Approximated Function Based Spectral Gradient Algorithm for Sparse Signal Recovery

Weifeng Wang [1, *] and Qiuyu Wang

[1]*College of Mathematics and Information Science, Henan University, China*

Abstract   Numerical algorithms for the l_0-norm regularized non-smooth non-convex minimization problems have recently became a topic of great interest within signal processing, compressive sensing, statistics, and machine learning. Nevertheless, the l_0-norm makes the problem combinatorial and generally computationally intractable. In this paper, we construct a new surrogate function to approximate l_0-norm regularization, and subsequently make the discrete optimization problem continuous and smooth. Then we use the well-known spectral gradient algorithm to solve the resulting smooth optimization problem. Experiments are provided which illustrate that this method is very promising.

1. Introduction

The focus of this paper is on the following structured minimization:

$$\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_0, \tag{1.1}$$

where $\|x\|_0 = \text{card}(x) = \sum_{i=1}^{n} 1(x_i \neq 0)$ denotes the cardinality or the number of nonzero components in x. The true x is often sparse in the sense that many of

---

*Correspondence to: College of Mathematics and Information Science, Henan University, Kaifeng 475000, China. Email:weifengwang2012@gmail.com

its components are zeros. The function $f : R^n \to R$ is continuously differentiable (may be non-convex) and bounded below, and the parameter $\mu > 0$ is used to trade off both terms for minimization. Given its structure, problem (1.1) covers a wide range of apparently related formulations in different scientific fields, such as signal/image processing, compressive sensing, statistics, machine learning, and so on.

To solve (1.1), a convex relaxation which can be solved efficiently is the so-called l_1-regularized least squares:

$$\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_1, \tag{1.2}$$

where $\|x\|_1$ is the sum of the absolute values of each component in x. The compressive sensing theory based on the fact that most signals are compressible, was led by Candès et al. [2, 3], Donoho [7]. It has been shown that under some reasonable conditions, problems (1.1) and (1.2) share the common solution with high probability [6]. For this reason, numerous algorithms to solve problem (1.2) have been proposed, analyzed, and tested in the last decade. For instance, by using an operator splitting technique, Hale, Yin and Zhang derive the iterative shrinkage/thresholding fixed-point continuation algorithm [11]. A closely related method is the fixed-point continuation and active set [16, 17], which solves a smooth subproblem to determine the magnitudes of the nonzero components of x based on an active set. Another closely related method is the sparse reconstruction algorithm [14], which involves minimizing a non-smooth convex problem with separable structures.

Unlike all the methods mentioned above, in this paper, we proposed a novel approximated function method to solve (1.1). More preciously, we construct an approximated smooth function to replace the discrete l_0-norm in (1.1) which based on a small parameter β. We also show that the proposed smooth function trends to the l_0-norm as $\beta \to \infty$. Moreover, we use the well-known spectral gradient method to solve the resulting smooth minimization, and incorporate a non-monotone line search technique to accelerate its convergence. Finally, we do numerical experiments to recover a large sparse signal from its limited measurement, which illustrate that the constructed function works well, and furthermore indicate that the proposed algorithm is encouraging.

The rest of the paper is organized as follows. In section 2, we recall some existing surrogate functions to the l_0-norm and construct our new surrogate function immediately. In section 3, we describe the spectral gradient method for unconstrained optimization problem, and then list the full steps of our algorithm. Furthermore, to illustrate the feasibility and efficiency of the proposed method, we report the numerical experiments in section 4. Finally, the conclusion is drawn in the last section.

2. Motivation and Surrogate Function

This section is devoted to stating our motivations, reviewing some existing surrogate functions to l_0-norm, and constructing a new function. In the mean time, some nice properties of the surrogate functions are also presented.

