Weighted Thresholding Homotopy Method for Sparsity Constrained Optimization

Wenxing Zhu, Huating Huang, Lanfan Jiang, Jianli Chen
Center for Discrete Mathematics and Theoretical Computer Science
Fuzhou University, Fuzhou 350116, China

Abstract

Weighted or reweighted strategies have not been considered for sparsity constrained optimization. In this paper, we reformulate the sparsity constraint as an equivalent weighted l_1-norm constraint in the sparsity constrained optimization problem. To solve the reformulated problem, we investigate the problem in the Lagrange dual framework, and prove that the strong duality property holds. Then we propose a weighted thresholding method to solve the Lagrangian problem. Moreover, we analyze convergence of the method and derive an error bound of the solution under some assumptions. Furthermore, we propose a weighted thresholding homotopy method to get a suitable Lagrange multiplier and prove that this homotopy method can find an L-stationarity point of the primal problem. Computational experiments on compressed sensing and sparse logistic regression problems show that, in comparable running time the proposed weighted thresholding homotopy method works better than state-of-the-art methods on randomly generated instances of the compressed sensing problem and on the synthetic data of the logistic regression problem. Especially, the sparsity level improvement of exact recovery is 26.9% for compressed sensing, and the sparsity level improvement of exact classification is 38.5% for sparse logistic regression on randomly generated data.

Keywords: sparsity constraint, weighted thresholding, Lagrangian method, homotopy technique.
Subject Classifications: Integer: heuristic; Nonlinear: algorithms.
Area of Review: Programming.

1 Introduction

In this paper, we consider the following s-sparse constrained minimization problem

$$\begin{aligned} \min \ & f(x) \\ s.t. \ & \|x\|_0 \leq s \\ & l \leq x \leq u, \end{aligned} \tag{1}$$

where $l \leq 0$, $u \geq 0$, $x \in \mathbb{R}^n$. The notation $\|x\|_0$ denotes the number of nonzero components of x (called the l_0-norm of x, for simplicity), $s \ll n$ is a nonnegative integer, and $f(x)$ is a function whose gradient is Lipschitz-continuous with Lipschitz constant $L_f > 0$. If $f(x) = \|Ax - b\|^2$, then problem (1) is the linear compressed sensing problem, which can be found in signal processing [10], image and vision processing [17]. Another case is that $f(x)$ is the logistic regression loss function [16], which has been widely used in machine learning.

Over the past decades, many methods have been developed for the sparse optimization problem, although it is strongly NP-hard to find an approximate optimal solution within certain error bound [8]. These methods can be categorized roughly into two classes. One is solving the problem directly, which includes greedy methods and gradient projection methods. The other one is solving the l_0-regularized problem of problem (1) directly or indirectly.

In the first class of methods, the matching pursuit (MP) algorithm [18] is the earliest greedy method to seek a sparse solution of an underdetermined linear system, which is known as Compressed Sensing. At each iteration, this method explores support set by taking only one new component into the current support set, and solves a convex quadratic programming problem, which is time consuming. To improve reconstruction speed and precision of the method, many variants of the MP algorithm have been proposed, e.g., GraSP [1], CoSaMP [19] and OMP [23]. Among them, the Gradient Support Pursuit (GraSP) method by Bahmani et al. [1] can solve the problem with a more general objective function.

Gradient projection methods are based on the idea of the iterative hard thresholding (IHT) method [5], which takes a gradient descent step at the current solution and keeps the s largest components. To enhance convergence speed of the IHT method, Blumensath [4] designed an accelerated IHT method. Similar to the IHT method, Qiu and Dogandžić [21] proposed an Expectation-Conditional Maximisation Either (ECME) algorithm for sparse signal reconstruction. For the problem of minimizing a general function over sparse symmetric set, Beck et al. [2] introduced three types of optimality conditions: basic feasibility, L-stationary, and coordinatewise optimality, and used the IHT method to find an L-stationary point.

For the second class of methods, the l_0-regularized problem contains an l_0-norm function in the regularization term, which makes the problem hard to solve. To handle this issue, one approach is the IHT method [15], which keeps the l_0-norm and approximates $f(x)$ by a separable quadratic function, and then solves the resulting problem. Since the l_0-norm is the limit of l_p-norm as $p \to 0$, another approach is relaxing the l_0-norm as the l_p-norm ($0 < p \leq 1$), and solving the relaxation problem. Specifically, Xu et al. [26] developed an iterative half thresholding operator for the $l_{1/2}$-norm regularized problem. Xiao and Zhang [25] adopted a homotopy technique and proposed a proximal gradient method to solve the l_1-norm regularized least-squares problem. Another similar method is the Lasso homotopy algorithm [9].

It is worth mentioning that, the iterative reweighted approach has been successively applied to the l_0-norm minimization problem

$$\begin{aligned} \min \ & \|x\|_0 \\ s.t. \ & Ax = b \\ & l \leq x \leq u, \end{aligned}$$

which modifies the objective function as $\sum_{i=1}^{n} w_i^k |x_i|$, where w_i^k is a weight. A simple reweighted strategy [7] is setting the weight w_i^k as $w_i^k = \frac{1}{|x_i^{k-1}|+\delta}$, where $\delta > 0$. More reweighted methods can refer to [13, 28, 29], which show that reweighted l_1-norm methods outperform unreweighted ones empirically and theoretically. This motivates us using the weighted idea on problem (1).

However, no literature concerns the weighted l_1-norm strategy for the s-sparse constrained minimization problem (1). In this paper, we reformulate the problem equivalently as a weighted l_1-norm constrained optimization problem, and prove that the problem has the strong duality property. Then we propose a weighted thresholding method to solve the Lagrangian problem, in which the weight w and variable x are optimized simultaneously. We analyze the convergence of the method, and under some assumptions show a bound between an optimal solution of problem (1) and a solution generated by the method. Furthermore, to get a suitable Lagrange multiplier, we propose a weighted thresholding based homotopy method and prove that the homotopy method can find an L-stationarity point of the primal problem. Computational experiments indicate that, our method works better than state-of-the-art methods on randomly generated instances of compressed sensing and sparse logistic regression problems, and is comparable on real data sets of the logistic regression problem. Especially, the sparsity level improvement of exact recovery is 26.9% for compressed sensing, and the sparsity level improvement of exact classification is 38.5% for sparse logistic regression on randomly generated data.

The rest of this paper is organized as follows. In Section 2, we first give some notations. Then we propose a weighted thresholding method for the s-

parsity constrained optimization problem, where the weight w and variable x are optimized simultaneously. In Section 3, we investigate the Lagrangian of the problem and further develop a weighted thresholding homotopy method. We prove that any accumulation point generated by the method is an L-stationary point of problem (1). In Section 4, we compare experimental results of our method on compressed sensing and sparse logistic regression with state-of-the-art methods. Finally, conclusions are drawn in Section 5.

2 Equivalent formulation

2.1 Notations

Unless otherwise stated, $\|\cdot\|$ denotes the Euclidean norm. The transpose of a vector $x \in \mathbb{R}^n$ is denoted by x^T. Let $B = \{x \in \mathbb{R}^n : l \leq x \leq u\}$, and let $C_s = \{x \in \mathbb{R}^n : \|x\|_0 \leq s\}$, which is a set including all the s-sparse vectors. Let $\Pi_C(y)$ be the projection of $y \in \mathbb{R}^n$ onto the set C, i.e.,

$$\Pi_C(y) = \operatorname{argmin}\{\|x - y\|^2 : x \in C\}.$$

Given an index set $I \subseteq \{1, \ldots, n\}$, x_I denotes the subvector formed by the components of x indexed by I. The index set of nonzero components of a vector x is denoted by $I(x) = \{i : x_i \neq 0\}$ (called support set). Let $\bar{I}(x)$ be the complement set of $I(x)$, i.e., $\bar{I}(x) = \{i : x_i = 0\}$. We also define $A = \{i : w_i = 0\}$, where w_i is the weight corresponding to x_i, and define the complement of A as $\bar{A} = \{i : w_i = 1\}$. $w \circ x$ denotes the dot product of two vectors, i.e., $w \circ x = (w_1x_1, w_2x_2, \cdots, w_nx_n)^T$. For a set S, $|S|$ is the number of elements in S.

2.2 Equivalent formulation

First, we give an equivalent formulation of the sparsity constraint $x \in C_s$.

Lemma 1. *$x \in C_s$ is equivalent to that there exists $w \in \{0,1\}^n$ such that*

$$\begin{cases} \|w \circ x\|_1 = 0, \\ \|1 - w\|_0 \leq s. \end{cases} \tag{2}$$

Proof. If $x \in C_s$, then we assign the value of w as

$$\begin{cases} w_i = 0, & \text{if } \ i \in \{i : x_i \neq 0\}; \\ w_i = 1, & \text{if } \ i \in \{i : x_i = 0\}, \end{cases}$$

which satisfies (2). Conversely, suppose $w \in \{0,1\}^n$ and x satisfy (2). If $x \notin C_s$, then the number of nonzero components of x is larger than s. However, $\|1 - w\|_0 \leq s$ and $w \in \{0,1\}^n$ indicate that the number of zero components of w is not more than s. Hence $\|w \circ x\|_1 \neq 0$, which leads to a contradiction. □

Then, by Lemma 1 we can rewrite problem (1) equivalently as

$$\begin{aligned} \min \ & f(x) \\ s.t. \ & \|w \circ x\|_1 = 0 \\ & \|1 - w\|_0 \leq s \\ & w \in \{0,1\}^n \\ & l \leq x \leq u. \end{aligned} \tag{3}$$

For simplification, we define $\Omega = \{(x, w) : \|1 - w\|_0 \leq s, w \in \{0,1\}^n, l \leq x \leq u\}$.

2.3 Properties of Lagrange dual problem

Motivated by the Lagrangian method, for problem (3) we consider the Lagrangian problem

$$\min_{(x,w)\in\Omega} F_\lambda(x, w) = f(x) + \lambda\|w \circ x\|_1, \tag{4}$$

where $\lambda \geq 0$. The Lagrange dual function $g : \mathbb{R} \to \mathbb{R}$ is

$$g(\lambda) = \min_{(x,w)\in\Omega} f(x) + \lambda\|w \circ x\|_1,$$

and the Lagrange dual problem of problem (3) is

$$\max_{\lambda\geq 0} g(\lambda). \tag{5}$$

Next, we show some properties of the Lagrange dual problem, which will indicate the strong duality property. Obviously, we have

Lemma 2. *The Lagrange dual function $g(\lambda)$ is concave.*

Let

$$(x^*_\lambda, w^*_\lambda) \in \operatorname*{argmin}_{(x,w)\in\Omega} f(x) + \lambda\|w \circ x\|_1. \tag{6}$$

Then we have the following results.

