Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

An Alternating Proximal Splitting Method with Global Convergence for Nonconvex Structured Sparsity Optimization

Shubao Zhang[1] and **Hui Qian**[1*] and **Xiaojin Gong**[2]
[1]College of Computer Science and Technology
[2]Department of Information Science and Electronic Engineering
Zhejiang University, Hangzhou 310027, China
{bravemind, qianhui, gongxj}@zju.edu.cn

Abstract

In many learning tasks with structural properties, structured sparse modeling usually leads to better interpretability and higher generalization performance. While great efforts have focused on the convex regularization, recent studies show that nonconvex regularizers can outperform their convex counterparts in many situations. However, the resulting nonconvex optimization problems are still challenging, especially for the structured sparsity-inducing regularizers. In this paper, we propose a splitting method for solving nonconvex structured sparsity optimization problems. The proposed method alternates between a gradient step and an easily solvable proximal step, and thus enjoys low per-iteration computational complexity. We prove that the whole sequence generated by the proposed method converges to a critical point with at least sublinear convergence rate, relying on the Kurdyka-Łojasiewicz inequality. Experiments on both simulated and real-world data sets demonstrate the efficiency and efficacy of the proposed method.

Introduction

Learning or mining from high dimensional data is an important issue (Hastie, Tibshirani, and Friedman 2009). In this paper we are especially concerned with variable selection problems. Generally, the colinearity among the variables implies that the underlying model lies on an intrinsic low dimensional subspace. Thus, it is interesting and challenging to find a sparse representation for high dimensional data.

To achieve sparsity, regularization methods have been widely used in the literature. It admits a tradeoff between the empirical loss and regularization as:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + \tilde{g}(\mathbf{x}), \qquad (1)$$

where f is the loss, and $\tilde{g}$ is a regularizer on the parameter $\mathbf{x}$. A principled approach is learning with the ℓ_1 norm such as lasso (Tibshirani 1996), which encourages sparse estimate of $\mathbf{x}$. When the parameter has some intrinsic structures, it needs to use more sophisticated structured sparsity-inducing regularizers such as the (overlapping) group lasso (Yuan and Lin 2006; Zhao, Rocha, and Yu 2009), the generalized lasso (Tibshirani and Taylor 2011), etc. However, it is usually not easy to solve the structured sparsity optimization problem (1) directly, since the structured sparsity-inducing regularizers $\tilde{g}$ are often composite. For example, when solving the overlapping group lasso by using the proximal gradient algorithm (Yuan, Liu, and Ye 2013), the underlying proximal operator for $\tilde{g}$ is not easily computed, see Eq. (3).

Observe that in many cases, the composite regularization function can be decomposed into a simple regularization function g and a linear transformation $\mathbf{D}$, i.e., $\tilde{g}(\mathbf{x}) = g(\mathbf{D}\mathbf{x})$ where $\mathbf{D} \in \mathbb{R}^{p\times d}$. In this paper, we consider the following general optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + g(\mathbf{D}\mathbf{x}), \qquad (2)$$

where the linear operator $\mathbf{D}$ encodes the structural information of the parameter. The problem (2) can be convex or nonconvex.

Developing efficient algorithm for (2) is a hot area of research focus. Most of existing algorithms are mainly designed for the convex optimization problems with global optimal solution. Among the most successful ones are Nesterov's optimal first-order method (Nesterov 2007; Argyriou et al. 2011) and the alternating direction method of multipliers (ADMM) (Boyd et al. 2011). However, these methods can not be directly extended to the nonconvex case due to lack of convergence guarantee. Existing nonconvex approaches for (1) usually take the idea of concave-convex procedure (CCCP) (Yuille and Rangarajan 2001), majorization-minimization (MM) (Hunter and Li 2005), and concave duality (Zhang 2010b; Zhang et al. 2013), etc. For example, the iterative reweighted ℓ_1 and ℓ_2 methods (Candès, Wakin, and Boyd 2008; Daubechies et al. 2010) employ the idea of majorization-minimization; the DC programming method (Gasso, Rakotomamonjy, and Canu 2009) shares the same spirit with the concave-convex procedure. All these methods have a point in common that they solve the nonconvex optimization problem by solving a sequence of solvable convex problems. As a result, these methods are often computationally expensive for large scale problems. It is thus desirable to develop efficient algorithms that are scalable for large scale problems. In the spirit of the proximal algorithm, a general iterative shrinkage and thresholding (GIST) framework was recently proposed for a class of nonconvex sparisty opti-