2.1. *Motivation and ℓ_p-norm function ($0 < p < 1$)*

Instead of the ℓ_0-norm regularized problem (1.1), another very natural improvement is to apply the ℓ_p-norm ($0 < p < 1$) regularization [10], that is, to solve the following problem

$$\min_{x\in\mathbb{R}^n} f(x) := \frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_p^p,, \tag{2.1}$$

where $\|x\|_p$ is the so-called ℓ_p quasi-norm of $\mathbb{R}^n$, defined by

$$\|x\|_p = \Big(\sum_{i=1}^n |x_i|^p \Big)^{1/p}.$$

One can observe that as $p \downarrow 0$, (2.1) approaches the ℓ_0 problem (1.1). On the other hand, as $p \uparrow 1$, (1.2) approaches the ℓ_1-norm problem (1.2). In other words, problem (2.1) is an intermediate between (1.1) and (1.2). Nevertheless, as the ℓ_p-norm regularization is nonconvex, nonsmooth, and non-Lipschitz, the use of the standard ℓ_1-norm minimization problems' solvers are generally precluded. As shown in [9, 4] that finding the global minimizer of (2.1) is strongly Np-hard, but finding a local optimal solution of the problem could be done in polynomial time, and furthermore the lower bounds for the absolute value of nonzero entries in each local optimal solution can be established theoretically [5]. The curve of the ℓ_0-norm and the curves of ℓ_p-norm with different values of p are plotted in Figure 1.

It can be seen from Figure 1 that, as p decreases, the curve of ℓ_p-norm trends to the one of ℓ_0. Hence, it is reasonable to think that the smaller p may yield the more sparse solutions. However, the ℓ_p norm is non-differentiable at original point, which makes it is challenging to minimize. The important observation motivates us to construct an surrogate function which is continuous and differentiable and approximate the ℓ_0-norm to produce sparse solutions.

2.2. *Existing surrogate functions*

First we replace the cardinality term in (1.1) with a continuous and smooth surrogate penalty function $\omega(y)$, which makes the discrete optimization problem become continuous and smooth. The suitable surrogate function must range from 0 to 1 with the property that $w(0) = 0$ and $w(y) = 1$ when y is away from 0. Now,
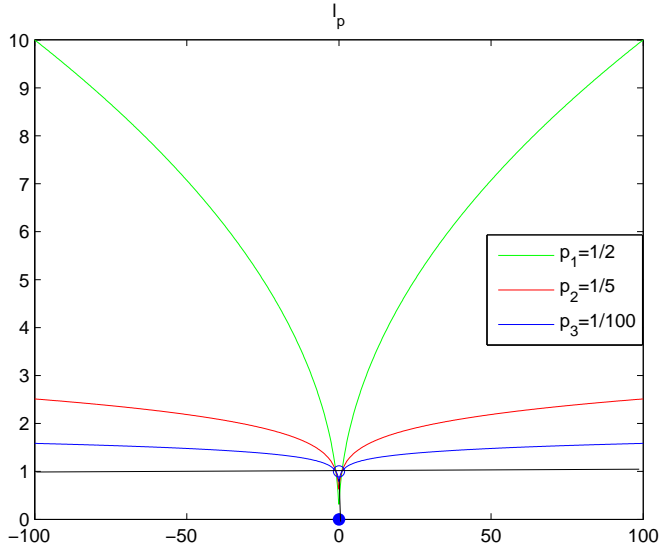
Figure 1. Comparing curves of ℓ_0-norm to ℓ_p-norm with $p = 1/2, 1/5$, and $1/100$.

we recall three well-known choices of the surrogate function which are designed in this literature.

1. Truncated l_γ penalty given by

$$\omega(\beta; \alpha, \gamma) = \begin{cases} (|\beta|/\alpha)^r & \text{if} |\beta| \leq \alpha \\ 1 & \text{otherwise} \end{cases} \tag{2.2}$$

for $\gamma > 0$ and $\alpha > 0$. Typically, $\gamma = 0.5, 1, 2$ can be considered. The magnitude of a controls the sharpness of the approximation.

2. The "weight elimination" penalty (A.S. Weigend, D.E. Rmelhart, and B.A. Huberman 1991) [15] in artificial neural network (ANN) provides another smooth approximation $\omega(x) = \tanh(\alpha \cdot |x|)^\gamma$ with $\gamma = -2$. The constant $\alpha > 0$ can adjust the precision of the approximation.

3. The hyperbolic tangent penalty

$$\omega(x) = \tanh(\alpha \cdot |x|)^\gamma,$$

for $\alpha \geq 0$ and $\gamma \geq 0$. When $\gamma = 2$, $\omega(x)$ is smooth. Its shape is similar to the "weight elimination" penalty, and with a faster progression in its approximation to l_0. It can be seen that a larger α provides a sharper discrimination between 0 and values that are close to 0.

