HHS Public Access

Author manuscript

JMLR Workshop Conf Proc. Author manuscript; available in PMC 2014 October 03.

Published in final edited form as:

JMLR Workshop Conf Proc. 2013 ; 28(2): 37–45.

A General Iterative Shrinkage and Thresholding Algorithm for Non-convex Regularized Optimization Problems

Pinghua Gong,

State Key Laboratory on Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Automation, Tsinghua University, Beijing 100084, China

Changshui Zhang,

State Key Laboratory on Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Automation, Tsinghua University, Beijing 100084, China

Zhaosong Lu,

Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada

Jianhua Z. Huang,

Department of Statistics, Texas A&M University, TX 77843, USA

Jieping Ye

Computer Science and Engineering, Arizona State University, Tempe, AZ 85287, USA

Pinghua Gong: gph08@mails.tsinghua.edu.cn; Changshui Zhang: zcs@mail.tsinghua.edu.cn; Zhaosong Lu: zhaosong@sfu.ca; Jianhua Z. Huang: jianhua@stat.tamu.edu; Jieping Ye: jieping.ye@asu.edu

Abstract

Non-convex sparsity-inducing penalties have recently received considerable attentions in sparse learning. Recent theoretical investigations have demonstrated their superiority over the convex counterparts in several sparse learning settings. However, solving the non-convex optimization problems associated with non-convex penalties remains a big challenge. A commonly used approach is the Multi-Stage (MS) convex relaxation (or DC programming), which relaxes the original non-convex problem to a sequence of convex problems. This approach is usually not very practical for large-scale problems because its computational cost is a multiple of solving a single convex problem. In this paper, we propose a General Iterative Shrinkage and Thresholding (GIST) algorithm to solve the nonconvex optimization problem for a large class of non-convex penalties. The GIST algorithm iteratively solves a proximal operator problem, which in turn has a closed-form solution for many commonly used penalties. At each outer iteration of the algorithm, we use a line search initialized by the Barzilai-Borwein (BB) rule that allows finding an appropriate step size quickly. The paper also presents a detailed convergence analysis of the GIST algorithm. The efficiency of the proposed algorithm is demonstrated by extensive experiments on large-scale data sets.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013.

1. Introduction

Learning sparse representations has important applications in many areas of science and engineering. The use of an l_0-norm regularizer leads to a sparse solution, however the l_0-norm regularized optimization problem is challenging to solve, due to the discontinuity and non-convexity of the l_0-norm regularizer. The l_1-norm regularizer, a continuous and convex surrogate, has been studied extensively in the literature (Tibshirani, 1996; Efron et al., 2004) and has been applied successfully to many applications including signal/image processing, biomedical informatics and computer vision (Shevade & Keerthi, 2003; Wright et al., 2008; Beck & Teboulle, 2009; Wright et al., 2009; Ye & Liu, 2012). Although the l_1-norm based sparse learning formulations have achieved great success, they have been shown to be suboptimal in many cases (Candes et al., 2008; Zhang, 2010b; 2012), since the l_1-norm is a loose approximation of the l_0-norm and often leads to an over-penalized problem. To address this issue, many non-convex regularizers, interpolated between the l_0-norm and the l_1-norm, have been proposed to better approximate the l_0-norm. They include l_q-norm ($0 < q < 1$) (Foucart & Lai, 2009), Smoothly Clipped Absolute Deviation (SCAD) (Fan & Li, 2001), Log-Sum Penalty (LSP) (Candes et al., 2008), Minimax Concave Penalty (MCP) (Zhang, 2010a), Geman Penalty (GP) (Geman & Yang, 1995; Trzasko & Manduca, 2009) and Capped-l_1 penalty (Zhang, 2010b; 2012; Gong et al., 2012a).

Although the non-convex regularizers (penalties) are appealing in sparse learning, it is challenging to solve the corresponding non-convex optimization problems. In this paper, we propose a General Iterative Shrinkage and Thresholding (GIST) algorithm for a large class of non-convex penalties. The key step of the proposed algorithm is to compute a proximal operator, which has a closed-form solution for many commonly used non-convex penalties. In our algorithm, we adopt the Barzilai-Borwein (BB) rule (Barzilai & Borwein, 1988) to initialize the line search step size at each iteration, which greatly accelerates the convergence speed. We also use a non-monotone line search criterion to further speed up the convergence of the algorithm. In addition, we present a detailed convergence analysis for the proposed algorithm. Extensive experiments on large-scale real-world data sets demonstrate the efficiency of the proposed algorithm.

2. The Proposed Algorithm: GIST

2.1. General Problems

—We consider solving the following general problem:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \{ f(\mathbf{w}) = l(\mathbf{w}) + r(\mathbf{w}) \} . \tag{1}$$

We make the following assumptions on the above formulation throughout the paper:

A1 $l(\mathbf{w})$ is continuously differentiable with Lipschitz continuous gradient, that is, there exists a positive constant $\beta(l)$ such that

$$\| \nabla l(\mathbf{w}) - \nabla l(\mathbf{u}) \| \leq \beta(l) \| \mathbf{w} - \mathbf{u} \|, \forall \mathbf{w}, \mathbf{u} \in \mathbb{R}^d .$$

(A2) $r(\mathbf{w})$ is a continuous function which is possibly *non-smooth* and *non-convex*, and can be rewritten as the difference of two convex functions, that is,

$$r(\mathbf{w}) = r_1(\mathbf{w}) - r_2(\mathbf{w}),$$

where $r_1(\mathbf{w})$ and $r_2(\mathbf{w})$ are convex functions.

(A3) $f(\mathbf{w})$ is bounded from below.

Remark 1: We say that $\mathbf{w}^\star$ is a critical point of problem (1), if the following holds (Toland, 1979; Wright et al., 2009):

$$\mathbf{0} \in \nabla l(\mathbf{w}^\star) + \partial r_1(\mathbf{w}^\star) - \partial r_2(\mathbf{w}^\star),$$

where $\partial r_1(\mathbf{w}^\star)$ is the sub-differential of the function $r_1(\mathbf{w})$ at $\mathbf{w} = \mathbf{w}^\star$, that is,

$$\partial r_1(\mathbf{w}^\star) = \{\mathbf{s}: r_1(\mathbf{w}) \geq r_1(\mathbf{w}^\star) + \langle \mathbf{s}, \mathbf{w} - \mathbf{w}^\star \rangle, \forall \mathbf{w} \in \mathbb{R}^d\}.$$

We should mention that the sub-differential is nonempty on any convex function; this is why we make the assumption that $r(\mathbf{w})$ can be rewritten as the difference of two convex functions.