[*] Corresponding author.

mization problems (Gong et al. 2013), which enjoys low per-iteration computational complexity. However, this method can not handle with the structured sparse learning problems such as the overlapping group lasso and generalized lasso, since the proximal maps for these regularizers are difficult to solve.

Moreover, to the best of our knowledge, none of the nonconvex methods mentioned above established the global convergence property, i.e., the whole sequence generated by the method converges to a critical point. These schemes only guarantee that the objective function value is monotonically decreasing over the iterations, and thus there exists a subsequence converging to the critical point if the sequence is bounded. Therefore, it is of great significance to develop nonconvex methods with global convergence, not only for theoretical importance, but also for practical computation as many intermediate results are useless for a method without global convergence property.

In this paper, we propose a splitting method for solving (2). Inspired by the great success of the Forward-Backward splitting method (also known as the proximal gradient method) for convex sparsity optimization (Beck and Teboulle 2009; Combettes and Pesquet 2011), we develop an alternating Forward-Backward splitting scheme by combinning the idea of the alternating minimization and proximal algorithm. The proposed method alternates between two minimization subproblems that reduce to a gradient step and an easily solvable proximal step, and thus enjoys low per-iteration computational complexity. Furthermore, we prove that the whole sequence generated by the proposed method converges to a critical point. We also show that the convergence rate of the proposed method is at least sublinear.

It is worth pointing out that when g in (2) is the ℓ_0 norm, the proposed method can solve the ℓ_0 norm based structured sparsity optimization problems. In this paper, we propose the ℓ_0 norm penalized overlapping group lasso and the ℓ_0 norm penalized graph-guided fused logistic regression. To the best of our knowledge, our work is the first study in this issue. Moreover, with different choices of the loss f and the linear operator $\mathbf{D}$, this also can induce other kind of structured sparse modelings based on the ℓ_0 norm.

Problem Formulation

The following assumptions are made on the problem (2) throughout the paper:

(A1) $\mathbf{D}\mathbf{D}^T \succeq \mu_0 \mathbf{I}$.

(A2) f is a continuously differentiable function with Lipschitz continuous gradient, i.e.,

$$||\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})|| \leq L_f ||\mathbf{x} - \mathbf{y}||, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

g is a proper lower semi-continuous function, and the associated proximal operator

$$\mathbf{prox}_g^\lambda(\mathbf{u}) = \underset{\mathbf{z}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{z} - \mathbf{u}||_2^2 + \lambda g(\mathbf{z}) \qquad (3)$$

can be solved efficiently and exactly [*].

(A3) $f(\mathbf{x}) + g(\mathbf{D}\mathbf{x}) \to \infty$ iff $||\mathbf{x}|| \to \infty$.

Based on different choices of $f, g, \mathbf{D}$, the problem (2) covers many applications in machine learning, statistical estimation, signal processing, and computer vision literatures. For the choice of f, the least square and logistic loss functions are two most widely used ones satisfying (**A2**):

$$f(\mathbf{x}) = \frac{1}{2}||\mathbf{A}\mathbf{x} - \mathbf{y}||_2^2, \text{or} \sum_{i=1}^n \log(1 + \exp(-y_i \mathbf{a}_i^T \mathbf{x})),$$

where $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_n]^T$ is the data matrix and $\mathbf{y} = [y_1, \dots, y_n]$ is the response vector. It is easy to show that the Lipschitz constants in (**A2**) of the least square and logistic loss are lower bounded by $||\mathbf{A}^T\mathbf{A}||_2$ and $\sum_{i=1}^n ||\mathbf{a}_i||_2^2$ respectively.