2.3. A new surrogate function

The suitable surrogate function should satisfy at least the following properties. Firstly, the surrogate function should be continuous and smooth. Secondly, its values should be ranged from 0 to 1 with the property that $\omega(0) = 0$ and $\omega(y) = 1$ when y is away from 0. Now, we develop the following function

$$\omega(y) = \ln \frac{e\beta|y|^\gamma + 1}{\beta|y|^\gamma + 1},$$

where the constants $\beta > 0$ and $\gamma > 0$ is fixed. When $\gamma = 2$, the function reduces to

$$\omega(y) = \ln \frac{e\beta y^2 + 1}{\beta y^2 + 1}, \tag{2.3}$$

Clearly, it is continuous and smooth, as graphically shown at the right plot in Figure 2.

Figure 2 shows the curve of ℓ_0 function and the surrogate function (2.3) with different choices of β. As can be seen from the right plot, the curve approximates the ℓ_0 as the value of β increase. In other words, a bigger β may cause a higher precision of its approximation. In our experiments, we show that a small β can produce higher quality re-solutions.

3. Algorithm

In this section, based on the surrogate function as described in the previous section, we discuss the construction of our algorithm. Before stating the steps of our method, we first give a brief description of the well-known spectral gradient method for the following unconstrained optimization

$$\min_{x \in \mathbb{R}^n} f(x), \tag{3.4}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function. The spectral gradient method was originated by Barzilai and Borwein [1], and its iterative form is defined by

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \tag{3.5}$$

where $\nabla f(x_k)$ is the gradient of f at x_k and one of the choices of the scalar λ_k is given by

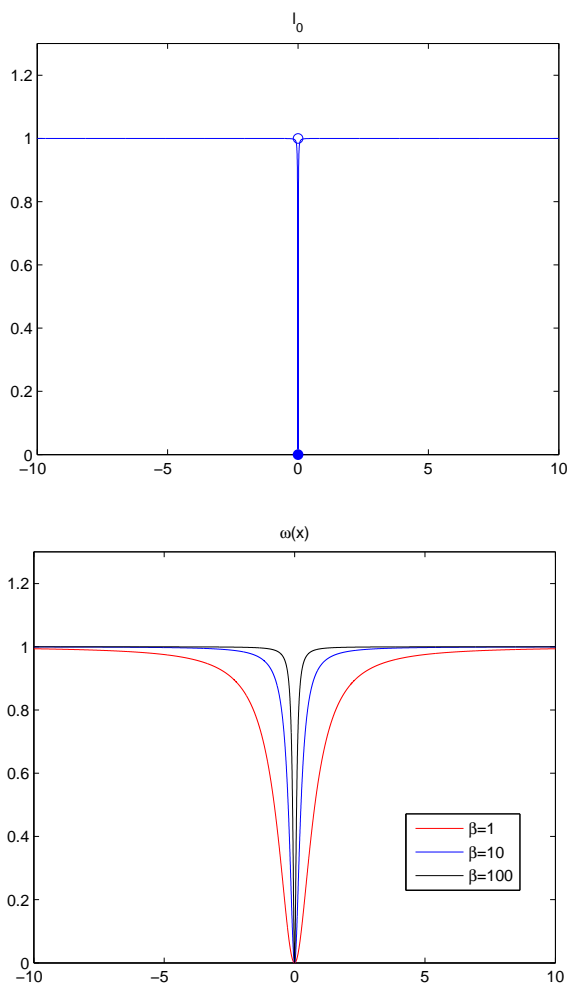$$\lambda_k = \frac{\|s_{k-1}\|_2^2}{y_{k-1}^r \cdot s_{k-1}}. \tag{3.6}$$

Figure 2. Top: the plot of l_0 function; Bottom: the plot of the surrogate function $\omega(y)$ in (2.3) with $\beta = 1, 10, 100$.