2.2. Some Examples

Many formulations in machine learning satisfy the assumptions above. The following least square and logistic loss functions are two commonly used ones which satisfy assumption **A1**:

$$l(\mathbf{w}) = \frac{1}{2n}\|X\mathbf{w} - \mathbf{y}\|^2 \text{or} \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i \mathbf{x}_i^T \mathbf{w})),$$

where $X = [\mathbf{x}_1^T; \cdots; \mathbf{x}_n^T] \in \mathbb{R}^{n \times d}$ is a data matrix and $\mathbf{y} = [y_1, \cdots, y_n]^T \in \mathbb{R}^n$ is a target vector. The regularizers (penalties) which satisfy the assumption **A2** are presented in Table 1. They are non-convex (except the l_1-norm) and extensively used in sparse learning. The functions $l(\mathbf{w})$ and $r(\mathbf{w})$ mentioned above are nonnegative. Hence, f is bounded from below and satisfies assumption **A3**.

2.3. Algorithm

Our proposed General Iterative Shrinkage and Thresholding (GIST) algorithm solves problem (1) by generating a sequence $\{\mathbf{w}^{(k)}\}$ via:

$$\mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w}} l(\mathbf{w}^{(k)}) + \langle \nabla l(\mathbf{w}^{(k)}), \mathbf{w} - \mathbf{w}^{(k)} \rangle + \frac{t^{(k)}}{2}\|\mathbf{w} - \mathbf{w}^{(k)}\|^2 + r(\mathbf{w}), \quad (2)$$

In fact, problem (2) is equivalent to the following proximal operator problem:

$$\mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w} - \mathbf{u}^{(k)}\|^2 + \frac{1}{t^{(k)}} r(\mathbf{w}),$$

where $\mathbf{u}^{(k)} = \mathbf{w}^{(k)} - \nabla l(\mathbf{w}^{(k)})/t^{(k)}$. Thus, in GIST we first perform a gradient descent along the direction $-\nabla l(\mathbf{w}^{(k)})$ with step size $1/t^{(k)}$ and then solve a proximal operator problem. For all the regularizers listed in Table 1, problem (2) has a closed-form solution (details are provided in the Appendix), although it may be a non-convex problem. For example, for the l_1 and Capped l_1 regularizers, we have closed-form solutions as follows:

$$\ell_1 : w_i^{(k+1)} = \mathrm{sign}(u_i^{(k)}) \max\left(0, \mid u_i^{(k)} \mid - \lambda/t^{(k)}\right),$$

$$\text{Capped } \ell_1 : w_i^{(k+1)} = \begin{cases} x_1, & \text{if } h_i(x_1) \le h_i(x_2), \\ x_2, & \text{otherwise,} \end{cases}$$

where $x_1 = \mathrm{sign}(u_i^{(k)}) \max(\mid u_i^{(k)} \mid, \theta)$, $x_2 = \mathrm{sign}(u_i^{(k)}) \min(\theta, [\mid u_i^{(k)} \mid - \lambda/t^{(k)}]_+)$ and $h_i(x) = 0.5(x - u_i^{(k)})^2 + \lambda/t^{(k)} \min(\mid x \mid, \theta)$. The detailed procedure of the GIST algorithm is presented in Algorithm 1. There are two issues that remain to be addressed: how to initialize $t^{(k)}$ (in Line 4) and how to select a line search criterion (in Line 8) at each outer iteration.

2.3.1. The Step Size Initialization: $1/t^{(k)}$—Intuitively, a good step size initialization strategy at each outer iteration can greatly reduce the line search cost (Lines 5–8) and hence is critical for the fast convergence of the algorithm. In this paper, we propose to initialize the step size by adopting the Barzilai-Borwein (BB) rule (Barzilai & Borwein, 1988), which uses a diagonal matrix $t^{(k)}I$ to approximate the Hessian matrix $\nabla^2 l(\mathbf{w})$ at $\mathbf{w} = \mathbf{w}^{(k)}$. Denote

$$\mathbf{x}^{(k)} = \mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}, \mathbf{y}^{(k)} = \nabla l(\mathbf{w}^{(k)}) - \nabla l(\mathbf{w}^{(k-1)}).$$

Then $t^{(k)}$ is initialized at the outer iteration k as

$$t^{(k)} = \arg\min_{t} \|t\mathbf{x}^{(k)} - \mathbf{y}^{(k)}\|^2 = \frac{\langle \mathbf{x}^{(k)}, \mathbf{y}^{(k)} \rangle}{\langle \mathbf{x}^{(k)}, \mathbf{x}^{(k)} \rangle}.$$

2.3.2. Line Search Criterion—One natural and commonly used line search criterion is to require that the objective function value is monotonically decreasing. More specifically, we propose to accept the step size $1/t^{(k)}$ at the outer iteration k if the following monotone line search criterion is satisfied:

$$f(\mathbf{w}^{(k+1)}) \le f(\mathbf{w}^{(k)}) - \frac{\sigma}{2} t^{(k)} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2, \quad (3)$$

where σ is a constant in the interval (0, 1).

A variant of the monotone criterion in Eq. (3) is a non-monotone line search criterion (Grippo et al., 1986; Grippo & Sciandrone, 2002; Wright et al., 2009). It possibly accepts the step size $1/t^{(k)}$ even if $\mathbf{w}^{(k+1)}$ yields a larger objective function value than $\mathbf{w}^{(k)}$. Specifically, we propose to accept the step size $1/t^{(k)}$, if $\mathbf{w}^{(k+1)}$ makes the objective function value smaller than the maximum over previous m ($m > 1$) iterations, that is,

$$f(\mathbf{w}^{(k+1)}) \leq \max_{i = \max(0, k-m+1), \cdots, k} f(\mathbf{w}^{(i)}) - \frac{\sigma}{2} t^{(k)} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2, \tag{4}$$

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

2.3.3. Convergence Analysis—Inspired by Wright et al. (2009); Lu (2012a), we present detailed convergence analysis under both monotone and non-monotone line search criteria. We first present a lemma which guarantees that the monotone line search criterion in Eq. (3) is satisfied. This is a basic support for the convergence of Algorithm 1.

