Iterative Soft/Hard Thresholding Homotopy Algorithm for Sparse Recovery

Yuling Jiao, Bangti Jin, Xiliang Lu

***Abstract*—In this note, we analyze an iterative soft / hard thresholding algorithm with homotopy continuation for recovering a sparse signal $x^\dagger$ from noisy data of a noise level ϵ. Under standard regularity and sparsity conditions, we design a path along which the algorithm will find a solution x^* which admits sharp reconstruction error $\|x^* - x^\dagger\|_{\ell^\infty} = O(\epsilon)$ with an iteration complexity $O(\frac{\ln \epsilon}{\ln \rho} np)$, where n and p are problem dimensionality and $\rho \in (0,1)$ controls the length of the path. Numerical examples are given to illustrate its performance.**

***Index Terms*—iterative soft/hard thresholding, homotopy, continuation, solution path, convergence**

I. Introduction

SPARSE recovery has attracted considerable attention in signal processing, machine learning, statistics and inverse problems over the last decade. Often the problem is formulated as

$$y = \Psi x^\dagger + \eta, \tag{1}$$

where $x^\dagger \in \mathbb{R}^p$ denotes the sparse signal to be recovered, $y \in \mathbb{R}^n$ is the noisy data with noise $\eta \in \mathbb{R}^n$ of level $\epsilon = \|\eta\|$, and the sensing matrix $\Psi \in \mathbb{R}^{n\times p}$ with $p \gg n$ has normalized columns $\{\psi_i\}$, i.e., $\|\psi_i\| = 1$, $i = 1, \ldots, p$. The sparsity can be enforced by either the ℓ^0 or ℓ^1 penalty, i.e.,

$$\min_{x\in\mathbb{R}^p} \tfrac{1}{2}\|\Psi x - y\|^2 + \lambda \|x\|_t, \quad t \in \{0, 1\}, \tag{2}$$

where $\lambda > 0$ is the regularization parameter, controlling the sparsity of the penalized solution.

Among existing algorithms for minimizing (2), iterative soft / hard thresholding (IST/IHT) algorithm [1]–[4] and their accelerated extension [5], [6] are extremely popular. These algorithms are of the form

$$x^{k+1} = T_{\tau_k \lambda}(x^k + \tau_k \Psi^t (y - \Psi x^k)), \tag{3}$$

where T_λ is a soft- or hard-thresholding operator, and τ_k is the stepsize. The convergence of such algorithms was analyzed in many works, predominantly under the condition that the stepsize τ_k is smaller than $2/\|\Psi\|^2$. Under this condition, the thresholding step is (asymptotically) contractive, leading directly to the desired convergence [1]–[4]. Meanwhile, a lot of experimental studies have shown that the continuation strategy along the parameter λ can substantially improve the performance of the algorithm in terms of computing time [6]–[10]. However, as pointed out by the review paper [11] "... the design of a robust, practical, and theoretically effective continuation algorithm remains an interesting open question ..." There have been several works aiming at filling in this gap [12]–[15]. In the works [12], [13], the proximal gradient method with continuation for ℓ^1 minimization was analyzed with linear search, under sparse restricted eigenvalue condition or restricted strong convexity condition. Recently, a Newton type method with continuation was also studied for both ℓ^1 and ℓ^0 minimization [14], [15]. In this work, we present a unified approach to study the IST/IHT algorithm with homotoy continuation and a fixed stepsize $\tau = 1$. The main challenge is to overcome the lack of monotonicity of function values caused by a fixed stepsize $\tau = 1$.

The overall procedure with continuation is given in Algorithm 1. Here λ_0 is an initial guess of λ, supposedly large, $\rho \in (0,1)$ is the decreasing factor for λ, and K_{max} is the maximum number of inner iterations (for a fixed λ). The choice of the final λ^* is given in (4) below. One distinct feature of the algorithm is that the inner iteration does not need to be solved exactly, and one inner iteration suffices the desired accuracy of the final solution x^*, cf. Theorem 2 below.