where $s_{k-1} = x_k - x_{k-1}$, and $y_{k-1} = \nabla f(x_k) - \nabla f(x_{k-1})$. The spectral gradient method is also named two-points method, Barziali-Borwein gradient method, and is received much attention in past years (see e.g., [12, 13, 8]). Given an initial point x_0, our algorithm generates a sequence x_k by

$$x_{k+1} = x_k + \alpha_k d_k, \quad k = 0, 1, 2 \tag{3.7}$$

where d_k is a descent direction of f at x_k, and α_k is a steplength determined by a line search. In our algorithm, the search direction currently is defined by

$$d_k = -\lambda_k \nabla f(x_k). \tag{3.8}$$

It is easy to see that $d_k^T \nabla f(x_k) < 0$ provided by $\lambda_k > 0$, which shows that d_k is a descent direction at current point. Along the search direction, in our algorithm, the steplength α_k is determined by the nonmonotone line search

$$f(x_k + \alpha_k d_k) \leq \max_{0 \leq j \leq m(k)} f(x_{k-j}) + \delta \alpha_k \nabla f(x_k)^T d_k \tag{3.9}$$

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

Based on the spectral gradient, we pay our attention to constructing our algorithm. Now we replace $\|x\|_0$ in (1.1) with the surrogate function $\omega(x)$, then we get the continuous and smooth function $f(x)$ as

$$f(x) := \frac{1}{2} \|Ax - b\|_2^2 + \mu \sum_{i=1}^{n} \omega(x_i).$$

It is easy to compute that the gradient at x is:

$$\nabla f(x) = A^T \cdot (Ax - b) + \mu \sum_{i=1}^{n} \omega'(x_i),$$

where

$$w'(x_i) = \frac{2\beta e x_i(e-1)}{(\beta e x_i^2 + 1)(\beta x_i^2 + 1)}.$$

To end this section, we list the iterative scheme of the Approximated Function Based Spectral Gradient Algorithm (referred to AFSG) as follows:

Algorithm 3.1 **(AFSG)**

Initialization: Choose x_0 and constant $\mu > 0$. Constants $\tilde{\alpha} > 0$, $\rho \in (0, 1)$, $\delta \in (0, 1)$, $h \in (0, 1]$ and positive integer $\tilde{m}$. Set $k = 0$.
Step 1. Stop if $\|d_k\|_2 = 0$. Otherwise, continue.
Step 2. Compute d_k via (3.8).
Step 3. Compute α_k via (3.9).
Step 4. Let $x_{k+1} = x_k + \alpha_k d_k$.
Step 5. Let $k = k + 1$. Go to Step 1.

Remark 3.1
To ensure λ_k is bounded from 0 and subsequently ensures that d_k is a descent direction per-iteration, the λ_k is forced as

$$\lambda_k = \min\{\lambda_{(\max)}, \max\{\lambda_k, \lambda_{(\min)}\}\},$$

where $0 < \lambda_{(\min)} \ll \lambda_{(\max)} < +\infty$ are fixed scalars.

We clearly see that our proposed algorithm is essentially the standard spectral gradient method for smooth unconstrained minimization. Hence, it global convergence can be followed directly in optimization literature. To end this section, we list the convergence theorem of the proposed algorithm. Its proof can be found in Theorem 2.1 in [13].

Theorem 3.1
Let f be continuously differently and x_k be generated by algorithm AFSG, then $\nabla f(x_k)$ for some k, or $\lim_{k \to \infty} \nabla f(x_k) = 0$.

4. Numerical experiments

In this section, we present numerical results to illustrate the feasibility and efficiency of the proposed method. We use our method to solve l_0-regularized least squares, which mainly appear in compressive sensing. All experiments are performed under Windows 7 and Matlab 7.11 ($R2010b$) running on a Lenovo laptop with an Intel Atom CPU at 2.2 GHz and 2GB of memory.

Test on l_0-regularized least squares. Let $\bar{x}$ be a sparse or a nearly sparse original signal, $A \in \mathbb{R}^{m \times n}$ ($m \ll n$) be a linear operator, $\varepsilon \in \mathbb{R}^m$ be a zero-mean Gaussian white noise, and $b \in \mathbb{R}^m$ be an observation that satisfies the relationship

$$b = A\bar{x} + \varepsilon.$$

We measure the quality of restoration x^* through the relative error to the original signal $\bar{x}$; that is,