Lemma 3. *If $\lambda_2 > \lambda_1 \geq 0$, then $\|w^*_{\lambda_1} \circ x^*_{\lambda_1}\|_1 \geq \|w^*_{\lambda_2} \circ x^*_{\lambda_2}\|_1$, and $f(x^*_{\lambda_1}) \leq f(x^*_{\lambda_2})$.*

Proof. We prove the first inequality by contradiction. Suppose that there exists $\lambda_2 > \lambda_1$ satisfying $\|w^*_{\lambda_1} \circ x^*_{\lambda_1}\|_1 < \|w^*_{\lambda_2} \circ x^*_{\lambda_2}\|_1$. Then for the λ_1, λ_2, there exists $t > 0$ such that

$$\|w^*_{\lambda_2} \circ x^*_{\lambda_2}\|_1 = \|w^*_{\lambda_1} \circ x^*_{\lambda_1}\|_1 + t.$$

Furthermore by (6), we have

$$\begin{aligned} f(x^*_{\lambda_2}) + \lambda_2(\|w^*_{\lambda_1} \circ x^*_{\lambda_1}\|_1 + t) &= f(x^*_{\lambda_2}) + \lambda_2\|w^*_{\lambda_2} \circ x^*_{\lambda_2}\|_1 \\ &\leq f(x^*_{\lambda_1}) + \lambda_2\|w^*_{\lambda_1} \circ x^*_{\lambda_1}\|_1 \end{aligned}$$

and

$$\begin{aligned} f(x^*_{\lambda_1}) + \lambda_1\|w^*_{\lambda_1} \circ x^*_{\lambda_1}\|_1 &\leq f(x^*_{\lambda_2}) + \lambda_1\|w^*_{\lambda_2} \circ x^*_{\lambda_2}\|_1 \\ &= f(x^*_{\lambda_2}) + \lambda_1(\|w^*_{\lambda_1} \circ x^*_{\lambda_1}\|_1 + t). \end{aligned} \tag{7}$$

The above two inequalities imply that

$$f(x^*_{\lambda_2}) + \lambda_2 t \leq f(x^*_{\lambda_1}) \leq f(x^*_{\lambda_2}) + \lambda_1 t.$$

Hence $\lambda_2 \leq \lambda_1$, which contradicts the assumption that $\lambda_2 > \lambda_1 \geq 0$. Thus, $\|w^*_{\lambda_1} \circ x^*_{\lambda_1}\|_1 \geq \|w^*_{\lambda_2} \circ x^*_{\lambda_2}\|_1$.

Finally, by (7), we obtain that

$$f(x^*_{\lambda_1}) - f(x^*_{\lambda_2}) \leq \lambda_1(\|w^*_{\lambda_2} \circ x^*_{\lambda_2}\|_1 - \|w^*_{\lambda_1} \circ x^*_{\lambda_1}\|_1).$$

Combining $\lambda_1 \geq 0$ with $\|w^*_{\lambda_1} \circ x^*_{\lambda_1}\|_1 \geq \|w^*_{\lambda_2} \circ x^*_{\lambda_2}\|_1$, it holds that $f(x^*_{\lambda_1}) \leq f(x^*_{\lambda_2})$. □

Lemma 3 indicates that, to obtain a smaller objective function value, λ should be smaller. However, to meet the condition $\|w \circ x\|_1 = 0$, λ should not be too small. If λ is too small such that $\|w^*_\lambda \circ x^*_\lambda\|_1 \neq 0$, where $(x^*_\lambda, w^*_\lambda)$ is an optimal solution of problem (4), then by Lemma 2 $\|x^*_\lambda\|_0 > s$, which leads to x^*_λ being not an optimal solution of problem (1). Next, we prove that if $\|w^*_\lambda \circ x^*_\lambda\|_1 = 0$, where $(x^*_\lambda, w^*_\lambda)$ is an optimal solution of problem (4), then x^*_λ is an optimal solution of problem (1).

Corollary 1. *If there exists λ such that $\|w^*_\lambda \circ x^*_\lambda\|_1 = 0$. Then problem* (1) *has the strong duality property:*

$$\min_{\|x\|_0 \leq s, l \leq x \leq u} f(x) = \max_{\lambda \geq 0} \min_{(x,w) \in \Omega} f(x) + \lambda\|w \circ x\|_1$$

Proof. Since $(x^*_\lambda, w^*_\lambda) \in \Omega$ and $\|w^*_\lambda \circ x^*_\lambda\|_1 = 0$, by Lemma 1 x^*_λ is a feasible solution of problem (1). Hence it holds that

$$\max_{\lambda \geq 0} \min_{(x,w) \in \Omega} f(x) + \lambda\|w \circ x\|_1 \geq f(x^*_\lambda) \geq \min_{\|x\|_0 \leq s, l \leq x \leq u} f(x).$$

Then, according to the weak duality theorem

$$\min_{\|x\|_0 \leq s, l \leq x \leq u} f(x) \geq \max_{\lambda \geq 0} \min_{(x,w) \in \Omega} f(x) + \lambda\|w \circ x\|_1,$$

we know the strong duality property holds. □

From the above analysis, it is important that there exists λ such that $\|w_\lambda^* \circ x_\lambda^*\|_1 = 0$. Next, we prove that there exists $\lambda \neq +\infty$ such that $\|w_\lambda^* \circ x_\lambda^*\|_1 = 0$. Firstly, we give a property on the optimal solution of problem (4).

Lemma 4. *Suppose that $(x_\lambda^*, w_\lambda^*)$ is an optimal solution of problem (4). Then*

$$([x_\lambda^*]_i, [w_\lambda^*]_i) = \begin{cases} ([\Pi_B(y)]_i, 0), & \text{if } i \in A; \\ ([\Pi_B(\text{soft}(y))]_i, 1), & \text{otherwise.} \end{cases} \tag{8}$$

where $y = x_\lambda^* - \frac{1}{L}\nabla f(x_\lambda^*)$, and $\text{soft}(y) = \text{sign}(y) \circ \max\{|y| - \lambda/L, 0\}$ is a soft thresholding operator.

Proof. Since $(x_\lambda^*, w_\lambda^*)$ is an optimal solution of problem (4), we have

$$f(x) + \lambda\|w \circ x\|_1 \geq f(x_\lambda^*) + \lambda\|w_\lambda^* \circ x_\lambda^*\|_1$$

for any feasible solution (x, w) of problem (4). Furthermore, since $\nabla f(x)$ is Lipschitz continuous, it holds that

$$f(x_\lambda^*) + \langle \nabla f(x_\lambda^*), x - x_\lambda^* \rangle + \frac{L}{2}\|x - x_\lambda^*\|^2 \geq f(x).$$

The above two inequalities imply that for any feasible solution (x, w) of problem (4),

$$\begin{aligned} &\langle \nabla f(x_\lambda^*), x - x_\lambda^* \rangle + \frac{L}{2}\|x - x_\lambda^*\|^2 + \lambda\|w \circ x\|_1 \\ \geq &0 + \lambda\|w_\lambda^* \circ x_\lambda^*\|_1 \\ = &\langle \nabla f(x_\lambda^*), x_\lambda^* - x_\lambda^* \rangle + \frac{L}{2}\|x_\lambda^* - x_\lambda^*\|^2 + \lambda\|w_\lambda^* \circ x_\lambda^*\|_1. \end{aligned}$$

Thus,

$$(x_\lambda^*, w_\lambda^*) \in \underset{(x,w)\in\Omega}{\text{argmin}} \left\{ \langle \nabla f(x_\lambda^*), x - x_\lambda^* \rangle + \frac{L}{2}\|x - x_\lambda^*\|^2 + \lambda\|w \circ x\|_1 \right\}.$$

Note that

$$\begin{aligned} &\|x - (x_\lambda^* - \frac{1}{L}\nabla f(x_\lambda^*))\|^2 \\ = &\left[\frac{1}{L^2}\|\nabla f(x_\lambda^*)\|^2 - \frac{2}{L}\langle \nabla f(x_\lambda^*), x_\lambda^* \rangle \right] + \frac{2}{L}\left[\langle \nabla f(x_\lambda^*), x \rangle + \frac{L}{2}\|x - x_\lambda^*\|^2 \right]. \end{aligned}$$

Hence,

$$([x_\lambda^*]_i, [w_\lambda^*]_i) \in \underset{(x,w)\in\Omega}{\text{argmin}} \left\{ \frac{L}{2}\|x - (x_\lambda^* - \frac{1}{L}\nabla f(x_\lambda^*))\|^2 + \lambda\|w \circ x\|_1 \right\}.$$

So

$$([x_\lambda^*]_i, [w_\lambda^*]_i) = \begin{cases} ([\Pi_B(y)]_i, 0), & \text{if } \ i \in A; \\ ([\Pi_B(\text{soft}(y))]_i, 1), & \text{otherwise,} \end{cases}$$

where $y = x_\lambda^* - \frac{1}{L}\nabla f(x_\lambda^*)$ and $\text{soft}(y) = \text{sign}(y) \circ \max\{|y| - \lambda/L, 0\}$ is a soft thresholding operator (the proof can be found in Lemma 6(i)). □

Next, based on Lemma 4, we prove that there exists $\lambda \neq +\infty$ such that $\|w_\lambda^* \circ x_\lambda^*\|_1 = 0$. Since $\nabla f(x)$ is Lipschitz continuous, $\|\nabla f(x)\|_\infty$ is bounded for all $x \in B$. Without loss of generality, denote $\lambda_f = \max_{x \in B} \|\nabla f(x)\|_\infty$, then $\lambda_f \neq +\infty$. Moreover, we have the following result.

Corollary 2. *For any $\lambda > \lambda_f$, $\|w_\lambda^* \circ x_\lambda^*\|_1 = 0$, where $(x_\lambda^*, w_\lambda^*)$ is an optimal solution of problem* (4).