Algorithm 1 Iterative Soft/Hard-Thresholding with Continuation (ISTC/IHTC)

Input: $\Psi \in \mathbb{R}^{n\times p}$, y, λ_0, $\rho \in (0,1)$, λ^*, $K_{max} \in \mathbb{N}$, $x(\lambda_0) = 0$.
for $s = 1, 2, \ldots$ **do**
  Let $\lambda_s = \rho\lambda_{s-1}$, $x^0 = x(\lambda_{s-1})$.
  If $\lambda_s < \lambda^*$, stop and output $x^* = x^0$.
  for $k = 0, 1, \ldots, K_{max} - 1$ **do**
    $x^{k+1} = T_\lambda(x^k + \Psi^t(y - \Psi x^k))$.
  end for
  Set $x(\lambda_s) = x^{K_{max}}$
end for

In Theorem 2, we shall prove that under standard conditions on the matrix Ψ, ISTC/IHTC always converges without the line search step, i.e., stepsize τ is fixed at 1.

II. Convergence analysis

The starting point of our analysis is the next lemma.

Lemma 1: For any $x, y \in \mathbb{R}$, there holds

$$|T_\lambda(x + y) - x| \le \begin{cases} |y| + \lambda & \text{IST}, \\ |y| + \sqrt{2\lambda} & \text{IHT}. \end{cases}$$

School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, 430063, P.R. China. (yulingjiaomath@whu.edu.cn)

Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK. (bangti.jin@gmail.com, b.jin@ucl.ac.uk)

Corresponding author. School of Mathematics and Statistics and Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan 430072, P.R. China. (xllv.math@whu.edu.cn)

Proof: By the definition of the thresholding operator T_λ

$$|T_\lambda(x+y)-x| \le |T_\lambda(x+y)-(x+y)|+|y| \le \begin{cases} |y|+\lambda & \text{IST}, \\ |y|+\sqrt{2\lambda} & \text{IHT}, \end{cases}$$

which completes the proof of the lemma. ∎

Let the true signal $x^\dagger$ be T-sparse with a support $\mathcal{A}^\dagger$. The mutual coherence (MC) μ of the matrix Ψ is defined by $\mu = \max_{i\neq j}|\langle\psi_i,\psi_j\rangle|$ [16].

Assumption 2.1: The MC μ of Ψ satisfies $2\mu T < 1$.

The proper choice of the parameter λ is essential for successful reconstruction. It is well known that the choice $\lambda = O(\epsilon)$ for the ℓ_1 penalty and $\lambda = O(\epsilon^2)$ for the ℓ_0 penalty ensures $\|x-x^\dagger\|_{\ell^\infty} = O(\epsilon)$, under Assumption 2.1 [15], [17]. Below we consider the following choice

$$\lambda^* = \begin{cases} C_1\epsilon, \text{ with } C_1 > \frac{1}{1-2\mu T}, & \text{for ISTC}, \\ C_0\epsilon^2, \text{ with } C_0 > \frac{1}{2(1-2\mu T)^2}, & \text{for IHTC}. \end{cases} \tag{4}$$

Now we can state the global convergence of Algorithm 1. In particular, it does not require a line search step.

Theorem 2: Let Assumption 2.1 hold, and λ^* be chosen by (4). Suppose that λ_0 is large, $K_{max}\in\mathbb{N}$, and

$$\rho \in \begin{cases} [2\mu T C_1/(C_1-1),1), & \text{for ISTC}, \\ [(\frac{2\mu T}{1-1/(2C_0)^{1/2}})^2,1), & \text{for IHTC}. \end{cases}$$

Then Algorithm 1 is well-defined, and the solution x^* satisfies:

(i) $\operatorname{supp}(x^*)\subset\mathcal{A}^\dagger$,

(ii) there holds the error estimate

$$\|x^*-x^\dagger\|_{\ell^\infty} \le \begin{cases} (C_1-1)\epsilon/(\mu T), & \text{for ISTC}, \\ (\sqrt{2C_0}-1)\epsilon/(\mu T), & \text{for IHTC}. \end{cases}$$

Further, if $\min_{i\in\mathcal{A}^\dagger}|x_i^\dagger|$ is large enough, then $\operatorname{supp}(x^*)=\mathcal{A}^\dagger$.