Lemma 1: Let the assumptions **A1**–**A3** hold and the constant $\sigma \in (0, 1)$ be given. Then for any integer $k \geq 0$, the monotone line search criterion in Eq. (3) is satisfied whenever $t^{(k)} \geq \beta(l)/(1 - \sigma)$.

Proof: Since $\mathbf{w}^{(k+1)}$ is a minimizer of problem (2), we have

$$\langle \nabla l(\mathbf{w}^{(k)}), \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)} \rangle + \frac{t^{(k)}}{2} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2 + r(\mathbf{w}^{(k+1)}) \leq r(\mathbf{w}^{(k)}). \tag{5}$$

It follows from assumption **A1** that

$$l(\mathbf{w}^{(k+1)}) \leq l(\mathbf{w}^{(k)}) + \langle \nabla l(\mathbf{w}^{(k)}), \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)} \rangle + \frac{\beta(l)}{2} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2. \tag{6}$$

Combining Eq. (5) and Eq. (6), we have

$$l(\mathbf{w}^{(k+1)}) + r(\mathbf{w}^{(k+1)}) \leq l(\mathbf{w}^{(k)}) + r(\mathbf{w}^{(k)}) - \frac{t^{(k)} - \beta(l)}{2} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2.$$

It follows that

$$f(\mathbf{w}^{(k+1)}) \leq f(\mathbf{w}^{(k)}) - \frac{t^{(k)} - \beta(l)}{2} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2.$$

Therefore, the line search criterion in Eq. (3) is satisfied whenever $(t^{(k)} - \beta(l))/2 \geq \sigma t^{(k)}/2$, i.e., $t^{(k)} \geq \beta(l)/(1 - \sigma)$. This completes the proof the lemma.

Next, we summarize the boundedness of $t^{(k)}$ in the following lemma.

Lemma 2: For any $k \geq 0$, $t^{(k)}$ is bounded under the monotone line search criterion in Eq. (3).

Proof: It is trivial to show that $t^{(k)}$ is bounded from below, since $t^{(k)} \geq t_{\min}$ ($t_{\min}$ is defined in Algorithm 1). Next we prove that $t^{(k)}$ is bounded from above by contradiction. Assume that there exists a $k \geq 0$, such that $t^{(k)}$ is unbounded from above. Without loss of generality, we assume that $t^{(k)}$ increases monotonically to $+\infty$ and $t^{(k)} \geq \eta\beta(l)/(1-\sigma)$. Thus, the value $t = t^{(k)}/\eta \geq \beta(l)/(1-\sigma)$ must have been tried at iteration k and does not satisfy the line search criterion in Eq. (3). But Lemma 1 states that $t = t^{(k)}/\eta \geq \beta(l)/(1-\sigma)$ is guaranteed to satisfy the line search criterion in Eq. (3). This leads to a contradiction. Thus, $t^{(k)}$ is bounded from above.

Remark 2: We note that if Eq. (3) holds, Eq. (4) is guaranteed to be satisfied. Thus, the same conclusions in Lemma 1 and Lemma 2 also hold under the the non-monotone line search criterion in Eq. (4).

Based on Lemma 1 and Lemma 2, we present our convergence result in the following theorem.

Theorem 1: Let the assumptions **A1**–**A3** hold and the monotone line search criterion in Eq. (3) be satisfied. Then all limit points of the sequence $\{\mathbf{w}^{(k)}\}$ generated by Algorithm 1 are critical points of problem (1).

Proof: Based on Lemma 1, the monotone line search criterion in Eq. (3) is satisfied and hence

$$f(\mathbf{w}^{(k+1)}) \leq f(\mathbf{w}^{(k)}), \forall k \geq 0,$$

which implies that the sequence $f(\mathbf{w}^{(k)})_{k=0,1,\ldots}$ is monotonically decreasing. Let $\mathbf{w}^{\star}$ be a limit point of the sequence $\{\mathbf{w}^{(k)}\}$, that is, there exists a subsequence $\mathcal{K}$ such that

$$\lim_{k \in \mathcal{K} \to \infty} \mathbf{w}^{(k)} = \mathbf{w}^{\star}.$$

Since f is bounded from below, together with the fact that $\{f(\mathbf{w}^{(k)})\}$ is monotonically decreasing, $\lim_{k\to\infty} f(\mathbf{w}^{(k)})$ exists. Observing that f is continuous, we have

$$\lim_{k \to \infty} f(\mathbf{w}^{(k)}) = \lim_{k \in \mathcal{K} \to \infty} f(\mathbf{w}^{(k)}) = f(\mathbf{w}^{\star}).$$

Taking limits on both sides of Eq. (3) with $k \in \mathcal{K}$, we have

$$\lim_{k \in \mathcal{K} \to \infty} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| = 0. \tag{7}$$

Considering that the minimizer $\mathbf{w}^{(k+1)}$ is also a critical point of problem (2) and $r(\mathbf{w}) = r_1(\mathbf{w}) - r_2(\mathbf{w})$, we have

$$\mathbf{0} \in \nabla l(\mathbf{w}^{(k)}) + t^{(k)}(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}) + \partial r_1(\mathbf{w}^{(k+1)}) - \partial r_2(\mathbf{w}^{(k+1)}).$$

Taking limits on both sides of the above equation with $k \in \mathcal{K}$, by considering the semi-continuity of $\partial r_1(\cdot)$ and $\partial r_2(\cdot)$, the boundedness of $t^{(k)}$ (based on Lemma 2) and Eq. (7), we obtain

$$\mathbf{0} \in \nabla l(\mathbf{w}^{\bigstar}) + \partial r_1(\mathbf{w}^{\bigstar}) - \partial r_2(\mathbf{w}^{\bigstar}),$$

Therefore, $\mathbf{w}^{\bigstar}$ is a critical point of problem (1). This completes the proof of Theorem 1.

Based on Eq. (7), we know that $\lim_{k \in \mathcal{K} \to \infty} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2 = 0$ is a necessary optimality condition of Algorithm 1. Thus, $\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2$ is a quantity to measure the convergence of the sequence $\{\mathbf{w}^{(k)}\}$ to a critical point. We present the convergence rate in terms of $\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2$ in the following theorem.