Proof. We prove it by contradiction. Suppose that there exists $\lambda > \lambda_f$ such that $\|w_\lambda^* \circ x_\lambda^*\|_1 \neq 0$. Then $\|w_\lambda^* \circ x_\lambda^*\|_1 > 0$. By Lemma 4, $(x_\lambda^*, w_\lambda^*)$ satisfies

$$([x_\lambda^*]_i, [w_\lambda^*]_i) = \begin{cases} ([\Pi_B(y)]_i, 0), & \text{if } \ i \in A; \\ ([\Pi_B(\text{soft}(y))]_i, 1), & \text{otherwise,} \end{cases}$$

where $y = x_\lambda^* - \frac{1}{L}\nabla f(x_\lambda^*)$ and $\text{soft}(y) = \text{sign}(y) \circ \max\{|y| - \lambda/L, 0\}$ is a soft thresholding operator. By the assumption $\|w_\lambda^* \circ x_\lambda^*\|_1 \neq 0$ and the above soft thresholding operator, there exists $i \in \bar{A}$ such that $[x_\lambda^*]_i = [\Pi_B(\text{soft}(y))]_i \neq 0$. Then $|[x_\lambda^*]_i| \leq [|y| - \lambda/L]_i = |[x_\lambda^* - \frac{1}{L}\nabla f(x_\lambda^*)]_i| - \lambda/L$, which leads to $|[\nabla f(x_\lambda^*)]_i| \geq \lambda$. So $|[\nabla f(x_\lambda^*)]_i| \geq \lambda$. However, since $\lambda_f \geq \|\nabla f(x_\lambda^*)\|_\infty$, hence $\lambda_f \geq \lambda$, which contradicts the assumption that $\lambda > \lambda_f$. □

Remark. Corollary 2 indicates that, there exists $\lambda \neq +\infty$ such that $\|w_\lambda^* \circ x_\lambda^*\|_1 = 0$, where $(x_\lambda^*, w_\lambda^*)$ is an optimal solution of problem (4). Since $\lambda_f \neq +\infty$, we let $\lambda_{f+} > \lambda_f$ and $\lambda_{f+} \neq +\infty$. Moreover, let λ^* be the smallest value in the set $J = \{\lambda : \lambda \in \text{argmax}_{\lambda \geq 0} \ g(\lambda)\}$. Then $\lambda^* \leq \lambda_{f+} \neq +\infty$ by Corollaries 1 and 2. Next, we give another relation between λ and $(x_\lambda^*, w_\lambda^*)$.

Lemma 5. *Let λ^* be the smallest value in the set $J = \{\lambda : \lambda \in \text{argmax}_{\lambda \geq 0} g(\lambda)\}$. Then $(\lambda^* - \lambda)\|w_\lambda^* \circ x_\lambda^*\|_1 \ \geq 0$.*

Proof. Let (x^*, w^*) be a global minimizer of $F_{\lambda^*}(x, w)$ for $(x, w) \in \Omega$. Then for any feasible solution $(x, w) \in \Omega$, we have

$$f(x) + \lambda^* \|w \circ x\|_1 \geq f(x^*) + \lambda^* \|w^* \circ x^*\|_1. \tag{9}$$

In addition, by the definition of λ^*, $g(\lambda^*) \geq g(\lambda)$. That is,

$$f(x^*) + \lambda^* \|w^* \circ x^*\|_1 \geq f(x_\lambda^*) + \lambda \|w_\lambda^* \circ x_\lambda^*\|_1.$$

Substituting (x, w) by $(x_\lambda^*, w_\lambda^*)$ in (9), the above two inequalities imply that

$$\lambda^* \|w_\lambda^* \circ x_\lambda^*\|_1 \geq \lambda \|w_\lambda^* \circ x_\lambda^*\|_1,$$

which is $(\lambda^* - \lambda)\|w_\lambda^* \circ x_\lambda^*\|_1 \ \geq 0$. □

Since $\|w \circ x\|_1 \geq 0$ for all $(x, w) \in \Omega$, Lemma 5 indicates that, if $\lambda > \lambda^*$, then $\|w_\lambda^* \circ x_\lambda^*\|_1 = 0$, which means w_λ^* and x_λ^* are complementary. Moreover, Lemma 5 implies the following corollary.

Corollary 3. *Suppose for some* $\lambda_1 \geq 0$, $\|w_{\lambda_1}^* \circ x_{\lambda_1}^*\|_1 = 0$. *Then for any* $\lambda_2 > \lambda_1$, $\|w_{\lambda_2}^* \circ x_{\lambda_2}^*\|_1 = 0$ *and* $f(x_{\lambda_1}^*) = f(x_{\lambda_2}^*)$.

Proof. By Lemma 5, we can obtain that for any $\lambda_2 > \lambda_1$, $\|w_{\lambda_2}^* \circ x_{\lambda_2}^*\|_1 = 0$. Combining the following two inequalities

$$\begin{cases} f(x_{\lambda_2}^*) + \lambda_2 \|w_{\lambda_2}^* \circ x_{\lambda_2}^*\|_1 \leq f(x_{\lambda_1}^*) + \lambda_2 \|w_{\lambda_1}^* \circ x_{\lambda_1}^*\|_1 \\ f(x_{\lambda_1}^*) + \lambda_1 \|w_{\lambda_1}^* \circ x_{\lambda_1}^*\|_1 \leq f(x_{\lambda_2}^*) + \lambda_1 \|w_{\lambda_2}^* \circ x_{\lambda_2}^*\|_1 \end{cases} \tag{10}$$

and $\|w_{\lambda_1}^* \circ x_{\lambda_1}^*\|_1 = 0$, $\|w_{\lambda_2}^* \circ x_{\lambda_2}^*\|_1 = 0$, we get $f(x_{\lambda_1}^*) = f(x_{\lambda_2}^*)$. □

Since the Lagrange dual function $g(\lambda)$ is concave and by the properties above, $g(\lambda)$ can roughly be depicted as in Fig. 1, where $\lambda^* \neq +\infty$ is the point such that (x^*, w^*) is an optimal solution of problem (4). So it is necessary to solve problem (4).

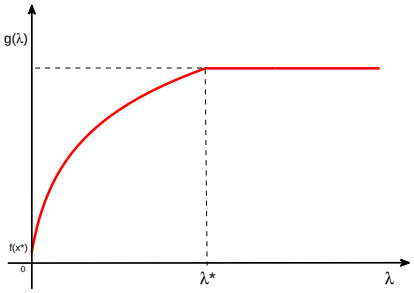

Figure 1: The function $g(\lambda)$ for a fixed sparsity level s.

3 Weighted thresholding method

3.1 Weighted thresholding operator

To solve problem (4), we use the second order Taylor expansion to approximate $f(x)$ at the current solution x^k, and get

$$H(x, w, x^k) = f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2}\|x - x^k\|^2 + \lambda \|w \circ x\|_1.$$

Then minimizing $H(x, w, x^k)$ over $(x, w) \in \Omega$ is the same as

$$\min_{(x,w)\in\Omega} \ \phi(x, w) = \frac{L}{2}\|x - y^k\|^2 + \lambda \|w \circ x\|_1, \tag{11}$$

where $y^k = x^k - \frac{1}{L}\nabla f(x^k)$.

Obviously, $\phi(x, w)$ is a separable function. Let $\phi(x, w) = \sum_{i=1}^{n} \phi_i(x_i, w_i)$, where

$$\phi_i(x_i, w_i) = \frac{L}{2}|x_i - y_i^k|^2 + \lambda|w_i x_i|.$$

Then we can obtain the following result.

Lemma 6. *Problem* (11) *has a closed form solution* (x^*, w^*)*:*

$$(x_i^*, w_i^*) = \begin{cases} ([\Pi_B(y^k)]_i, 0), & \text{if } \ i \in A; \\ ([\Pi_B(\text{soft}(y^k))]_i, 1), & \text{otherwise,} \end{cases} \tag{12}$$

$i = 1, \cdots, n$. *In* (12),

$$\text{soft}(y^k) = \text{sign}(y^k) \circ \max\{|y^k| - \lambda/L, 0\}$$

is the soft thresholding operator [27], and $A \subseteq \{1, 2, \ldots, n\}$ *is the index set corresponding to the* s *largest values of* $\{v_i\}_{i=1}^n$*, where* $v_i = \phi_i([\Pi_B(\text{soft}(y^k))]_i, 1) - \phi_i([\Pi_B(y^k)]_i, 0)$.

Proof. First, we consider the problem

$$\min \ \phi(x, w) = \frac{L}{2}\|x - y^k\|^2 + \lambda\|w \circ x\|_1 \ \ s.t. \ \ w \in \{0, 1\}^n \ , \ x \in B, \tag{13}$$

where $B = \{x \in \mathbb{R}^n : l \leq x \leq u\}$, which is obtained by omitting the constraint $\|1 - w\|_0 \leq s$ in problem (11). Obviously, this problem can be decomposed equivalently into

$$\min \ \phi_i(x_i, w_i) = \frac{L}{2}|x_i - y_i^k|^2 + \lambda w_i|x_i| \ \ s.t. \ w_i \in \{0, 1\}, \ l_i \leq x_i \leq u_i, \tag{14}$$

$i = 1, 2, ..., n$.

We solve problem (14) by analyzing the following two cases.

Case 1). If $w_i = 0$, then problem (14) is

$$\min_{l_i \leq x_i \leq u_i} \quad \frac{L}{2}|x_i - y_i^k|^2, \tag{15}$$

whose minimal solution is $x_i = [\Pi_B(y^k)]_i$, and the minimum value is $\phi_i([\Pi_B($ $y^k)]_i, 0)$.

Case 2). If $w_i = 1$, then problem (14) is

$$\min_{l_i \leq x_i \leq u_i} \quad \frac{L}{2}|x_i - y_i^k|^2 + \lambda|x_i|, \tag{16}$$

whose minimal solution is $x_i = [\Pi_B(\text{soft}(y^k))]_i$, and the corresponding minimal value is $\phi_i([\Pi_B(\text{soft}(y^k))]_i, 1)$.

Since the objective function of problem (15) is not greater than that of problem (16), it holds that

$$\phi_i([\Pi_B(y^k)]_i, 0) \leq \phi_i([\Pi_B(\text{soft}(y^k))]_i, 1)$$

for all $i = 1, 2, \cdots, n$. And for $\lambda > 0$, the equality holds if and only if $y_i^k = 0$.

Let $A \subseteq \{1, 2, \ldots, n\}$ be the index set corresponding to the s largest values of $\{v_i\}_{i=1}^n$, where $v_i = \phi_i([\Pi_B(\text{soft}(y^k))]_i, 1) - \phi_i([\Pi_B(y^k)]_i, 0)$. Next, we prove that if we set $w_i = 0$ for all $i \in A$, and take $w_i = 1$ for the others, then we can obtain a minimal solution of problem (11). That is to say, equation (12) is a minimal solution of problem (11).