Proof: We only prove the theorem for ISTC, and that for IHTC is analogously. The choice C_1 in (4) implies $C_1>1$ and $\frac{2\mu T}{1-1/C_1}<1$, and thus the choice of ρ makes sense. Let $E^k = \|x^k-x^\dagger\|_{\ell^\infty}$, and $\alpha = \frac{1-1/C_1}{\mu T}$. Consider one IST iteration from x^k to x^{k+1}.

The key step to the convergence is the following implication: with $\mathcal{A}^k=\operatorname{supp}(x^k)$

$$\begin{aligned} &\mathcal{A}^k\subset\mathcal{A}^\dagger \text{ and } E^k\le\alpha\lambda \\ \Rightarrow &\mathcal{A}^{k+1}\subset\mathcal{A}^\dagger \text{ and } E^{k+1}\le\alpha\rho\lambda \quad \forall\lambda\ge\lambda^*. \end{aligned} \tag{5}$$

Now we show this claim. It follows from (1) and $\|\Psi_i\|=1$ the following componentwise expression for the update

$$\begin{aligned} x_i^{k+1} &= T_\lambda(x_i^k+\Psi_i^t(y-\Psi x^k)) \\ &= T_\lambda(x_i^\dagger+\Psi_i^t(\Psi_{\mathcal{A}^\dagger\cup\mathcal{A}^k\setminus\{i\}}(x^\dagger-x^k)_{\mathcal{A}^\dagger\cup\mathcal{A}^k\setminus\{i\}}+\eta)). \end{aligned}$$

From the hypothesis in (5), $\mathcal{A}^k\subset\mathcal{A}^\dagger$, $E^k\le\alpha\lambda$, $\lambda\ge\lambda^*$ and (4), we deduce that for any $i\in\mathcal{I}^\dagger$

$$\begin{aligned} &|x_i^\dagger+\Psi_i^t(\Psi_{\mathcal{A}^\dagger\cup\mathcal{A}^k\setminus\{i\}}(x^\dagger-x^k)_{\mathcal{A}^\dagger\cup\mathcal{A}^k\setminus\{i\}}+\eta)| \\ \le&|\Psi_i^t(\Psi_{\mathcal{A}^\dagger}(x^\dagger-x^k)_{\mathcal{A}^\dagger}|+|\Psi_i^t\eta| \\ \le&\mu T E^k+\epsilon\le(\tfrac{1}{C_1}+\mu T\alpha)\lambda=\lambda, \end{aligned}$$

by the definition of α, and the second inequality follows from [15, Lemma 2.1]. Hence, $|x_i^{k+1}|\le|T_\lambda(\mu T E^k+\epsilon)|=0$, which implies directly $\mathcal{A}^{k+1}\subset\mathcal{A}^\dagger$. Meanwhile, under (5) and (4), for any $i\in\mathcal{A}^\dagger$, by Lemma 1, we deduce

$$\begin{aligned} |x_i^{k+1}-x_i^\dagger| &\le \lambda+|\Psi_i^t(\Psi_{\mathcal{A}^\dagger\setminus\{i\}}(x^\dagger-x^k)_{\mathcal{A}^\dagger\setminus\{i\}}|+|\Psi_i^t\eta| \\ &\le \lambda+\mu(T-1)E^k+\epsilon\le\lambda+\mu T\alpha\lambda+\tfrac{1}{C}\lambda \\ &= (1+\tfrac{1}{C}+\alpha\mu T)\lambda = 2\lambda\le\alpha\rho\lambda. \end{aligned}$$

Then the assertion $E^{k+1}\le\alpha\rho\lambda$ follows, showing the claim.