Theorem 2: Let $\{\mathbf{w}^{(k)}\}$ be the sequence generated by Algorithm 1 with the monotone line search criterion in Eq. (3) satisfied. Then for every $n \geq 1$, we have

$$\min_{0 \leq k \leq n} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2 \leq \frac{2(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{\bigstar}))}{n\sigma t_{\min}},$$

where $\mathbf{w}^{\bigstar}$ is a limit point of the sequence $\{\mathbf{w}^{(k)}\}$.

Proof: Based on Eq. (3) with $t^{(k)} \geq t_{\min}$, we have

$$\frac{\sigma t_{\min}}{2}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2 \leq f(\mathbf{w}^{(k)}) - f(\mathbf{w}^{(k+1)}).$$

Summing the above inequality over $k = 0, \cdots, n$, we obtain

$$\frac{\sigma t_{\min}}{2} \sum_{k=0}^{n} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2 \leq f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(n+1)}),$$

which implies that

$$\min_{0 \le k \le n} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2 \le \frac{2(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(n+1)}))}{n\sigma t_{\min}} \le \frac{2(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{\bigstar}))}{n\sigma t_{\min}}.$$

This completes the proof of the theorem.

Under the non-monotone line search criterion in Eq. (4), we have a similar convergence result in the following theorem (the proof uses an extension of argument for Theorem 1 and is omitted).

Theorem 3: Let the assumptions **A1**–**A3** hold and the non-monotone line search criterion in Eq. (4) be satisfied. Then all limit points of the sequence $\{\mathbf{w}^{(k)}\}$ generated by Algorithm 1 are critical points of problem (1).

Note that Theorem 1/Theorem 3 makes sense only if $\{\mathbf{w}^{(k)}\}$ has limit points. By considering one more mild assumption:

A4 $f(\mathbf{w}) \to +\infty$ when $\|\mathbf{w}\| \to +\infty$,

we summarize the existence of limit points in the following theorem (the proof is omitted):

Theorem 4: Let the assumptions **A1**–**A4** hold and the monotone/non-monotone line search criterion in Eq. (3)/Eq. (4) be satisfied. Then the sequence $\{\mathbf{w}^{(k)}\}$ generated by Algorithm 1 has at least one limit point.

2.3.4. Discussions—Observe that $l(\mathbf{w}^{(k)}) + \langle \nabla l(\mathbf{w}^{(k)}), \mathbf{w} - \mathbf{w}^{(k)} \rangle + \frac{t^{(k)}}{2}\|\mathbf{w} - \mathbf{w}^{(k)}\|^2$ can be viewed as an approximation of $l(\mathbf{w})$ at $\mathbf{w} = \mathbf{w}^{(k)}$. The GIST algorithm minimizes an approximate surrogate instead of the objective function in problem (1) at each outer iteration. We further observe that if $t^{(k)} \ge \beta(l)/(1-\sigma) > \beta(l)$ [the sufficient condition of Eq. (3)], we obtain

$$l(\mathbf{w}) \le l(\mathbf{w}^{(k)}) + \langle \nabla l(\mathbf{w}^{(k)}), \mathbf{w} - \mathbf{w}^{(k)} \rangle + \frac{t^{(k)}}{2}\|\mathbf{w} - \mathbf{w}^{(k)}\|^2, \forall \mathbf{w} \in \mathbb{R}^d.$$

It follows that

$$f(\mathbf{w}) = l(\mathbf{w}) + r(\mathbf{w}) \le M(\mathbf{w}, \mathbf{w}^{(k)}), \forall \mathbf{w} \in \mathbb{R}^d,$$

where $M(\mathbf{w}, \mathbf{w}^{(k)})$ denotes the objective function of problem (2). We can easily show that

$$f(\mathbf{w}^{(k)}) = M(\mathbf{w}^{(k)}, \mathbf{w}^{(k)}).$$

Thus, the GIST algorithm is equivalent to solving a sequence of minimization problems:

$$\mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w}} M(\mathbf{w}, \mathbf{w}^{(k)}),\ k = 0, 1, 2, \cdots$$

and can be interpreted as the well-known Majorization and Minimization (MM) technique (Hunter & Lange, 2000).

Note that we focus on the vector case in this paper and the proposed GIST algorithm can be easily extended to the matrix case.

3. Related Work

In this section, we discuss some related algorithms. One commonly used approach to solve problem (1) is the Multi-Stage (MS) convex relaxation (or CCCP, or DC programming) (Zhang, 2010b; Yuille & Rangarajan, 2003; Gasso et al., 2009). It equivalently rewrites problem (1) as

$$\min_{\mathbf{w} \in \mathbb{R}^d} f_1(\mathbf{w}) - f_2(\mathbf{w}),$$

where $f_1(\mathbf{w})$ and $f_2(\mathbf{w})$ are both convex functions. The MS algorithm solves problem (1) by generating a sequence $\{\mathbf{w}^{(k)}\}$ as

$$\mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} f_1(\mathbf{w}) - f_2(\mathbf{w}^{(k)}) - \langle \mathbf{s}_2(\mathbf{w}^{(k)}), \mathbf{w} - \mathbf{w}^{(k)} \rangle, \tag{8}$$

where $\mathbf{s}_2(\mathbf{w}^{(k)})$ denotes a sub-gradient of $f_2(\mathbf{w})$ at $\mathbf{w} = \mathbf{w}^{(k)}$. Obviously, the objective function in problem (8) is convex. The MS algorithm involves solving a sequence of convex optimization problems as in problem (8). In general, there is no closed-form solution to problem (8) and the computational cost of the MS algorithm is k times that of solving problem (8), where k is the number of outer iterations. This is computationally expensive especially for large scale problems.

A class of related algorithms called iterative shrinkage and thresholding (IST), which are also known as different names such as fixed point iteration and forward-backward splitting (Daubechies et al., 2004; Combettes & Wajs, 2005; Hale et al., 2007; Beck & Teboulle, 2009; Wright et al., 2009; Liu et al., 2009), have been extensively applied to solve problem (1). The key step is by generating a sequence $\{\mathbf{w}^{(k)}\}$ via solving problem (2). However, they require that the regularizer $r(\mathbf{w})$ is *convex* and some of them even require that both $l(\mathbf{w})$ and $r(\mathbf{w})$ are convex. Our proposed GIST algorithm is a more general framework, which can deal with a wider range of problems including both convex and non-convex cases.