Suppose that (x, w) is a feasible solution of problem (11). Denote $F = \{i : w_i = 0\}$. Then $|F| \leq s = |A|$ implies that

$$|A \cap \bar{F}| = |A| - |A \cap F| \geq |F| - |F \cap A| = |F \cap \bar{A}|.$$

For convenience, we set $\check{x}_i = [\Pi_B(y^k)]_i$ and $\tilde{x}_i = [\Pi_B(\text{soft}(y^k))]_i$, then $\phi_i(\check{x}_i, 0) = \min_{l_i \leq x_i \leq u_i} \phi_i(x_i, 0)$ and $\phi_i(\tilde{x}_i, 1) = \min_{l_i \leq x_i \leq u_i} \phi_i(x_i, 1)$ by cases 1) and 2). By the definition of (x^*, w^*) and the analyses of cases 1) and 2), we can obtain

$$\begin{aligned}
&\phi(x, w) - \phi(x^*, w^*) \\
=&(\sum_{i \in A \cap F} + \sum_{i \in A \cap \bar{F}} + \sum_{i \in \bar{A} \cap F} + \sum_{i \in \bar{A} \cap \bar{F}})(\phi_i(x_i, w_i) - \phi_i(x_i^*, w_i^*)) \\
=& \sum_{i \in A \cap F} (\phi_i(x_i, 0) - \phi_i(\check{x}_i, 0)) + \sum_{i \in A \cap \bar{F}} (\phi_i(x_i, 1) - \phi_i(\check{x}_i, 0)) \\
&+ \sum_{i \in \bar{A} \cap F} (\phi_i(x_i, 0) - \phi_i(\tilde{x}_i, 1)) + \sum_{i \in \bar{A} \cap \bar{F}} (\phi_i(x_i, 1) - \phi_i(\tilde{x}_i, 1)) \\
\geq& \sum_{i \in A \cap \bar{F}} (\phi_i(x_i, 1) - \phi_i(\check{x}_i, 0)) + \sum_{i \in \bar{A} \cap F} (\phi_i(x_i, 0) - \phi_i(\tilde{x}_i, 1)) \\
\geq& \sum_{i \in A \cap \bar{F}} (\phi_i(\tilde{x}_i, 1) - \phi_i(\check{x}_i, 0)) + \sum_{i \in \bar{A} \cap F} (\phi_i(\check{x}_i, 0) - \phi_i(\tilde{x}_i, 1)) \\
=& \sum_{i \in A \cap \bar{F}} (\phi_i(\tilde{x}_i, 1) - \phi_i(\check{x}_i, 0)) - \sum_{i \in \bar{A} \cap F} (\phi_i(\tilde{x}_i, 1) - \phi_i(\check{x}_i, 0)) \geq 0,
\end{aligned}$$

where the last inequality comes from the definition of A and the relation $|A \cap \bar{F}| > |F \cap \bar{A}|$. Hence, (x^*, w^*) is an optimal solution of problem (11). □

For the convenience of description, we call the closed form solution in Lemma 6 as a weighted thresholding operator, and denote it by $T_{\lambda,L}(x^k)$:

$$[T_{\lambda,L}(x^k)]_i = \begin{cases} ([\Pi_B(y^k)]_i, 0), & \text{if } \ i \in A; \\ ([\Pi_B(\text{soft}(y^k))]_i, 1), & \text{otherwise}, \end{cases} \tag{17}$$

$i = 1, \cdots, n$, where $y^k = x^k - \frac{1}{L}\nabla f(x^k)$, $A \subseteq \{1, 2, \ldots, n\}$ is the index set corresponding to the s largest values of $\{v_i\}_{i=1}^n$, $v_i = \phi_i([\Pi_B(\text{soft}(y^k))]_i, 1) - \phi_i([\Pi_B(y^k)]_i, 0)$.

3.2 Weighted thresholding method

In this subsection, we propose a weighted thresholding method for problem (4).

Algorithm 1: Weighted Thresholding Method

Input: s, λ, L, x^0. //$L > L_f$
Output: $\hat{x}$, $\hat{w}$.
1: initialize $k \leftarrow 0$;
2: **repeat**
3: $\quad (x^{k+1}, w^{k+1}) \leftarrow T_{\lambda,L}(x^k)$;
4: $\quad k \leftarrow k+1$;
5: **until** some termination condition is reached
6: $\hat{x} \leftarrow x^k$, $\hat{w} \leftarrow w^k$.

Remark. In Algorithm 1, the Lipschitz constant L_f is generally unknown, and it might be hard to calculate directly a fixed value of L such that $L > L_f$. Hence, we will use a line search strategy to explore a feasible value of L in Section 5 in detail.

Theorem 1. *Let $\eta = L - L_f > 0$, and let $\{x^k\}$ be the sequence generated by Algorithm 1. We have*

(i) the sequence $\{F_\lambda(x^k, w^k)\}$ is nonincreasing and satisfies that

$$F_\lambda(x^k, w^k) - F_\lambda(x^{k+1}, w^{k+1}) \geq \frac{\eta}{2}\|x^{k+1} - x^k\|^2; \tag{18}$$

(ii) the sequence $\{x^k\}$ is convergent.

Proof. (i) Since $\nabla f(x)$ is Lipschitz continuous, it holds that

$$f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^T(x^{k+1} - x^k) + \frac{L_f}{2}\|x^{k+1} - x^k\|^2.$$

Then by $L > L_f$, we have

$$\begin{aligned}
F_\lambda(x^{k+1}, w^{k+1}) &= f(x^{k+1}) + \lambda\|w^{k+1} \circ x^{k+1}\|_1 \\
&\leq \overbrace{f(x^k) + \nabla f(x^k)^T(x^{k+1} - x^k) + \frac{L_f}{2}\|x^{k+1} - x^k\|^2 + \lambda\|w^{k+1} \circ x^{k+1}\|_1}^{a} \\
&\leq \underbrace{f(x^k) + \nabla f(x^k)^T(x^{k+1} - x^k) + \frac{L}{2}\|x^{k+1} - x^k\|^2 + \lambda\|w^{k+1} \circ x^{k+1}\|_1}_{b} \\
&\leq f(x^k) + \lambda\|w^k \circ x^k\|_1 = F_\lambda(x^k, w^k),
\end{aligned}$$

where the last inequality follows from that (x^{k+1}, w^{k+1}) is an optimal solution of problem (11). The above inequalities imply that

$$\begin{aligned}
F_\lambda(x^k, w^k) - F_\lambda(x^{k+1}, w^{k+1}) \geq b - a &= \frac{(L - L_f)}{2}\|x^{k+1} - x^k\|^2 \\
&\geq \frac{\eta}{2}\|x^{k+1} - x^k\|^2.
\end{aligned} \tag{19}$$

(ii) Since f is bounded below, $F_\lambda(x,w)$ is also bounded below, which together with the monotonicity of $\{F_\lambda(x^k,w^k)\}$ implies that the sequence $\{F_\lambda(x^k,w^k)\}$ is convergent. Furthermore, by (19), we obtain

$$\lim_{k\to\infty}\|x^{k+1}-x^k\|=0,$$

which together with the boundedness of $\{x^k\}$ implies that $\{x^k\}$ converges. □

Suppose that $(x_\lambda^*,w_\lambda^*)$ is an optimal solution of problem (4) and $\{(x^k,w^k)\}$ is the sequence generated by Algorithm 1. Specially, once there exists $k\in\mathbb{N}$ such that $(x^k,w^k)=(x_\lambda^*,w_\lambda^*)$, we can get that the sequence $\{(x^k,w^k)\}$ is convergent to $(x_\lambda^*,w_\lambda^*)$ by the following lemma.

Lemma 7. *Suppose $(x_\lambda^*,w_\lambda^*)$ is an optimal solution of problem (4) and $\{(x^k,w^k\}$ is the sequence generated by Algotithm1. Then we have that, once there exists $k\in\mathbb{N}$ such that $(x^k,w^k)=(x_\lambda^*,w_\lambda^*)$, then $(x^{k+1},w^{k+1})=(x_\lambda^*,w_\lambda^*)$.*

Proof. Since $(x_\lambda^*,w_\lambda^*)$ is an optimal solution of problem (4), it holds that

$$(x_\lambda^*,w_\lambda^*)\in\underset{(x,w)\in\Omega}{\operatorname{argmin}}\left\{\langle\nabla f(x_\lambda^*),x-x_\lambda^*\rangle+\frac{L}{2}\|x-x_\lambda^*\|^2+\lambda\|w\circ x\|_1\right\},$$

by the proof Lemma 4. Thus, once there exists $k\in\mathbb{N}$ such that $(x^k,w^k)=(x_\lambda^*,w_\lambda^*)$, $(x^{k+1},w^{k+1})=(x_\lambda^*,w_\lambda^*)$ by step 3 of Algorithm 1. □

In Subsection 2.3, we have proved that problems (1) and (4) are equivalent if λ is given properly. Thus with a proper λ, if $(x_\lambda^*,w_\lambda^*)$ is an optimal solution of problem (4), then x_λ^* is also an optimal solution of problem (1), and $\|w_\lambda^*\circ x_\lambda^*\|_1=0$.

Next, we study the bound between an optimal solution x_λ^* of problem (1) and a solution generated by Algorithm 1, when $f(x)=\|Ax-b\|^2$. We recall the definition of restricted isometry property for matrix $A\in\mathbb{R}^{m\times n}$.

Definition 1. *(RIP [29]). For $s\in\{1,2,...,n\}$, the restricted isometry constant is the smallest positive number δ_s such that*

$$(1-\delta_s)\|x\|^2\le\|Ax\|^2\le(1+\delta_s)\|x\|^2$$

holds for all s-sparse vector $x\in\mathbb{R}^n$, i.e., $\|x\|_0\le s$.

Under the RIP property, we have the following result.

Theorem 2. *Suppose that $f(x)=\|Ax-b\|^2$, $L=1$, and x_λ^* is an optimal solution of problem (1). Let (x^k,w^k) be the solution obtained after the k-th iteration of Algorithm 1. If $\delta_{3s}<1/2$, then we have*

$$\|x^k-x_\lambda^*\|\le(2\delta_{3s})^k\|x^0-x^*\|+\frac{2\sqrt{1+\delta_{2s}}}{1-2\delta_{3s}}\|e^*\|-\varepsilon. \tag{20}$$

where $e^* = b - Ax^*_\lambda$, $\varepsilon = \sum_{i=1}^{k} \varepsilon_i$, *and* $\varepsilon_i = 2(2\delta_{3s})^{k-i}\lambda\frac{\|w^i \circ x^i\|_1}{\|x^i - x^*_\lambda\|}$.