Next we proceed by mathematics induction. Since λ_0 is large, it satisfies the condition in (5). Then for all s with $\lambda_s\ge\lambda^*$, there holds $\operatorname{supp}x(\lambda_s)\subset\mathcal{A}^\dagger$ and $\|x(\lambda_s)-x^\dagger\|_{\ell^\infty}\le\alpha\rho\lambda_s$. When Algorithm 1 terminates, the first two properties hold, i.e., $\operatorname{supp}x^*\subset\mathcal{A}^\dagger$ and $\|x^*-x^\dagger\|_{\ell^\infty}\le\alpha\lambda^* = (C_1-1)\epsilon/(\mu T)$. Likewise, if $\min_{i\in\mathcal{A}^\dagger}|x_i|>(C-1)\epsilon/(\mu T)$, property (ii) implies $\operatorname{supp}(x^*)=\mathcal{A}^\dagger$.

Last, we briefly discuss IHTC. For the choice C_0 in (4), $\rho\in[(\frac{2\mu T}{1-1/(2C_0)^{1/2}})^2,1)$ makes sense. With $\alpha=\frac{1-1/(2C_0)^{1/2}}{\mu T}$, a similar argument yields

$$\begin{aligned} &\mathcal{A}^k\subset\mathcal{A}^\dagger \text{ and } E^k\le\alpha\sqrt{2\lambda} \\ \Rightarrow&\mathcal{A}^{k+1}\subset\mathcal{A}^\dagger \text{ and } E^{k+1}\le\alpha\sqrt{2\rho\lambda}. \end{aligned}$$

The rest follows like before, and thus it is omitted. ∎

With the help of Theorem 2, we can estimate the complexity of Algorithm 1. At each iteration, one needs to compute matrix-vector product Ψx and $\Psi^t y$, and for each λ, the number of iterations is bounded by K_{max}. The overall computational cost depends on the decreasing factor ρ by $O(\frac{\ln\lambda^*}{\ln\rho}np) = O(\frac{\ln\epsilon}{\ln\rho}np)$.

III. Numerical Results and Discussions

Now we present numerical examples to show the performance of Algorithm 1. First, we give implementation details, e.g., data generation, parameter setting for the algorithm. Then our method is compared with several state-of-the-art algorithms in terms of CPU time and reconstruction error.

A. Implementation details

Following [6], the signals $x^\dagger$ are chosen as T-sparse with a dynamic range

$$DR := \frac{\max\{|x_i^\dagger|: x_i^\dagger\neq 0\}}{\min\{|x_i^\dagger|: x_i^\dagger\neq 0\}}.$$

The matrix $\Psi\in\mathbb{R}^{n\times p}$ is chosen to be either random Gaussian matrix, or random Bernoulli matrix, or the product of a partial FFT matrix and inverse Haar wavelet transform. The noise η has entries following i.i.d. $N(0,\sigma^2)$.

We fix the algorithm parameters as follows: $\lambda_0=\|\Psi^t y\|_\infty$ and $\lambda_0=\|\Psi^t y\|_\infty^2/2$ for ISTC and IHTC, respectively, [14], [15]; decreasing factor $\rho=0.8$; maximum iteration number $K_{max}=20$. Since the optimal λ^* depends on the noise level ϵ, which is often unknown in practice, we predefine a path $\Lambda=\{\lambda_s\}_{s=0}^N$ with $\lambda_t=\lambda_0\rho^t$ and $N=100$. Then we run Algorithm 1 on the path Λ and select the optimal λ^* by Bayesian information criterion [14]. All the computations were performed on a dual core laptop with 1.90 GHz and 8 GB RAM using `MATLAB` 2015b. The `MATLAB` package `ISHTC` for reproducing all the numerical results can be found at http://www0.cs.ucl.ac.uk/staff/b.jin/companioncode.html.