$$RelErr = \frac{\|x^* - \bar{x}\|_2}{\|\bar{x}\|_2}.$$

In this test, we use a random matrix A with independent identically distributed Gaussian entries. The ε is the additive Gaussian noise of zero mean and standard deviation σ. Given the storage limitations of the PC, we test a small size signal with $n = 2^{11}$, $m = 2^9$. The original signal contains randomly $p = 2^6$ nonzero elements. We also choose the noise level $\sigma = 10^{-3}$. The proposed algorithm starts at zero point and terminates when the relative change of two successive points are sufficiently small, i.e.,

$$\frac{\|x_k - x_{k-1}\|_2}{\|x_{k-1}\|_2} < tol.$$

In this experiment, we take $tol = 10^{-4}$, $\lambda_{(\min)} = 10^{-20}$, and $\lambda_{(\max)} = 10^{20}$. In the line search, we choose $\bar{\alpha} = 1$, $\rho = 0.5$, $\delta = 10^{-4}$, and $m = 5$. The original signal, the limited measurement, and the reconstructed signal are given in Figure 3.

Comparing the right plot with the left one in Figure 3, we see that the original sparse signal is restored almost completely in sense of RelErr is 0.84%. All the
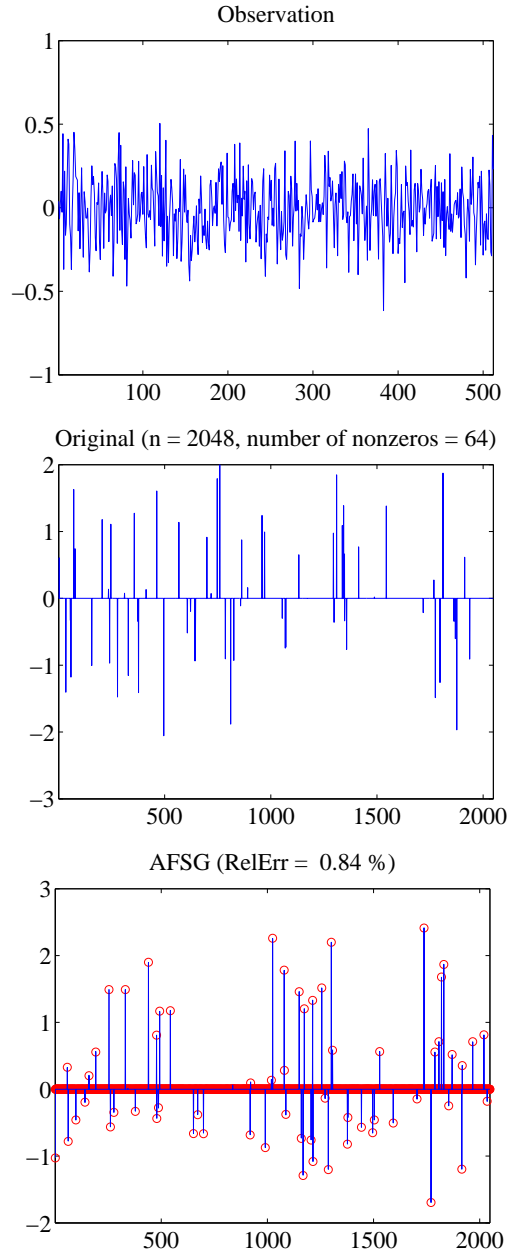
Figure 3. Top: original signal with length 4096 and 64 positive non-zero elements; Middle: the noisy measurement with length 512; Bottom: recovered signal by AFSG (red circle) versus original signal (blue peaks).

blue peaks are encircled by the red circles, which illustrates that the original signal has been found successfully. Overall, this experiment shows that the proposed method performs very well and provides an efficient approach to the recovery of large sparse non-negative signal.

4.1. *Test on AFSG and NBBL1*

In this subsection, we compare AFSG with the solver NBBL1 [18] with respect to relative errors and running time. Parameters' values are set just the same to the previous. The Matlab package of NBBL1 is obtained at URL: http://maths.henu.edu.cn/szdw/teachers/xyh.htm. We repeatedly run each code with different combinations of (n, m, k), and record the computing time (Time) and relative errors (RelErr) in Table I.

Table I. Test results of AFSG and NBBL1 with different combinations of (n, m, k).