Proof. Without loss of generality, suppose $x^i \neq x^*_\lambda$ for any $i \in \{1, 2, 3, ..., k\}$. By the optimality of x^k, we have

$$\frac{1}{2}\|x^k - y^{k-1}\|^2 + \lambda\|w^k \circ x^k\|_1 \leq \frac{1}{2}\|x^*_\lambda - y^{k-1}\|^2 + \lambda\|w^*_\lambda \circ x^*_\lambda\|_1,$$

where $y^{k-1} = x^{k-1} - \frac{1}{L}\nabla f(x^{k-1})$. Moreover,

$$\|x^k - y^{k-1}\|^2 = \|x^k - x^*_\lambda\|^2 + \|x^*_\lambda - y^{k-1}\|^2 + 2\langle x^k - x^*_\lambda, x^*_\lambda - y^{k-1}\rangle.$$

Hence

$$\begin{aligned}
&\|x^k - x^*_\lambda\|^2 \\
\leq& 2\langle x^k - x^*_\lambda, y^{k-1} - x^*_\lambda\rangle + 2\lambda(\|w^*_\lambda \circ x^*_\lambda\|_1 - \|w^k \circ x^k\|_1) \\
=& 2\langle x^k - x^*_\lambda, y^{k-1} - x^*_\lambda\rangle - 2\lambda\|w^k \circ x^k\|_1.
\end{aligned}$$

Furthermore, by setting $\varepsilon_k = 2\lambda\frac{\|w^k \circ x^k\|_1}{\|x^k - x^*_\lambda\|}$, we get

$$\begin{aligned}
&\|x^k - x^*_\lambda\| \\
\leq& 2\langle x^k - x^*_\lambda, \frac{y^{k-1} - x^*_\lambda}{\|x^k - x^*_\lambda\|}\rangle - \varepsilon_k \\
=& 2\langle x^{k-1} - A^T(Ax^{k-1} - b) - x^*_\lambda, \frac{x^k - x^*_\lambda}{\|x^k - x^*_\lambda\|}\rangle - \varepsilon_k \\
=& 2\langle x^{k-1} - A^T(Ax^{k-1} - (Ax^*_\lambda + e^*)) - x^*_\lambda, \frac{x^k - x^*_\lambda}{\|x^k - x^*_\lambda\|}\rangle - \varepsilon_k \\
\leq& 2(\delta_{3s}\|x^{k-1} - x^*_\lambda\| + \sqrt{1 + \delta_{2s}}\|e^*\|) - \varepsilon_k,
\end{aligned}$$

where the last inequality follows from Theorem 4 of [24]. Expanding the last term and computing the power series, the inequality (20) holds. □

So far, we have presented a weighted thresholding method for problem (4). Similar to the regularization based methods [22, 25, 26], it is difficult to choose a proper regularization parameter λ for problem (1). In the next section, we analyze properties of the Lagrange dual function of problem (3) for guiding the choice of the value of parameter λ.

4 Weighted thresholding homotopy method

4.1 Weighted thresholding homotopy method

Suppose that $(x^*_\lambda, w^*_\lambda)$ is an optimal solution of problem (4). Then by the properties of Lagrange dual problem (5) in Subsection 2.3, we know that

there exists $\lambda \neq +\infty$ such that $(x^*_\lambda, w^*_\lambda)$ is also an optimal solution of problem (1). Moreover, for any fixed λ, the feasible region of problem (4) becomes bigger with the increase of s, leading to that the minimum objective function value of problem (4), i.e., $g(\lambda)$, becomes smaller. Combining Fig. 1, the picture of $g(\lambda)$ about λ and s can be roughly depicted as Fig. 2, where $\lambda_{f+} \neq +\infty$ denotes a number that is bigger than λ_f.

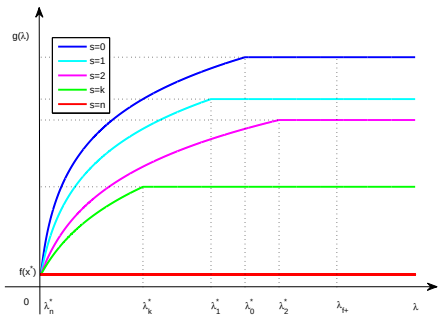

Figure 2: Function $g(\lambda)$ with respect to a series of sparsity levels s.

In Fig. 2, for any given s, $(x^*_\lambda, w^*_\lambda)$ is an optimal solution of problem (4) if $\lambda \geq \lambda_{f+}$ by Corollary 2, and then x^*_λ is also an optimal solution of problem (1) by Corollary 1. Unfortunately, λ_{f+} is unknown. So we use a homotopy technique to find a proper λ for the target sparsity level s.

The homotopy idea is not new. It has been widely used for the l_1-regularized least-squares problem [9, 12, 22, 25]. In this subsection, we develop a weighted thresholding method with homotopy technique which is denoted as WTH. The key idea of the WTH method is to solve problem (4) by tracing a path of solutions with varying (s, λ), where (s, λ) starts from $(0, \lambda_0)$ and gradually increase λ and s simultaneously, until target values of sparsity level s and parameter λ are reached. For each fixed $s = s_k$, $\lambda = \lambda_k$, we use the weighted thresholding method (Algorithm 1) to solve problem (4). The homotopy method is outlined in Algorithm 2.

Algorithm 2: Weighted Thresholding Homotopy Method (WTH)

Input: L, x^0, λ_0, N is a positive integer; // $L > L_f$

Output: $\hat{x}$, $\hat{w}$;

1: initialize $k \leftarrow 0$, $s_0 \leftarrow 0$, $\rho \geq 1$, $\epsilon > 0$, $\gamma > 0$, $\Delta \leftarrow \lceil s/N \rceil$;
2: **for** $k = 1 : N$ **do**
3: $\quad i \leftarrow 0$;
4: $\quad x^{k,i} \leftarrow x^{k-1}$;
5: $\quad s_k \leftarrow \min\{s_{k-1} + \Delta, s\}$;
6: $\quad \lambda_k \leftarrow \rho\lambda_{k-1}$;
7: $\quad$ **repeat**
8: $\quad\quad (x^{k,i+1}, w^{k,i+1}) \leftarrow T_{\lambda_k, L}(x^{k,i})$;
9: $\quad\quad i \leftarrow i + 1$;
10: $\quad$ **until** $\|x^{k,i} - x^{k,i+1}\| / \max\{\|x^{k,i}\|, 10^{-6}\} < \epsilon$
11: $\quad x^k \leftarrow x^{k,i}$, $w^k \leftarrow w^{k,i}$;
12: **end for**
13: $\hat{x} \leftarrow x^N$, $\hat{w} \leftarrow w^N$.

Next, we prove the convergence of Algorithm 2. Firstly, we introduce the definition of L-stationary point of problem (1).

Definition 2. *(L-Stationarity [2]). Suppose that $L > 0$. A vector $x \in C_s \cap B$ is called an L-stationary point of problem* (1) *if*

$$x \in \Pi_{C_s \cap B}\Big(x - \frac{1}{L}\nabla f(x)\Big).$$

Theorem 3. *Let $\{x^{k,i}\}$ be the sequence generated by Algorithm 2. Then:*

(i) the sequence $\{F_{\lambda_k}(x^{k,i}, w^{k,i})\}$ is nonincreasing and $\{x^{k,i}\}$ converges for any fixed k.

(ii) let $\lim_{i\to\infty} x^{N,i} = \hat{x}$. If $\|\hat{w} \circ \hat{x}\|_1 = 0$, then $\hat{x}$ is an L-stationary point of problem (1).

Proof. (i) Since the inner loop of Algorithm 2 calls the Weighted thresholding method (Algorithm 1), by Theorem 1 the sequence $\{F_{\lambda_k}(x^{k,i}, w^{k,i})\}$ is nonincreasing and $\{x^{k,i}\}$ converges for each fixed k.

(ii) If $\|\hat{w} \circ \hat{x}\|_1 = 0$, then by line 8 of Algorithm 2, $\hat{w} \in \{0,1\}^n$ and $\|1 - \hat{w}\|_0 \leq s$. Furthermore, by Lemma 1, $\|\hat{x}\|_0 \leq s$. Combining $l \leq \hat{x} \leq u$, we get that $\hat{x}$ is a feasible solution of problem (1). Since $\lim_{i\to\infty} x^{N,i} = \hat{x}$, by line 8 of Algorithm 2 and Lemma 6, we can get that

$$\hat{x} \in \operatorname*{argmin}_{(x,w)\in\Omega} \frac{L}{2}\|x - \hat{y}\|^2 + \lambda_N \|w \circ x\|_1, \tag{21}$$

where $\hat{y} = \hat{x} - \frac{1}{L}\nabla f(\hat{x})$ and $\Omega = \{(x, w) : \|1 - w\|_0 \leq s, w \in \{0,1\}^n, l \leq x \leq u\}$. Next, we prove that $\hat{x}$ is an L-stationary point of problem (1). By the definition of L-stationary point, we just need to prove that

$$\hat{x} \in \Pi_{C_s \cap B}\Big(\hat{x} - \frac{1}{L}\nabla f(\hat{x})\Big) = \Pi_{C_s \cap B}(\hat{y}),$$

where $\Pi_{C_s \cap B}(\hat{y}) = \text{argmin}\{\|x - \hat{y}\|^2 : x \in C_s \cap B\}$. That is to say, $\hat{x}$ is the solution of the following problem

$$\min\{\|x - \hat{y}\|^2 : \|x\|_0 \leq s, l \leq x \leq u\}.$$

Suppose by contradiction that there exists $\tilde{x}$ such that $\|\tilde{x} - \hat{y}\|^2 < \|\hat{x} - \hat{y}\|^2$, where $\|\tilde{x}\|_0 \leq s$ and $l \leq \tilde{x} \leq u$. Since $\|\tilde{x}\|_0 \leq s$, we can construct $\tilde{w}$ as

$$\begin{cases} \tilde{w}_i = 0, & \text{if } \tilde{x}_i = 1; \\ \tilde{w}_i = 1, & \text{if } \tilde{x}_i \neq 0. \end{cases}$$

Then $\tilde{w} \in \{0,1\}^n$ and $\|1 - \tilde{w}\|_0 \leq s$. So $(\tilde{x}, \tilde{w}) \in \Omega$, and

$$\begin{aligned} &\frac{L}{2}\|\tilde{x} - \hat{y}\|^2 + \lambda_N \|\tilde{w} \circ \tilde{x}\|_1 = \frac{L}{2}\|\tilde{x} - \hat{y}\|^2 + 0 \\ &< \frac{L}{2}\|\hat{x} - \hat{y}\|^2 + 0 \leq \frac{L}{2}\|\hat{x} - \hat{y}\|^2 + \lambda_N \|\hat{w} \circ \hat{x}\|_1, \end{aligned}$$

which contradicts Equation (21). Hence the conclusion holds. □

4.2 The choice of λ_0

If $s = 0$, then $w_i = 1, i = 1, 2, ..., n$, meet the constraints of problem (4), and problem (4) becomes an l_1-norm regularized problem

$$\begin{cases} \min \ \psi(x) = f(x) + \lambda \|x\|_1 \\ s.t. \ \ l \leq x \leq u. \end{cases} \tag{22}$$

Then, we prove that for $\lambda \geq \|\nabla f(0)\|_\infty$, $x = 0$ is the optimal solution of problem (22). That is to say, $\lambda = \|\nabla f(0)\|_\infty$ is big enough such that $(x, w) = (0, 1)$ is the optimal solution problem (4) with $s = 0$. Thus, $x = 0$ is the optimal solution of problem (1) with $s = 0$.