TABLE I: Numerical results (CPU time and errors), with random Gaussian Ψ, of size $p = 1024$, 2048, 4096, 8192, 16384, 24576, $n = \lfloor p/4 \rfloor$, $T = \lfloor n/32 \rfloor$, with $DR = 100$ and $\sigma =$ 1e-2.

p	method	time(s)	Relative ℓ^2 error	Absolute ℓ^∞ error
	ISTC	0.03	7.42e-4	4.53e-2
	HPG	0.06	7.06e-4	4.41e-2
1024	SpaRSA	0.07	7.19e-4	4.37e-2
	GPSR	0.07	7.50e-4	4.65e-2
	ISTC	0.10	8.03e-4	4.87e-2
	HPG	0.19	7.79e-4	4.84e-2
2048	SpaRSA	0.27	7.80e-4	4.82e-2
	GPSR	0.23	8.19e-4	5.04e-2
	ISTC	0.37	8.68e-4	4.74e-2
	HPG	0.63	8.36e-4	4.65e-2
4096	SpaRSA	0.98	8.36e-4	4.65e-2
	GPSR	0.83	8.82e-4	4.88e-2
	ISTC	1.45	8.17e-4	5.16e-2
	HPG	2.38	7.93e-4	5.09e-2
8192	SpaRSA	3.94	7.93e-4	5.09e-2
	GPSR	3.29	8.34e-4	5.33e-2
	ISTC	5.75	9.10e-4	5.50e-2
	HPG	9.90	8.82e-4	5.40e-2
16384	SpaRSA	15.71	8.82e-4	5.39e-2
	GPSR	13.16	9.32e-4	5.64e-2
	ISTC	13.83	8.84e-4	6.21e-2
	HPG	22.42	8.57e-4	6.00e-2
24576	SpaRSA	35.76	8.57e-4	6.00e-2
	GPSR	29.66	9.05e-4	6.36e-2

TABLE II: Numerical results (CPU time and errors), with random Gaussian Ψ, of size $p =$ 10000, 12000, 14000, 16000, 18000, 20000, $n = \lfloor p/4 \rfloor$, $T = \lfloor n/40 \rfloor$, with $DR = 100$ and $\sigma =$ 5e-2.

p	method	time(s)	Relative ℓ^2 error	Absolute ℓ^∞ error
	ISTC	1.79	4.21e-3	2.66e-1
	HPG	3.06	4.14e-3	2.66e-1
10000	SpaRSA	6.02	4.13e-3	2.63e-1
	GPSR	5.10	4.25e-3	2.71e-1
	ISTC	2.54	4.54e-3	2.72e-1
	HPG	4.38	4.45e-3	2.68e-1
12000	SpaRSA	8.57	4.44e-3	2.67e-1
	GPSR	7.33	4.57e-3	2.75e-1
	ISTC	3.44	4.30e-3	2.71e-1
	HPG	5.81	4.21e-3	2.68e-1
14000	SpaRSA	11.7	4.21e-3	2.67e-1
	GPSR	9.92	4.32e-3	2.75e-1
	ISTC	4.47	4.09e-3	2.68e-1
	HPG	7.49	4.02e-3	2.66e-1
16000	SpaRSA	15.1	4.01e-3	2.65e-1
	GPSR	12.9	4.14e-3	2.72e-1
	ISTC	5.65	4.34e-3	2.88e-1
	HPG	10.1	4.25e-3	2.85e-1
18000	SpaRSA	19.4	4.25e-3	2.84e-1
	GPSR	16.4	4.36e-3	2.91e-1
	ISTC	7.01	4.53e-3	2.86e-1
	HPG	12.1	4.43e-3	2.81e-1
20000	SpaRSA	23.9	4.43e-3	2.82e-1
	GPSR	20.2	4.55e-3	2.88e-1

B. Comparison of ISTC with ℓ^1 solvers

We first compare ISTC with three state-of-the-art ℓ^1 solvers: GPSR (with Barzilai-Borwein rule) [8] (available at at http://www.lx.it.pt/mtf/GPSR/), SpaRSA [9] (available at http://www.lx.it.pt/mtf/SpaRSA/), proximal-gradient homotopy method (PGH) [12] (available at https://www.microsoft.com/en-us/download/details.aspx?id=52421). All these three methods employ continuation.