Another related algorithm called a Variant of Iterative Reweighted L_a (VIRL) is recently proposed to solve the following optimization problem (Lu, 2012a):

$$\min_{\mathbf{w}\in\mathbb{R}^d}\left\{f(\mathbf{w})=l(\mathbf{w})+\lambda\sum_{i=1}^{d}(|w_i|^{\alpha}+\varepsilon_i)^{q/\alpha}\right\},$$

where $a \geq 1$, $0 < q < 1$, $\varepsilon_i > 0$. VIRL solves the above problem by generating a sequence $\{\mathbf{w}^{(k)}\}$ as

$$\mathbf{w}^{(k+1)}=\arg\min_{\mathbf{w}\in\mathbb{R}^d} l(\mathbf{w}^{(k)})+(\nabla l(\mathbf{w}^{(k)}),\mathbf{w}-\mathbf{w}^{(k)})+\frac{t^{(k)}}{2}\|\mathbf{w}-\mathbf{w}^{(k)}\|^2 + \frac{\lambda q}{\alpha}\sum_{i=1}^{d}(|w_i^k|^{\alpha}+\varepsilon_i)^{q/\alpha-1}|w_i|^{\alpha}.$$

In VIRL, $t^{(k-1)}$ is chosen as the initialization of $t^{(k)}$. The line search step in VIRL finds the smallest integer l with $t^{(k)}=t^{(k-1)}\eta^l$ ($\eta>1$) such that

$$f(\mathbf{w}^{(k+1)})\leq f(\mathbf{w}^{(k)})-\frac{\sigma}{2}\|\mathbf{w}^{(k+1)}-\mathbf{w}^{(k)}\|^2(\sigma>0).$$

The most related algorithm to our propose GIST is the Sequential Convex Programming (SCP) proposed by Lu (2012b). SCP solves problem (1) by generating a sequence $\{\mathbf{w}^{(k)}\}$ as

$$\mathbf{w}^{(k+1)}=\arg\min_{\mathbf{w}\in\mathbb{R}^d} l(\mathbf{w}^{(k)})+\langle\nabla l(\mathbf{w}^k),\mathbf{w}-\mathbf{w}^{(k)}\rangle+\frac{t^{(k)}}{2}\|\mathbf{w}-\mathbf{w}^{(k)}\|^2+r_1(\mathbf{w})-r_2(\mathbf{w}^{(k)})-\langle\mathbf{s}_2,\mathbf{w}-\mathbf{w}^{(k)}\rangle,$$

where $\mathbf{s}_2$ is a sub-gradient of $r_2(\mathbf{w})$ at $\mathbf{w}=\mathbf{w}^{(k)}$. Our algorithm differs from SCP in that the original regularizer $r(\mathbf{w})=r_1(\mathbf{w})-r_2(\mathbf{w})$ is used in the proximal operator in problem (2), while $r_1(\mathbf{w})$ minus a locally linear approximation for $r_2(\mathbf{w})$ is adopted in SCP. We will show in the experiments that our proposed GIST algorithm is more efficient than SCP.

4. Experiments

4.1. Experimental Setup

We evaluate our GIST algorithm by considering the Capped l_1 regularized logistic regression problem, that is $l(\mathbf{w})=\frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i\mathbf{x}_i^T\mathbf{w}))$ and $r(\mathbf{w})=\lambda\sum_{i=1}^{d}\min(|w_i|,\theta)$. We compare our GIST algorithm with the Multi-Stage (MS) algorithm and the SCP algorithm in different settings using twelve data sets summarized in Table 2. These data sets are high dimensional and sparse. Two of them (news20, real-sim)[1] have been preprocessed as two-class data sets (Lin et al., 2008). The other ten[2] are multi-class data sets. We transform the

[1] http://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/
[2] http://www.shi-zhong.com/software/docdata.zip

multi-class data sets into two-class by labeling the first half of all classes as positive class, and the remaining classes as the negative class.

All algorithms are implemented in Matlab and executed on an Intel(R) Core(TM)2 Quad CPU (Q6600 @2.4GHz) with 8GB memory. We set $\sigma = 10^{-5}$, $m = 5$, $\eta = 2$, $1/t_{\min} = t_{\max} = 10^{30}$ and choose the starting points $\mathbf{w}^{(0)}$ of all algorithms as zero vectors. We terminate all algorithms if the relative change of the two consecutive objective function values is less than 10^{-5} or the number of iterations exceeds 1000. The Matlab codes of the GIST algorithm are available online (Gong et al., 2013).

4.2. Experimental Evaluation and Analysis

We report the objective function value vs. CPU time plots with different parameter settings in Figure 1. From these figures, we have the following observations: (1) Both GISTbb-Monotone and GISTbb-Nonmonotone decrease the objective function value rapidly and they always have the fastest convergence speed, which shows that adopting the BB rule to initialize $t^{(k)}$ indeed greatly accelerates the convergence speed. Moreover, both GISTbb-Monotone and GISTbb-Nonmonotone algorithms achieve the smallest objective function values. (2) GISTbb-Nonmonotone may give rise to an increasing objective function value but finally converges and has a faster overall convergence speed than GISTbb-Monotone in most cases, which indicates that the non-monotone line search criterion can further accelerate the convergence speed. (3) SCPbb-Nonmonotone is comparable to GISTbb-Nonmonotone in several cases, however, it converges much slower and achieves much larger objective function values than those of GISTbb-Nonmonotone in the remaining cases. This demonstrates the superiority of using the original regularizer $r(\mathbf{w}) = r_1(\mathbf{w}) - r_2(\mathbf{w})$ in the proximal operator in problem (2). (4) GIST-1 has a faster convergence speed than GIST-$t^{(k-1)}$ in most cases, which demonstrates that it is a bad strategy to use $t^{(k-1)}$ to initialize $t^{(k)}$. This is because $\{t^{(k)}\}$ increases monotonically in this way, making the step size $1/t^{(k)}$ monotonically decreasing when the algorithm proceeds.