Lemma 8. *Suppose that $f(x)$ is a differentiable convex function. For $\lambda \geq \|\nabla f(0)\|_\infty$, $x = 0$ is the optimal solution of problem* (22).

Proof. Let $D = \{x \in \mathbb{R}^n : l \leq x \leq u\}$. Obviously, D is a convex set, and $0 \in D$. Note that in (22), $\psi(x) = f(x) + \lambda\|x\|_1$ is convex. Suppose g is a subgradient of $\|x\|_1$ at $x = 0$, then $\nabla f(0) + \lambda g \in \partial\psi(0)$, and

$$\psi(x) \geq \psi(0) + (\nabla f(0) + \lambda g)^T (x - 0), \ \textit{for any } x \in D.$$

So if $\nabla f(0) + \lambda g = 0$, then $x = 0$ is the optimal solution of problem (22).

Note that, if $l_i \neq 0$ and $u_i \neq 0$, then $g_i \in [-1, 1]$; if $l_i = 0, u_i \neq 0$, then $g_i \in (-\infty, 1]$; if $l_i \neq 0, u_i = 0$, then $g_i \in [-1, +\infty)$. Thus, for any $\lambda \geq \|\nabla f(0)\|_\infty$, there always exists g_i such that

$$\nabla_i f(0) + \lambda g_i = 0.$$

Hence $x = 0$ is the optimal solution of problem (22) for $\lambda \geq \|\nabla f(0)\|_\infty$. □

Based on the above analysis, we set $\lambda_0 = \|\nabla f(0)\|_\infty$ in our computational experiments.

5 Computational experiments

In Algorithm 2, a fixed L is used throughout all iterations. However, it is hard to calculate the minimum Lipschitz constant L_f for a general function [3]. Moreover, it may be too conservative if L_f is set too large [15]. Therefore, we adopt a line search technique to update dynamically the value of L [15]. The practical WTH Method is outlined as Algorithm 3, and called the PWTH algorithm. The line search technique for updating L is between lines 9 and 11.

Algorithm 3: $\{\hat{x}, \hat{w}, \hat{L}\} \leftarrow PWTH(L_1, \lambda_0, x^0)$

Input: L_1, x^0, λ_0, N is a positive integer; $//L_1 \in [L_{min}, L_{max}]$;
Output: $\hat{x}$, $\hat{w}$;

1: initialize $k \leftarrow 0$, $s_0 \leftarrow 0$, $\Delta \leftarrow \lceil s/N \rceil$, $\rho \geq 1$, $\epsilon > 0$, $\gamma > 1$;
2: **for** $k = 1 : N$ **do**
3: $\quad i \leftarrow 0$;
4: $\quad x^{k,i} \leftarrow x^{k-1}$;
5: $\quad s_k \leftarrow \min\{s_{k-1} + \Delta, s\}$;
6: $\quad \lambda_k \leftarrow \rho\lambda_{k-1}$;
7: $\quad$ **repeat**
8: $\quad\quad (x^{k,i+1}, w^{k,i+1}) \leftarrow T_{\lambda_k, L_k}(x^{k,i})$;
9: $\quad\quad L_{k,i} \leftarrow L_k$;
10: $\quad\quad$ **while** $f(x^{k,i+1}) > f(x^{k,i}) + \nabla f(x^{k,i})^T(x^{k,i+1} - x^{k,i}) + \frac{L_{k,i}}{2}\|x^{k,i+1} - x^{k,i}\|^2$ **do**
11: $\quad\quad\quad L_{k,i} \leftarrow \min\{\gamma L_{k,i}, L_{max}\}$; $(x^{k,i+1}, w^{k,i+1}) \leftarrow T_{\lambda_k, L_{k,i}}(x^{k,i})$;
12: $\quad\quad$ **end while**
13: $\quad\quad L_{k,i+1} \leftarrow L_{k,i}$;
14: $\quad\quad i \leftarrow i + 1$;
15: $\quad$ **until** $\|x^{k,i} - x^{k,i+1}\| / \max\{\|x^{k,i}\|, 10^{-6}\} < \epsilon$
16: $\quad x^k \leftarrow x^{k,i}$, $w^k \leftarrow w^{k,i}$, $L_{k+1} \leftarrow L_{k,i}$;
17: **end for**
18: $\hat{x} \leftarrow x^N$, $\hat{w} \leftarrow w^N$, $\hat{L} \leftarrow L_{N+1}$.

We compare our Algorithm 3 with some state-of-the-art methods on the compressed sensing and sparse logistic regression problems, by implementing them on the same sets of test instances. All experiments were performed on a personal computer with an Intel(R) Core(TM) i3-6100 CPU(3.70GHz) and 4.00GB memory, using MATLAB R2016b. We set $l = -\infty$, $u = +\infty$. Unless otherwise stated, all parameters were set as default for the compared methods in the experiments.

5.1 Compressed sensing

In this subsection, we compare our PWTH algorithm with the NIHT method [6] and its accelerated version–the AIHT method [4][1], the ECME method and its accelerated version–the DORE method [21][2]. All of them were tested on the compressed sensing problem, to find an s-sparse solution of the underdetermined linear system

$$\min_{x\in\mathbb{R}^n}\{\frac{1}{2}\|Ax-b\|_2^2 : \|x\|_0 \le s\}.$$

5.1.1 Parameter settings and test instances generation

In the experiments, all nonzero components of the sparse signal x^* (with size $n = 5000$) were drawn identically from the Gaussian distribution, and A (with size 1000×5000) was the Gaussian matrix, whose components obeyed the Gaussian distribution. Then the observations could be obtained by $b = Ax^* + z$, where z denoted the measurement noise. In the noisy case, z was distributed randomly and uniformly over the interval $[-0.01, 0.01]$. In the noiseless case, $z = 0$.

Similar to [21], the mean squared error (MSE) of a signal estimate $\hat{x}$ is defined as

$$MSE = \frac{1}{n}\|\hat{x} - x^*\|_2^2.$$

And the data fidelity of $A\hat{x} - b$ is defined as

$$DF = \|A\hat{x} - b\|^2.$$

We considered a signal x^* being exactly recovered if MSE is less than 10^{-6}.

We set γ=2, and set the initial signal $x^0 = 0$ and

$$L_1 = \max_{1\le j\le n}\|A_j\|^2,$$

which is similar to [25]. As for the termination condition, we used

$$\frac{\|x^{k+1} - x^k\|}{\max\{\|x^k\|, 10^{-6}\}} \le \epsilon,$$

where ϵ is a small constant.

In our algorithm, we initialized s as $s_0 = 0$, and set

$$s_k = \min\{s_{k-1} + \Delta, s\},$$

[1]MATLAB packages for NIHT and AIHT: `http://www.personal.soton.ac.uk/tb1m08/sparsify/sparsify.html`.

[2]MATLAB packages for ECME and DORE: `http://home.eng.iastate.edu/~ald/DORE.htm`.

Table 1: Results of the PWTH algorithm with different ascent speeds (ρ) and outer loop numbers (N).

N		1				2			
ρ	nnz	MSE	DF	RC	time (s)	MSE	DF	RC	time (s)
1	350	0.027627	3392.3895	0	0.95	0.0077919	1007.8058	0.3	1.13
2	350	0.027627	3392.3895	0	0.93	0.0077919	1007.8058	0.3	1.13
5	350	0.027627	3392.3895	0	0.95	0.0077919	1007.8058	0.3	1.14
10	350	0.027627	3392.3895	0	0.96	0.0077919	1007.8058	0.3	1.13
		3				4			
1	350	0.0030167	440.4982	0.6	1.35	0.0020213	327.9655	0.8	2
2	350	0.0030167	440.4982	0.6	1.36	0.0020213	327.9655	0.8	1.98
5	350	0.0030167	440.4982	0.6	1.35	0.0020213	327.9655	0.8	2.02
10	350	0.0030167	440.4982	0.6	1.38	0.0020213	327.9655	0.8	1.99
		7				10			
1	350	0.0012767	239.4922	0.8	2.24	0.002562	293.3209	0.8	2.74
2	350	0.0012767	239.4922	0.8	2.27	0.002562	293.3209	0.8	2.71
5	350	0.0012767	239.4922	0.8	2.27	0.002562	293.3209	0.8	2.73
10	350	0.0012767	239.4922	0.8	2.23	0.002562	293.3209	0.8	2.82
		13				16			
1	350	0.00094499	121.3322	0.9	3.21	0.0025633	327.4766	0.7	3.61
2	350	0.00094499	121.3322	0.9	3.2	0.0025633	327.4766	0.7	3.61
5	350	0.00094499	121.3322	0.9	3.22	0.0025633	327.4766	0.7	3.6
10	350	0.00094499	121.3322	0.9	3.2	0.0025633	327.4766	0.7	3.61
		19				22			
1	350	0.00097549	181.9931	0.8	4.18	0.001734	214.2616	0.8	4.65
2	350	0.00097549	181.9931	0.8	4.22	0.001734	214.2616	0.8	4.65
5	350	0.00097549	181.9931	0.8	4.27	0.001734	214.2616	0.8	4.65
10	350	0.00097549	181.9931	0.8	4.19	0.001734	214.2616	0.8	4.66
		25				28			
1	350	0.0010133	118.0747	0.9	4.82	0.00043542	80.297	0.9	5.38
2	350	0.0010133	118.0747	0.9	4.78	0.00043542	80.297	0.9	5.26
5	350	0.0010133	118.0747	0.9	4.86	0.00043542	80.297	0.9	5.25
10	350	0.0010133	118.0747	0.9	4.8	0.00043542	80.297	0.9	5.38

where $\Delta = \lceil s/N \rceil$ and $k = 1, 2, ..., N$.

To see the effect of parameters ρ and N on our PWTH algorithm, we generated 10 instances with the sparsity level $s = 350$ in the noiseless case, and test the performance of our algorithm. In the experiment, when $k = 1, 2, ..., N-1$, we set $\epsilon = 10^{-5}$, and when $k = N$, $\epsilon = 10^{-8}$. The average number of nonzero components (nnz), MSE, DF and running time are listed in Table 1, where RC denotes the recovery rate.