The numerical results (CPU times, relative ℓ_2 errors, and absolute ℓ_∞ errors) are computed from 10 independent realizations of the problem setup for random Gaussian or random Bernoulli sensing matrices with different parameter tuples (n, p, T, DR, σ) are shown in Tables I-II, respectively. It is observed that ISTC yields reconstructions that are comparable with that by other methods, e.g., GPSR, SpaRSA and PGH but usually two-three times faster. Further, we observe that it scales well with the problem size.

C. Comparison of IHTC with greedy solvers

Now we compare IHTC with three greedy methods for the ℓ^0 problem for recovering a 1D signal and a benchmark MRI image. These include OMP [18] (available at https://sparselab.stanford.edu/SparseLab_files/Download_files/SparseLab21-Core.zip), accelerated IHT (AIHT) [19] (available at http://www.personal.soton.ac.uk/tb1m08/sparsify/sparsify_0_5.zip), CoSaMP [20] (available at http://mdav.ece.gatech.edu/software/SSCoSaMP-1.0.zip).

The underlying 1D signal and 2D Phantom image are compressible under a wavelet basis. Therefor, the observation data can be chosen as the wavelet coefficients sampled by the product of a partial FFT matrix and inverse Haar wavelet transform. For the 1D signal, the matrix Ψ is of size 665×1024, and it consists of applying a partial FFT and an inverse two level Harr wavelet transform, and the signal under wavelet transformation has 247 nonzero entries and $\sigma =$ 1e-4. The results are shown in Fig. 1 and Table III. Fig. 1 shows that the reconstruction by IHTC is visually more appealing than that of the others. The reconstructions by AIHT and CoSaMP suffer from pronounced oscillations. This is further confirmed by the PSNR value defined by

$$\mathrm{PSNR} = 10 \cdot \log \frac{V^2}{\mathrm{MSE}}$$

where V is the maximum absolute value of the true signal, and MSE is the mean squared error of the reconstruction. Table III also presents the CPU time of the 1D example, showing that IHTC is the fastest one.

For the 2D MRI image, the matrix Ψ amounts to a partial FFT and an inverse wavelet transform, and it has a size 34489×262144. The image under eight level Haar wavelet transformation has 7926 nonzero entries and $\sigma =$ 3e-2. The numerical results are shown in Fig. 2 and Table IV. All methods produce comparable results, but IHTC is fastest.

TABLE III: 1D signal with $n = 665$, $p = 1024$, $T = 247$, σ=1e-4.