			AFSG		NBBL1	
n	m	k	Time	RelErr	Time	RelErr
8192	1938	256	32.4074	2.38e-2	71.2432	2.44e-2
4096	819	128	10.6549	2.35e-2	20.0149	2.24e-2
2048	409	64	2.4336	2.15e-2	5.2416	2.14e-2
1024	204	32	0.6864	1.18e-2	0.9672	2.87e-2

From table I we can see that the running time required by AFSG is less than the one by NBBL1, which shows that AFSG is the faster. Meanwhile, the relative errors obtained by AFSG is also smaller than the ones by NBBL1. In sum, the performance comparisons verified that the proposed algorithm AFSG is superior to NBBL1 in recovering a large and sparse signal.

5. Conclusions

In this paper, we proposed a new method to solve the l_0-norm regularized non-smooth non-convex minimization problems. Different from some existing algorithms, the proposed method makes the discrete optimization problem continuous and smooth by using an approximated function. The theoretical guarantee of the replacement ℓ_0-norm by a smooth function to yield sparse solutions is not analyzed in this paper. However, the numerical experiments illustrate that the approach is promising. Finally, we hope that this novel method even its modifications can further be used in some other fields, including statistics, machine learning, and image processing.

Acknowledgements

REFERENCES

1. J. Barzilai, J.M. Borwein, *Two point step size gradient method*, IMA J. Numer. Anal. 8 (1988) , 141-148.
2. E. Candès, J. Romberg, and T. Tao, *Stable signal recovery from imcomplete and inaccurate information*, Commun. Pure Appl. Math., 59 (2005), 1207-1233.
3. E. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequence information*, IEEE Trans. Inform. Theory, 52 (2006), 489-509.
4. X. Chen, D. Ge, Z. Wang, and Y. Ye, *Complexity of unconstrained $L_2 - L_p$ minimization*, Math. Program., DOI 10.1007/s10107-012-0613-0.
5. X. Chen, F. Xu, and Y. Ye, *Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p Minimization*, SIAM J. Sci. Comput., 32 (2010), 2832-2852.
6. D.L. Donoho, *For most large underdetermined systems of linear equations, the minimal l1-norm solution is also the sparsest solution*, Comm. Pure Appl. Math., 59 (2006), 907-934.
7. D.L. DONOHO, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), 1289-1306.
8. Y.H. Dai, W.W. Hager, K. Schittkowski, H. Zhang, *The cyclic Barzilai-Borwein method for unconstrained optimization*, IMA J. Numer. Anal., 26 (2006), 604-627.
9. D. Ge, X. Jiang, and Y. Ye, *A note on the complexity of Lp minimization*, Math. Program., 129(2011), 285-299.
10. R. Gribnoval and M. Nielsen, *Sparse decompositions in unions of bases*, IEEE Trans. Info. Theory, 49(2003), 3320-3325.
11. E.T. Hale, W. Yin and Y. Zhang, *Fixed-point continuation for ℓ_1-minimization: Methodology and convergence*, SAIM Journal on Optimization., 19 (2008), 1107-1130.
12. M. Raydan, *On the Barzilai and Borwein gradient method for the large scale unconstrained minimization problem*, IMA J. Numer. Anal., 13 (1993), 321-326.
13. M. Raydan, *The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem*, SIAM J. Optim., 7 (1997), 26-33.
14. S.J. Wright, R.D. Nowak and M.A.T. Figueuredo, *Sparse reconstruction by separable approximation*, in proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 2008, 3373-3376.
15. A.S. Weigend, D.E. Rumelhart, and B. A. Huberman (1991), *Generalization by weight-elimination with application to forecasting*, in Advances in Neural Information Processing Systems 3 (Denver 1990), R. P. Lippmann, J. E. Moody, and D. S. Touretzky, Editors, 875-882. Morgan Kaufmann, San Mateo, CA.
16. Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang, *A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation*, SIAM Journal on Scientific Computing, 32 (2010), 1832-1857.
17. Z. Wen, W. Yin, H. Zhang and D. Goldfarb, *On the convergence of an active-set method for ℓ_1 minimization*, Optimization Method and Software, 27 (2012), 1127-1146.
18. Y. Xiao, S.-Y. Wu and L. Qi, *Nonmonotone Barzilai-Borwein Gradient Algorithm for ℓ_1-Regularized Nonsmooth Minimization in Compressive Sensing*, avaiable at http://arxiv.org/abs/1207.4538.