For the choice of g, the most natural one is the ℓ_0 norm. But it is difficult to solve the ℓ_0 norm based optimization problems (NP-hard). A popular choice is the ℓ_1 norm, which is the tightest convex relaxation of the ℓ_0 norm. However, the convex models based on the ℓ_1 norm have been shown to be suboptimal in many cases (Candès, Wakin, and Boyd 2008; Zhang 2010b). Indeed, the ℓ_1 norm often leads to over-penalization, since it is a loose approximation of the ℓ_0 norm. To overcome this issue, many nonconvex regularizers have been proposed, including the smoothly clipped absolute deviation penalty (SCAD) (Fan and Li 2001), the log-sum penalty (LSP) (Candès, Wakin, and Boyd 2008), the minimax concave plus penalty (MCP) (Zhang 2010a), the capped-ℓ_1 penalty (Zhang 2010b), etc. These penalties are demonstrated to have attractive theoretical properties and wide practical applications, and can outperform their convex counterparts. Here we give two important structured sparsity examples below.

Example 1 (**Overlapping Group Lasso**) The group lasso is a way to select groups of features. Given some an a priori groups that may overlap, we can construct a binary matrix $\mathbf{D}$ for group configuration. By different setting of $\mathbf{D}$, the model (2) can handle non-overlapping group lasso and overlapping group lasso. When the size of $\mathbf{D}$ is large, it can be efficiently implemented by a sparse matrix. This kind of indexing matrix was previously used in (Argyriou et al. 2011). Let $\mathbf{z} = \mathbf{D}\mathbf{x}$. We can get a new groups partition $\mathcal{G} := \{\mathrm{g}_k\}$ of $\mathbf{z}$ that do not overlap. Now we can choose the $\ell_{2,1}$ norm as regularizer, i.e., $g(\mathbf{z}) = \sum_{k=1}^K ||\mathbf{z}_{\mathrm{g}_k}||_2$. However, the convex approach does not possess the group level selection consistency. To overcome this drawback, we can choose g from other nonconvex regularizers such as LSP, MCP, SCAD, and capped-ℓ_1 mixed with the ℓ_2 norm; please also refer to (Wang, Chen, and Li 2007; Huang, Wei, and Ma 2012). Very recently, Xiang, Shen, and Ye (2015) proposed the ℓ_0 norm constrained non-overlapping group lasso. In this paper, we propose the ℓ_0 norm penalized (overlapping) group lasso by setting $g(\mathbf{z}) = \sum_{k=1}^K \mathbf{I}(||\mathbf{z}_{\mathrm{g}_k}||_2 \neq 0)$. $\mathbf{I}(\cdot)$ denotes the indicator function. In this example f is the least square loss.

Now we show how to solve the proximal map (3) with mixed regularizers. We take $g(\mathbf{z}) = r(h(\mathbf{z}))$ where $h(\mathbf{z}) = [||\mathbf{z}_{\mathrm{g}_1}||_2; \dots; ||\mathbf{z}_{\mathrm{g}_K}||_2]$, and r is the penalty function such as

[*] It was pointed out in (Bolte, Sabach, and Teboulle 2014) that for proper lower semicontinuous functions, the proximal maps are well defined: the set $\mathbf{prox}_g^\lambda$ is nonempty and compact.

Table 1: Examples of one-dimensional regularizers and their corresponding proximal maps.

Regularizer	$\zeta(w_i)$	$\mathbf{prox}_{\zeta}^{\lambda}(s_i)$
ℓ_1-norm	$\|w_i\|$	$\mathrm{sign}(s_i)\max(\|s_i\|-\lambda, 0)$
ℓ_0-norm	$\mathbf{I}(\|w_i\| \neq 0)$	$\begin{cases} s_i & \text{if } \|s_i\| > \sqrt{2\lambda} \\ \{0, s_i\} & \text{if } \|s_i\| = \sqrt{2\lambda} \\ 0 & \text{if } \|s_i\| < \sqrt{2\lambda} \end{cases}$
capped-ℓ_1	$\min(\|w_i\|, \theta), (\theta > 0)$	$\begin{cases} \pi_1 = \mathrm{sign}(s_i)\max(\theta, \|s_i\|) \text{ when } \|w_i\| \geq \theta & \text{if } l_i(\pi_1) \leq l_i(\pi_2) \\ \pi_2 = \mathrm{sign}(s_i)\min(\theta, \max(0, \|s_i\|-\lambda)) \text{ when } \|w_i\| < \theta & \text{otherwise} \end{cases}$