5. Conclusions

We propose an efficient iterative shrinkage and thresholding algorithm to solve a general class of non-convex optimization problems encountered in sparse learning. A critical step of the proposed algorithm is the computation of a proximal operator, which has a closed-form solution for many commonly used formulations. We propose to initialize the step size at each iteration using the BB rule and employ both monotone and non-monotone criteria as line search conditions, which greatly accelerate the convergence speed. Moreover, we provide a detailed convergence analysis of the proposed algorithm, showing that the algorithm converges under both monotone and non-monotone line search criteria. Experiments results on large-scale data sets demonstrate the fast convergence of the proposed algorithm.

In our future work, we will focus on analyzing the theoretical performance (e.g., prediction error bound, parameter estimation error bound etc.) of the solution obtained by the GIST algorithm. In addition, we plan to apply the proposed algorithm to solve the multitask feature learning problem (Gong et al., 2012a;b).

Acknowledgments

This work is supported partly by 973 Program (2013CB329503), NSFC (Grant No. 91120301, 61075004, 61021063), NIH (R01 LM010730) and NSF (IIS-0953662, CCF-1025177, DMS1208952).

References

Barzilai J, Borwein JM. Two-point step size gradient methods. IMA Journal of Numerical Analysis. 8(1):141–148.1988;

Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences. 2(1):183–202.2009;

Candes EJ, Wakin MB, Boyd SP. Enhancing sparsity by reweighted l_1 minimization. Journal of Fourier Analysis and Applications. 14(5):877–905.2008;

Combettes PL, Wajs VR. Signal recovery by proximal forward-backward splitting. Multiscale Modeling & Simulation. 4(4):1168–1200.2005;

Daubechies I, Defrise M, De Mol C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Communications on pure and applied mathematics. 57(11):1413–1457.2004;

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. The Annals of statistics. 32(2):407–499.2004;

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 96(456):1348–1360.2001;

Foucart S, Lai MJ. Sparsest solutions of underdetermined linear systems via l_q-minimization for $0 < q \leq 1$. Applied and Computational Harmonic Analysis. 26(3):395–407.2009;

Gasso G, Rakotomamonjy A, Canu S. Recovering sparse signals with a certain family of nonconvex penalties and dc programming. IEEE Transactions on Signal Processing. 57(12):4686–4698.2009;

Geman D, Yang C. Nonlinear image recovery with half-quadratic regularization. IEEE Transactions on Image Processing. 4(7):932–946.1995; [PubMed: 18290044]

Gong P, Ye J, Zhang C. Multi-stage multi-task feature learning. NIPS. :1997–2005. 2012a.

Gong P, Ye J, Zhang C. Robust multi-task feature learning. SIGKDD. :895–903.2012b

Gong, P, Zhang, C, Lu, Z, Huang, J, Ye, J. GIST: General Iterative Shrinkage and Thresholding for Non-convex Sparse Learning. Tsinghua University; 2013. URL http://www.public.asu.edu/~jye02/Software/GIST

Grippo L, Sciandrone M. Nonmonotone globalization techniques for the barzilai-borwein gradient method. Computational Optimization and Applications. 23(2):143–169.2002;

Grippo L, Lampariello F, Lucidi S. A nonmonotone line search technique for newton's method. SIAM Journal on Numerical Analysis. 23(4):707–716.1986;

Hale, ET, Yin, W, Zhang, Y. CAAM TR07-07. Rice University; 2007. A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing.

Hunter DR, Lange K. Quantile regression via an mm algorithm. Journal of Computational and Graphical Statistics. 9(1):60–77.2000;

Lin CJ, Weng RC, Keerthi SS. Trust region newton method for logistic regression. Journal of Machine Learning Research. 9:627–650.2008;

Liu, J, Ji, S, Ye, J. SLEP: Sparse Learning with Efficient Projections. Arizona State University; 2009. URL http://www.public.asu.edu/~jye02/Software/SLEP

Lu Z. Iterative reweighted minimization methods for l_p regularized unconstrained nonlinear programming. 2012a

Lu Z. Sequential convex programming methods for a class of structured nonlinear programming. 2012b

Shevade SK, Keerthi SS. A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics. 19(17):2246–2253.2003; [PubMed: 14630653]

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 58(1):267–288.1996;

Toland JF. A duality principle for non-convex optimisation and the calculus of variations. Archive for Rational Mechanics and Analysis. 71(1):41–61.1979;

Trzasko J, Manduca A. Relaxed conditions for sparse signal recovery with general concave priors. IEEE Transactions on Signal Processing. 57(11):4347–4354.2009;

Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 31(2):210–227.2008;

Wright SJ, Nowak R, Figueiredo M. Sparse reconstruction by separable approximation. IEEE Transactions on Signal Processing. 57(7):2479–2493.2009;

Ye J, Liu J. Sparse methods for biomedical data. ACM SIGKDD Explorations Newsletter. 14(1):4–15.2012;

Yuille AL, Rangarajan A. The concave-convex procedure. Neural Computation. 15(4):915–936.2003; [PubMed: 12689392]

Zhang CH. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics. 38(2):894–942.2010a;

Zhang T. Analysis of multi-stage convex relaxation for sparse regularization. Journal of Machine Learning Research. 11:1081–1107.2010b;

Zhang T. Multi-stage convex relaxation for feature selection. Bernoulli. 2012

Appendix: Solutions to Problem (2)

Observe that $r(\mathbf{w}) = \sum_{i=1}^{d} r_i(w_i)$ and problem (2) can be equivalently decomposed into d independent univariate optimization problems:

$$w_i^{(k+1)} = \arg\min_{w_i} h_i(w_i) = \frac{1}{2}\left(w_i - u_i^{(k)}\right)^2 + \frac{1}{t^{(k)}} r_i(w_i),$$

where $i = 1, \cdots, d$ and $u_i^{(k)}$ is the i-th entry of $\mathbf{u}^{(k)} = \mathbf{w}^{(k)} - \nabla l(\mathbf{w}^{(k)})/t^{(k)}$. To simplify the notations, we unclutter the above equation by removing the subscripts and supscripts as follows:

$$w^{(k+1)} = \arg\min_{w} h_i(w) = \frac{1}{2}(w-u)^2 + \frac{1}{t} r_i(w)\,. \tag{9}$$

