z_l|$$

$$\Rightarrow 1 \geq \sum_{i} \text{sign}(z_i s_i z_k) |z_i|$$

$$= \text{sign}(z_k s_k) \sum_{i} s_i z_i = \text{sign}(z_k s_k) (s^T z),$$

we recover the optimality condition in (3).

To show that strict inequality in (3) implies that $|s^T z| < 1$, assume that $|s^T z| \geq 1$, but $\text{sign}(z_k s_k) (s^T z) < 1$ for all $k$. This implies that $\text{sign}(z_k s_k) = -\text{sign}(s^T z)$ for all $k$. However, if all terms $s_i z_i$ have the same sign, then $\text{sign}(z_k s_k) = \text{sign}(s^T z)$, giving a contradiction. Thus, strict inequality in (3) implies that $|s^T z| < 1$.

We find that

$$s^T z = s^T A_{I'}^{-1} a_j = h^T a_j$$

where $h = (A_{I'}^{-1})^s$. This gives that if $|h^T a_j| < 1$ for all $j \in I'$, then $A_{I'}^s h = s$ and $\|A_{I'}^s h\|_{\infty} < 1$. The vector $h$ is thus the dual certificate of the basis-pursuit solution [27], giving us that $x = h^T P$.

\qed
7. REFERENCES