From Table 1, we can observe that ρ has almost no effect on the quality of the results and running times. By observing Table 1, we can also find that the algorithm has good performance when N is greater than 3. Hence we just set $N = 10$ and $\rho = 1$ in the next experiments. In order to save running time, ϵ is allowed to be some bigger value in the first $N-1$ iterations of outer loop. More specially, we set $\epsilon = 10^{-5} * 10^{\lfloor (N-k)/2 \rfloor}$ $(k = 1, 2, ..., N-1)$ and $\epsilon = 10^{-8}$ in the last outer loop in the following experiments.

5.1.2 Comparison with other methods

In this experiment, we used the same parameters for the noiseless and noise cases for our PWTH algorithm. And all parameters were set as default for the compared methods. We generated 100 test instances for the noisy and noiseless cases, respectively. Fig. 3 shows the average DFs, MSEs, CPU times and the recovery rates of signal as functions of the sparsity level s in the noiseless and noise cases for all compared methods, respectively.

From Fig. $3(a)$, it can be found that for each algorithm, DF becomes larger with the increase of sparsity level s. And for any given sparsity level s, PWTH can obtain smaller DF comparing with other methods. Especially, the objective function values obtained by our PWTH algorithm are less than 10^{-10} when the sparsity level s is less than 320, while the objective function values obtained by the other methods increase gradually after the sparsity level is more than 260. In the noise case, Fig. $3(b)$ shows the similar trend.

Fig. $3(c)$ has the same trend as Fig. $3(a)$. For any given sparsity level s, PWTH can obtain smaller MSE comparing with the other methods. Hence we can conclude that the PWTH algorithm outperforms the other methods in both DF and MSE.

Fig. $3(e)$ shows that all methods can exactly recover sparse signals when the sparsity level s is not greater than 260. In particular, the PWTH algorithm still has high probability of recovery rate when the sparsity level s is 330. While all other methods have low probability of recovery rate when the sparsity level is greater than 300. That is to say, the PWTH algorithm can recover signals with higher sparsity level. The sparsity level improvement of exact recovery is 26.9%.

From Fig. $3(g)$, we can see that when the sparsity level s is not greater than 320, the CPU times of all methods are short except ECME. When the sparsity level s is not greater than 280, AIHT, PWTH and DORE are fast. The running times do not exceed 1 second. When the sparsity level s is in the interval $[280, 320]$, the CPU time of PWTH is between those of AIHT and NIHT, and it still runs fast. Combining Fig. $3(e)$ with Fig. $3(g)$, we can deduce that PWTH can recover signals with higher probability of recovery rate in a short time when the sparsity level s is not greater than 320. The running time of PWTH becomes slower gradually when $s \in [330, 400]$, but all the methods cannot recover the signal.

For the noisy case, we can get similar results from Figs. $3(b)(d)(f)(h)$, and we can conclude that our PWTH algorithm outperforms the other methods.

5.2 Sparse logistic regression

In this subsection, we compare our PWTH algorithm with the GraSP method [1][3] for finding a sparse solution of the logistic regression problem.

Given m samples $\{a_1, a_2, ..., a_m\}$ with n features, there is a label $y_i \in \{0, 1\}$ for every sample $a_i, i = 1, 2, ..., m$. The average logistic loss function is defined as

$$l_{avg}(x) = \frac{1}{m} \sum \log p(a_i, y_i, x),$$

where $p(a_i, y_i, x)$ is the logistic loss function with respect to $x \in \mathbb{R}^p$, which

[3]MATLAB package for GraSP: `http://sbahmani.ece.gatech.edu/GraSP.html`.

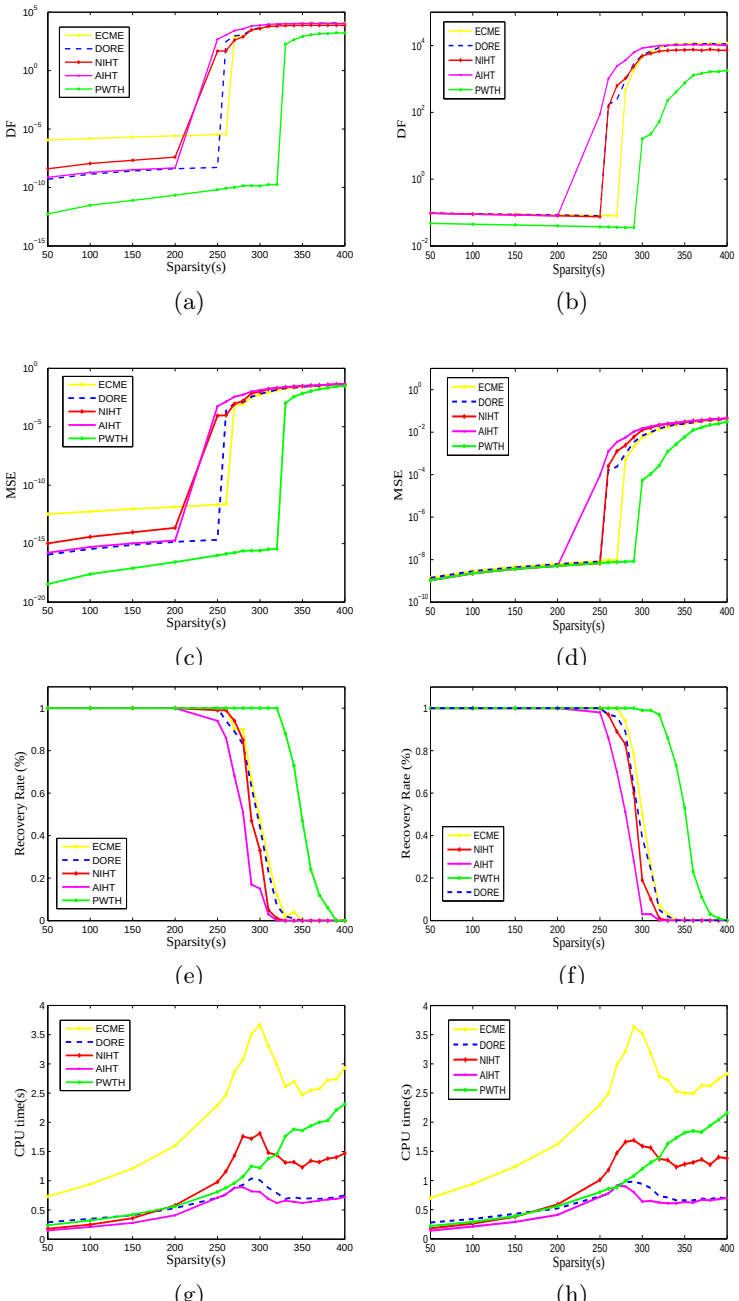

Figure 3: (a)(b) DFs, (c)(d) MSEs, (e)(f) recovery rates, and (g)(h) CPU times as functions of the sparsity level s, where (a),(c),(e) and (g) are for noiseless signal, and (b),(d),(f) and (h) are for noise signal, respectively.

is given by the conditional probability

$$p(a_i, y_i, x) = \Pr\{y_i|a_i; x\} = \frac{\exp(y_i(\langle a_i, x\rangle))}{1 + \exp(\langle a_i, x\rangle)}.$$

Then the sparse logistic regression problem is defined as

$$\min\{l_{avg}(x) : \|x\|_0 \leq s\}.$$

Next, we define the error rate, which is similar to [1, 14] and used to evaluate the solution quality. We first define the logistic classifier

$$\phi(a_i) = sgn(\langle x, a_i\rangle),$$

where

$$sgn(t) = \begin{cases} 1, & \text{if } \ t > 0; \\ 0, & \text{otherwise.} \end{cases}$$

Then the error rate is defined as

$$error = \{\sum \|\phi(a_i) - y_i\|_0/n\} \times 100\%.$$

Since the goal of sparse logistic regression is to classify different samples, we consider that the error rate is the most important assessment criterion.

Similar to the GraSP method with the debiasing process [1] (denoted as GraSPd), we also used the debiasing process for our PWTH algorithm in the experiments, and the corresponding algorithm is denoted as PWTHd. That is to say, after obtaining a current solution $\|x^{k,i}\| \leq s$, we minimized the objective function $f(x)$ restricting to the support set of $x^{k,i}$, and then got a new solution for the next iteration. In the experiments, we set $L_1 = 1$,$\gamma = 2$, $N = 10$, $\rho = 2$, $x^0 = 0$, and $\lambda_0 = \|\nabla f(0)\|_\infty$. In the PWTH algorithm, we set $\epsilon = 10^{-2}$ when $k = 1, 2, ..., N-1$ and $\epsilon = 10^{-3}$ when $k = N$. In the PWTH algorithm with debiasing process, we set $\epsilon = 10^{-3}$ when $k = 1, 2, ..., N-1$ and $\epsilon = 10^{-4}$ when $k = N$.

5.2.1 Synthetic Data

We generated 100 synthetic data in the way described in [1]. The sparse parameter x^* was a vector with dimension $p = 1000$, and with $s \in \{50, 60, 70, ..., 200\}$ nonzero entries generated independently from the standard Gaussian distribution. Each data sample was an independent instance of random vector $a_i \sim \aleph(0, 1), i = 1, 2, ..., n$, and the corresponding data label $y_i \in \{0, 1\}$ was generated randomly according to the Bernoulli distribution

$$\Pr\{y_i = 0|a_i\} = 1/(1 + \exp(\langle x, a_i\rangle)).$$

Fig. 4 depicts the average values of the logistic function l_{avg}, the average error rates (error), the average CPU times, the average number of nonzero

components (nnz) of obtained solutions as functions of the sparsity level s. From Fig. 4(b), we can see that all methods can obtain feasible solutions, and from Fig. 4(d) we can see that the GraSP method is faster than our PWTH algorithm. It is worth mentioning that, we can see from Figs. 4(a) and 4(c) that PWTH outperforms GraSP in terms of l_{avg} and error rate (error). So does the variant–PWTHd. The error rate is decreasing with the increase of sparsity level s for every method. Specifically, the error rate of our PWTHd is 0 when $s \geq 80$, while the error rate of GraSPd is 0 when $s \geq 130$, the sparsity level improvement of exact classification is 38.5%. In general, our method obtains better solution quality in terms of average logistic loss and error rate with or without adding a debiasing process.