- **l_1-norm**: $w^{(k+1)} = \text{sign}(u)\max(0, |u| - \lambda/t)$.
- **LSP**: We can obtain an optimal solution of problem (9) via: $w^{(k+1)} = \text{sign}(u)x$, where x is an optimal solution of the following problem:

$$x = \arg\min_{w} \frac{1}{2}(w - |u|)^2 + \frac{\lambda}{t}\log(1 + w/\theta) \quad s.t.\ w \geq 0.$$

Noting that the objective function above is differentiable in the interval $[0, +\infty)$ and the minimum of the above problem is either a stationary point (the first derivative is zero) or an endpoint of the feasible region, we have

$$x = \arg\min_{w \in \mathcal{C}} \frac{1}{2}(w - |u|)^2 + \frac{\lambda}{t}\log(1 + w/\theta),$$

where $\mathcal{C}$ is a set composed of 3 elements or 1 element. If $t^2(|u| - \theta)^2 - 4t(\lambda - t|u|\theta) \geq 0$,

$$\mathcal{C} = \left\{0, \left[\frac{t(|u| - 0) + \sqrt{t^2(|u| - \theta)^2 - 4t(\lambda - t|u|\theta)}}{2t}\right]_+ \left[\frac{t(|u| - \theta) - \sqrt{t^2(|u| - \theta)^2 - 4t(\lambda - t|u|\theta)}}{2t}\right]_+\right\}. \text{ Otherwise, } \mathcal{C} = \{0\}.$$

Otherwise, $\mathcal{C} = \{0\}$.

- **SCAD**: We can recast problem (9) into the following three problems:

$$x_1 = \arg\min_w \frac{1}{2}(w-u)^2 + \frac{\lambda}{t}|w| \quad s.t. \ |w| \leq \lambda,$$

$$x_2 = \arg\min_w \frac{1}{2}(w-u)^2 + \frac{-w^2 + 2\theta(\lambda/t)|w| - (\lambda/t)^2}{2(\theta-1)} s.t. \ \lambda \leq |w| \leq \theta\lambda,$$

$$x_3 = \arg\min_w \frac{1}{2}(w-u)^2 + \frac{(\theta+1)\lambda^2}{2t^2} s.t. \ |w| \geq \theta\lambda.$$

We can easily obtain that (x_2 is obtained using the similar idea as **LSP** by considering that $\theta > 2$):

$$x_1 = \text{sign}(u)\min(\lambda, \max(0, |u| - \lambda/t)),$$

$$x_2 = \text{sign}(u)\min(\theta\lambda, \max(\lambda, \frac{t|u|(\theta-1) - \theta\lambda}{t(\theta-2)})),$$

$$x_3 = \text{sign}(u)\max(\theta\lambda, |u|).$$

Thus, we have

$$w^{(k+1)} = \arg\min_y h_i(y) \quad s.t. \ y \in \{x_1, x_2, x_3\}.$$

- **MCP**: Similar to SCAD, we can recast problem (9) into the following two problems:

$$x_1 = \arg\min_w \frac{1}{2}(w-u)^2 + \frac{\lambda}{t}|w| - \frac{w^2}{2\theta} s.t. \ |w| \leq \theta\lambda,$$

$$x_2 = \arg\min_w \frac{1}{2}(w-u)^2 + \frac{\theta(\lambda/t)^2}{2} s.t. \ |w| \geq \theta\lambda.$$

We can easily obtain that

$$x_1 = \text{sign}(u)z, \ x_2 = \text{sign}(u)\max(\theta\lambda, |u|),$$

where $z = \arg\min_{w \in \mathscr{C}} \frac{1}{2}(w - |u|)^2 + \frac{\lambda}{t}w - \frac{w^2}{2\theta}$; $\mathscr{C}$, if $\theta - 1 \neq 0$, and $\mathscr{C} = \{0, \theta\lambda\}$

$$= \left\{0, \theta\lambda, \min\left(\theta\lambda, \max\left(0, \frac{\theta(t|u| - \lambda)}{t(\theta - 1)}\right)\right)\right\}$$

otherwise. Thus, we have

$$w^{(k+1)} = \begin{cases} x_1, & \text{if } h_i(x_1) \leq h_i(x_2) \\ x_2, & \text{otherwise.} \end{cases}$$

- **Capped l_1**: We can recast problem (9) into the following two problems:

$$x_1 = \arg\min_{w} \frac{1}{2}(w - u)^2 + \frac{\lambda}{t}\theta \quad s.t. \ |w| \geq \theta,$$

$$x_2 = \arg\min_{w} \frac{1}{2}(w - u)^2 + \frac{\lambda}{t}|w| \quad s.t. \ |w| \leq \theta.$$

We can easily obtain that

$$x_1 = \text{sign}(u) \max(\theta, |u|),$$

$$x_2 = \text{sign}(u) \min(\theta, \max(0, |u| - \lambda/t)).$$

Thus, we have

$$w^{(k+1)} = \begin{cases} x_1, & \text{if } h_i(x_1) \leq h_i(x_2), \\ x_2, & \text{otherwise.} \end{cases}$$

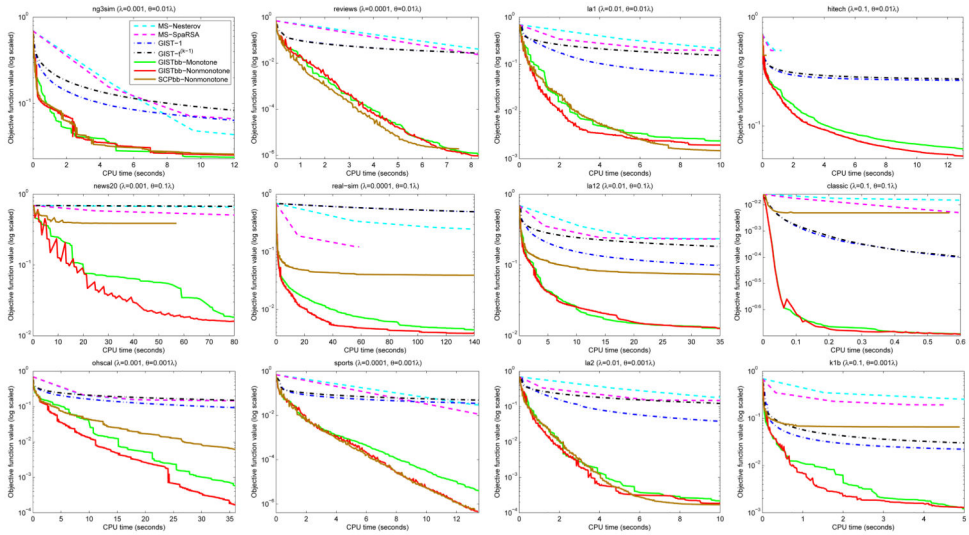

Figure 1.