method	CPU time	PSNR
IHTC	0.38	51
OMP	0.39	49
AIHT	1.01	34
CoSaMP	0.41	26

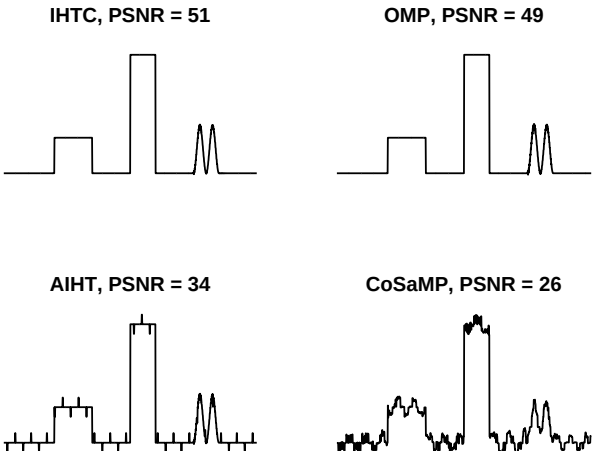

Fig. 1: Reconstructed signals and their PSNR values

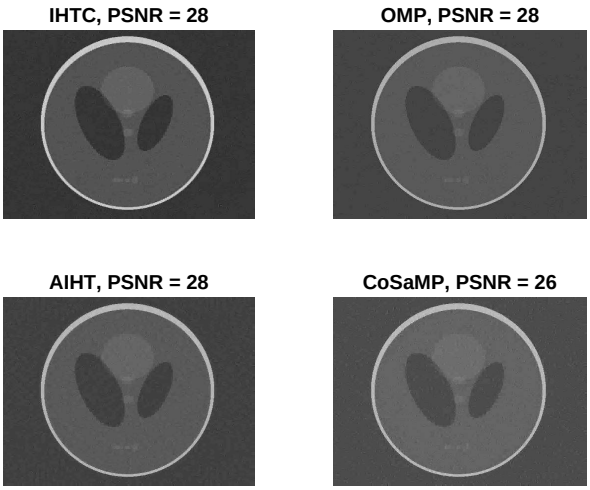

Fig. 2: Reconstructed MRI images and their PSNR values

TABLE IV: 2D image with $n = 34489$, $p = 262144$, $T = 7926$, σ=3e-2.

method	CPU time	PSNR
IHTC	10.9	28
OMP	1580	28
AIHT	23.4	28
CoSaMP	27.0	26

IV. Conclusion

In this paper, we analyze an iterative soft / hard thresholding algorithm with homotopy continuation for sparse recovery from noisy and underdetermined data. Under standard regularity condition and sparsity assumptions, sharp reconstruction errors can be obtained with an iteration complexity $O(\frac{\ln \epsilon}{\ln \rho} np)$. Numerical results indicated its competitiveness with state-of-the-art sparse solvers in terms of reconstruction error and computing time. Last, we note that the results can be extended straightforwardly to other thresholding operators, e.g., MCP [21] or SCAD [22].

Acknowledgements

The research of Y. Jiao is partially supported by National Science Foundation of China No. 11501579, B. Jin by EPSRC grant EP/M025160/1, and X. Lu by National Science Foundation of China No. 11471253.

References

[1] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[2] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.

[3] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *J. Fourier Anal. Appl.*, vol. 14, no. 5-6, pp. 629–654, 2008.

[4] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods," *Math. Program.*, vol. 137, no. 1-2, pp. 91–129, 2013.

[5] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[6] S. Becker, J. Bobin, and E. Candés, "NESTA: a fast and accurate first-order method for sparse recovery," *SIAM J. Imag. Sci.*, vol. 4, no. 1, pp. 1–39, 2011.

[7] E. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for ℓ_1-minimization: Methodology and convergence," *SIAM J. Optim.*, vol. 19, no. 3, pp. 1107–1130, 2008.

[8] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Proc.*, vol. 1, no. 4, pp. 586–597, 2007.

[9] S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Proc.*, vol. 57, no. 7, pp. 2479–2493, 2009.

[10] D. A. Lorenz, "Constructing test instances for basis pursuit denoising," *IEEE Trans. Signal Proc.*, vol. 5, no. 61, pp. 1210–1214, 2013.

[11] J. Tropp and S. Wright, "Computational methods for sparse solution of linear inverse problems," *Proc. IEEE*, vol. 98, no. 6, pp. 948–958, 2010.

[12] L. Xiao and T. Zhang, "A proximal-gradient homotopy method for the sparse least-squares problem," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1062–1091, 2013.

[13] A. Agawal, S. Negahban, and M. J. Wainwright, "Fast global convergence of gradient methods for high-dimensional statistical recovery," *Ann. Stat.*, vol. 40, no. 5, pp. 2452–2482, 2012.

[14] Q. Fan, Y. Jiao, and X. Lu, "A primal dual active set algorithm with continuation for compressed sensing," *IEEE Trans. Signal Proc.*, vol. 62, no. 23, pp. 6276–6285, 2014.

[15] Y. Jiao, B. Jin, and X. Lu, "A primal dual active set with continuation algorithm for the ℓ^0-regularized optimization problem," *Appl. Comput. Harmon. Anal.*, vol. 39, no. 3, pp. 400–426, 2015.

[16] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[17] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[18] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[19] T. Blumensath, "Accelerated iterative hard thresholding," *Signal Proc.*, vol. 92, no. 3, pp. 752–756, 2012.

[20] D. Needell and J. A. Tropp, "CoSaMP: iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2009.

[21] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Stat.*, vol. 38, no. 2, pp. 894–942, 2010.

[22] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.