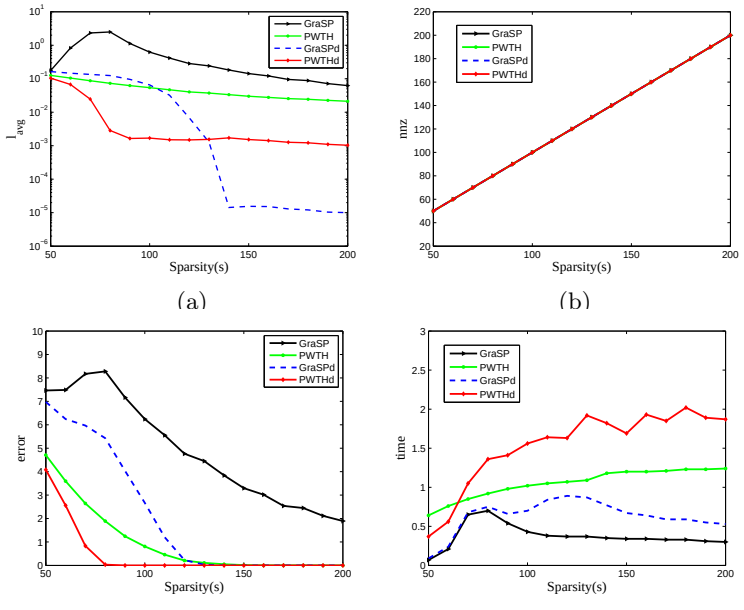

Figure 4: (a) objective function values, (b) nnzs, (c) errors, and (d) CPU times as functions of the sparsity s for sparse logistic regression.

5.2.2 Real Data

We conducted experiments on four real data sets, which are from the NIPS 2003 Workshop on feature extraction [11] and the UCI machine learning benchmark repository [20]. They are ARCENE, DEXTER, Ionosphere and Advertisements data.

Table 2 lists the results of the respective methods on the ARCENE and DEXTER data. The ARCENE data consists of 100 samples and 10000 features. And the DEXTER data consists of 300 samples and 20000 features. These two data sets have more features than samples. For the ARCENE

data, although the PWTH algorithm is slower than the GraSP method, it achieves lower logistic loss and error rate than GraSP. When the sparsity level is not less than 30, the error rates of GraSPd and PWTHd are equal to 0, and the corresponding objective function values are quite small. As for the DEXTER data, GraSPd and PWHTd achieve almost the same solution quality, but GraSPd runs faster.

Table 3 lists the results of the respective methods on the Ionosphere and Advertisements data. The Ionosphere data consists of 351 samples and 34 features. And the Advertisements data consists of 2359 samples and 1558 features. These two data sets have more samples than features. For the Ionosphere data, all methods run fast. Specifically, the running times of GraSP and GraSPd are almost 0, but PWTH and PWTHd achieve almost lower logistic loss and error rate than GraSP and GraSPd, respectively. As for the Advertisements data, GraSP and GraSPd are time-consuming when the sparsity level is not less than 36. In general, PWTH and PWTHd have better solution quality than GraSP and GraSPd in terms of error rate, respectively.

Overall, our PWTH algorithm performs well both on the synthetic and the real data.

Table 2: Test results on the ARCENE and DEXTER data .

		ARCENE			DEXTER		
nnz	algorithm	lavg	error	time (s)	lavg	error	time(s)
10	GraSP	56.9393	56	1.06	0.1964	16.3333	1.14
	PWTH	0.32933	15	0.88	0.34449	17.3333	3.7
	GraSPd	0.50116	26	6.35	0.23798	21	0.4
	PWTHd	0.30936	15	0.34	0.30941	14.6667	0.59
30	GraSP	1289.3621	56	0.85	0.71709	3	1.16
	PWTH	0.25574	8	2.28	0.12399	4	5.47
	GraSPd	1.31E-07	0	0.28	0.0069319	0.66667	2.44
	PWTHd	4.60E-06	0	0.85	0.076531	3.6667	0.8
60	GraSP	120.1791	56	5.01	0.28263	1	1.96
	PWTH	0.20736	2	3.32	0.0783	2	3.04
	GraSPd	3.33E-07	0	0.33	1.21E-07	0	0.22
	PWTHd	3.72E-06	0	1	4.91E-06	0	0.98
100	GraSP	35.7298	41	2.34	0.027757	1	4.43
	PWTH	0.20507	5	5.49	0.044399	0.33333	3.95
	GraSPd	1.05E-07	0	0.24	1.36E-07	0	0.26
	PWTHd	4.24E-06	0	1.19	1.24E-05	0	1.1

6 Conclusions

Since reweighted methods can enhance the sparsity of a solution and improve the signal recovery performance, there are many reweighted methods to solve sparse optimization problems. However, there is no literature concerning the

Table 3: Test results on the Ionosphere and Advertisements data.

	Ionosphere				Advertisements			
algorithm	nnz	lavg	error	time (s)	nnz	lavg	error	time (s)
GraSP	3	0.51701	21.9373	0	3	0.48251	8.5206	0.39
PWTH		0.4444	16.5242	0.03		0.39233	7.9695	7.89
GraSPd		0.5161	22.2222	0		0.63997	6.1043	0.2
PWTHd		0.47149	20.7977	0.05		0.62597	6.9945	0.44
GraSP	11	0.3856	14.245	0.01	36	0.47938	8.6901	18.76
PWTH		0.33831	11.396	0.05		0.36607	8.0543	12.5
GraSPd		0.38133	14.245	0.01		0.38454	4.5358	25.64
PWTHd		0.33681	11.9658	0.07		0.10863	2.8402	2.19
GraSP	14	0.35704	14.245	0.01	67	0.71588	8.3086	19.84
PWTH		0.32526	11.6809	0.04		0.35715	8.0967	19.91
GraSPd		0.31984	11.9658	0.01		0.31631	4.7054	59.3
PWTHd		0.3153	11.9658	0.07		0.10403	2.6706	2.76
GraSP	24	0.35814	12.5356	0.01	197	0.3083	3.9423	381.7
PWTH		0.2991	11.396	0.06		0.36204	8.0543	31.83
GraSPd		0.28185	11.9658	0.01		0.17538	1.2293	303.67
PWTHd		0.28354	10.8262	0.07		0.035593	1.1869	16.01

reweighted strategy for sparsity constrained optimization. In this paper, we reformulated the sparsity constraint as an equivalent weighted l_1-norm constraint in the sparsity constrained optimization problem. To solve the reformulated problem, we investigated the problem in the Lagrange dual framework, and proved that the problem has the strong duality property. We proposed a weighted thresholding method for the Lagrangian problem, which optimizes the weight w and the variable x simultaneously. Moreover, we analyzed the convergence of this method and provided an error bound of the obtained solution under some assumptions.

Basing on the homotopy technique, we proposed the weighted thresholding homotopy (PWTH) method which overcomes the difficulty of choosing the Lagrange multiplier. Moreover, we proved that the method can converge to an L-stationary point under some conditions. We compared our PWTH method with state-of-the-art methods on the compressed sensing and sparse logistic regression problems, by implementing the methods on the same sets of test instances. Computational experiments show that our method performs well both in running time and solution quality. Specifically, our method can solve the problem with higher sparsity level.

References

[1] S. Bahmani, B. Raj, and P. Boufounos. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 14(1):807–841, 2013.

[2] A. Beck and N. Hallak. On the minimization over sparse symmetric sets: projections, optimality conditions, and algorithms. *Mathematics of Operations Research*, 41(1):196–223, 2016.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[4] T. Blumensath. Accelerated iterative hard thresholding. *Signal Processing*, 92(3):752–756, 2012.

[5] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, 2008.

[6] T. Blumensath and M. E. Davies. Normalised itertive hard thresholding: guaranteed stability and performance. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):298–309, 2010.

[7] E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008.

[8] Y. Chen and M. Wang. Worst-case hardness of approximation for sparse optimization with l_0 norm. *http://www.optimization-online.org/DB_FILE/2016/02/5334.pdf*, 2016.

[9] D. L. Donoho and Y. Tsaig. Fast solution of l_1-norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54(11):4789–4812, 2008.

[10] M. Elad. Sparse and redundant representations: From theory to applications in signal and image processing. *Springer-Verlag New York*, 2010.

[11] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the nips 2003 feature selection challenge. *In L. K. Saul, Y. Weiss, and L. Bottou, editors, Advances in Neural Information Processing System 17, pages 545-552. MIT-Press, Cambridge, MA*, 2005.

[12] Y. Jiao, B. Jin, and X. Lu. A primal dual active set with continuation algorithm for the l_0-regularized optimization problem. *Applied and Computational Harmonic Analysis*, 39(3):400–426, 2015.

[13] M. A. Khajehnejad, W. Xu, A. S. Avestimehr, and B. Hassibi. Weighted l_1 minimization for sparse recovery with prior information. In *IEEE International Conference on Symposium on Information Theory*, pages 483–487, June 2009.

[14] K. Koh, S. Kim, and S. Boyd. An interior-point method for large-scale l_1-regularized logistic regression. *Journal of Machine Learning Research*, 8:1519–1555, 2007.

[15] Z. Lu. Iterative hard thresholding methods for l_0 regularized convex cone programming. *Mathematical Programming*, 147(1-2):125–154, 2014.

[16] Z. Lu and Y. Zhang. Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization*, 23(4):2448–2478, 2013.

[17] J. Mairal, F. Bach, and J. Ponce. Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

[18] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[19] D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[20] D. Newman, S. Hrttich, C. Blake, and C. Merz. Uci repository of machine learing databases. *Available at www.ics.uci.edu/ mlearn/MLRepository.html*, 1998.

[21] K. Qiu and A. Dogandžić. Sparse signal reconstruction via ecme hard thresholding. *IEEE Transactions on Signal Processing*, 60(9):4551–4569, 2012.

[22] C. Soussen, J. Idier, J. Duan, and D. Brie. Homotopy based algorithms for l_0-regularized least-squares. *IEEE Transactions on Signal Processing*, 63(13):3301–3316, 2015.

[23] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

[24] S. Xiang, X. Shen, and J. Ye. Efficient nonconvex sparse group feature selection via continuous and discrete optimization. *Artificial Intelligence*, 224:28–50, 2015.

[25] L. Xiao and T. Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.

[26] Z. Xu, X. Chang, F. Xu, and H. Zhang. $l_{1/2}$ regularization: a thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems*, 23(7):1013–1027, 2012.

[27] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang. A survey of sparse representation: Algorithms and applications. *IEEE Access*, 3:490–530, 2015.

[28] Y. Zhao and D. Li. Reweighted l_1-minimization for sparse solutions to underdetermined linear systems. *SIAM Journal on Optimization*, 22(3):1065–1088, 2012.

[29] S. Zhou, N. Xiu, Wang Y, and L. Kong. Exact recovery for sparse signal via weighted l_1 minimization. *arXiv: 1312.2358v2*, 2014.