Objective function value vs. CPU time plots. MS-Nesterov/MS-SpaRSA: The Multi-Stage algorithm using the Nesterov/SpaRSA method to solve problem (8); GIST-1/GIST-$t^{(k-1)}$/GISTbb-Monotone/GISTbb-Nonmonotone: The GIST algorithm using $1/t^{(k-1)}$/BB rule/BB rule to initialize $t^{(k)}$ and Eq. (3)/Eq. (3)/Eq. (3)/Eq. (4) as the line search criterion; SCPbb-Nonmonotone: The SCP algorithm using the BB rule to initialize $t^{(k)}$ and Eq. (4) as the line search criterion. Note that on data sets ‘hitech’ and ‘real-sim’, MS algorithms stop early (the SCP algorithm has similar behaviors on data sets ‘hitech’ and ‘news20’), because they satisfy the termination condition that the relative change of the two consecutive objective function values is less than 10^{-5}. However, their objective function values are much larger than those of GISTbb-Monotone and GISTbb-Nonmonotone.

Table 1

Examples of regularizers (penalties) $r(\mathbf{w})$ satisfying the assumption **A2** and the corresponding convex functions $r_1(\mathbf{w})$ and $r_2(\mathbf{w})$. $\lambda > 0$ is the regularization parameter; $r(\mathbf{w}) = \Sigma_i r_i(w_i)$, $r_1(\mathbf{w}) = \Sigma_i r_{1,i}(w_i)$, $r_2(\mathbf{w}) = \Sigma_i r_{2,i}(w_i)$, $[x]^+ = \max(0, x)$.

Name	$r_i(w_i)$	$r_{1,i}(w_i)$	$r_{2,i}(w_i)$
l_1-norm	$\lambda\|w_i\|$	$\lambda\|w_i\|$	0
LSP	$\lambda \log(1 + \|w_i\|/\theta)$ $(\theta > 0)$	$\lambda\|w_i\|$	$\lambda\,(\|w_i\| - \log(1 + \|w_i\|/\theta))$
SCAD	$\lambda\int_0^{\|w_i\|} \min\left(1, \frac{[\theta\lambda - x]_+}{(\theta-1)\lambda}\right)dx$ $(\theta > 2)$ $= \begin{cases} \lambda\|w_i\|, & \text{if } \|w_i\| \le \lambda, \\ \frac{-w_i^2 + 2\theta\lambda\|w_i\| - \lambda^2}{2(\theta-1)}, & \text{if } \lambda < \|w_i\| \le \theta\lambda, \\ (\theta+1)\lambda^2/2, & \text{if } \|w_i\| > \theta\lambda. \end{cases}$	$\lambda\|w_i\|$	$\lambda\int_0^{\|w_i\|} \frac{[\min(\theta\lambda, x) - \lambda]_+}{(\theta-1)\lambda}dx$ $= \begin{cases} 0, & \text{if } \|w_i\| \le \lambda, \\ \frac{w_i^2 - 2\lambda\|w_i + \lambda^2\|}{2(\theta-1)}, & \text{if } \lambda < \|w_i\| \theta\lambda, \\ \lambda\|w_i\| - \frac{(\theta+1)\lambda^2}{2}, & \text{if } \|w_i\| > \theta\lambda. \end{cases}$
MCP	$\lambda\int_0^{\|w_i\|} [1 - \frac{x}{\theta\lambda}]_+ dx$ $(\theta > 0)$ $= \begin{cases} \lambda\|w_i\| - w_i^2/(2\theta), & \text{if } \|w_i\| \le \theta\lambda, \\ \theta\lambda^2/2, & \text{if } \|w_i\| > \theta\lambda. \end{cases}$	$\lambda\|w_i\|$	$\lambda\int_0^{\|w_i\|} \min(1, x/(\theta\lambda))dx$ $= \begin{cases} w_i^2/(2\theta), & \text{if } \|w_i\| \le \theta\lambda, \\ \lambda\|w_i\| - \theta\lambda^2/2, & \text{if } \|w_i\| > \theta\lambda. \end{cases}$
Capped l_1	$\lambda \min(\|w_i\|, \theta)$ $(\theta > 0)$	$\lambda\|w_i\|$	$\lambda\,[\|w_i\| - \theta]_+$

Table 2

Data sets statistics: n is the number of samples and d is the dimensionality of the data.

No.	1	2	3	4	5	6	7	8	9	10	11	12
datasets	**classic**	**hitech**	**k1b**	**la12**	**la1**	**la2**	**news20**	**ng3sim**	**ohscal**	**real-sim**	**reviews**	**sports**
n	7094	2301	2340	2301	3204	3075	19996	2998	11162	72309	4069	8580
d	41681	10080	21839	31472	31472	31472	1355191	15810	11465	20958	18482	14866

JMLR Workshop Conf Proc. Author manuscript; available in PMC 2014 October 03.

Algorithm 1

GIST: General Iterative Shrinkage and Thresholding Algorithm

Choose parameters $\eta > 1$ and $t_{\min}$, $t_{\max}$ with $0 < t_{\min} < t_{\max}$;

Initialize iteration counter $k \leftarrow 0$ and a bounded starting point $\mathbf{w}^{(0)}$;

repeat

  $t^{(k)} \in [t_{\min}, t_{\max}]$;

  repeat

    $$\mathbf{w}^{(k+1)} \leftarrow \arg\ \min_{\mathbf{w}}\ l(\mathbf{w}^{(k)}) + \langle \nabla l(\mathbf{w}^{(k)}), \mathbf{w} - \mathbf{w}^{(k)} \rangle + \frac{t^{(k)}}{2}\|\mathbf{w} - \mathbf{w}^{(k)}\|^2 + r(\mathbf{w});$$

    $t^{(k)} \leftarrow \eta t^{(k)}$;

  until some line search criterion is satisfied

  $k \leftarrow k + 1$

until some stopping criterion is satisfied