the ℓ_1 or ℓ_0 norm, LSP, SCAD, MCP and capped-ℓ_1. First we assume that $||\mathbf{z}_{\mathrm{g}_k}||_2$ is known. Then $g(\mathbf{z})$ is also fixed. Let $\mathbf{z}^*$ denote the optimal solution of (3). The optimality condition of (3) implies that $\mathbf{z}^*_{\mathrm{g}_k}$ and $\mathbf{u}_{\mathrm{g}_k}$ are in the same direction. Thus the optimal solution of (3) can be obtained as

$$z_j^* = \begin{cases} u_j & j \notin \mathrm{g}_k, \\ \frac{u_j ||\mathbf{z}^*_{\mathrm{g}_k}||_2}{||\mathbf{u}_{\mathrm{g}_k}||_2} & j \in \mathrm{g}_k. \end{cases} \tag{4}$$

When $||\mathbf{u}_{\mathrm{g}_k}||_2 = 0$, we have $\mathbf{z}^*_{\mathrm{g}_k} = \mathbf{u}_{\mathrm{g}_k}$. From (4), we have $||\mathbf{z}^* - \mathbf{u}||_2^2 = (||\mathbf{z}^*_{\mathrm{g}_k}||_2 - ||\mathbf{u}_{\mathrm{g}_k}||_2)^2$. Let $w_k = ||\mathbf{z}^*_{\mathrm{g}_k}||_2, s_k = ||\mathbf{u}_{\mathrm{g}_k}||_2$. Then the proximal map (3) becomes the following problem

$$\min_{\mathbf{w}} \{ l(\mathbf{w}) := \frac{1}{2}||\mathbf{w} - \mathbf{s}||_2^2 + \lambda r(\mathbf{w}) \}, \tag{5}$$

where $r(\mathbf{w}) = \sum_{i=1}^{K} \zeta(w_i)$ and $l(\mathbf{w}) = \sum_{i=1}^{K} l_i(w_i)$. Table 1 lists the proximal solution of the ℓ_0, ℓ_1 norm and the capped-ℓ_1 penalty function. The proximal map for other nonconvex regularizers such as LSP, SCAD, MCP can be found in (Gong et al. 2013).

Example 2 (**Graph-guided Fused Logistic Regression**) The graph-guided fused logistic regression exploits some graph structure to select relevant feature variables jointly and improve classification performance. Assume that an a priori graph $\mathcal{G} := \{V, E\}$ with nodes V and edges E is given, where each node on the graph corresponds to a feature variable. We can construct an incidence matrix $\mathbf{D}$ whose each row with two nonzero elements $1, -1$ corresponds to an edge on the graph. The regularizer g can be chosen as the convex ℓ_1 norm or the nonconvex LSP, MCP, SCAD, and capped-ℓ_1. In this paper, we propose the ℓ_0 norm penalized graph-guided fused logistic regression, which is not discussed before. In this example f is the logistic loss.

Proposed Method

Due to nonsmoothness and nonseparability of the regularizer, it is a challenge to solve (2) directly. We deal with the regularizer by using the variable splitting and penalty techniques.

By letting $\mathbf{z} = \mathbf{D}\mathbf{x}$, the problem (2) can be rewritten as

$$\min_{\mathbf{x},\mathbf{z}} f(\mathbf{x}) + g(\mathbf{z}) \quad \text{s.t.} \quad \mathbf{z} = \mathbf{D}\mathbf{x}. \tag{6}$$

Using the penalty method, we obtain an approximation of the above problem as

$$\min_{\mathbf{x},\mathbf{z}} \{ F(\mathbf{x}, \mathbf{z}) = f(\mathbf{x}) + g(\mathbf{z}) + \frac{\rho}{2}||\mathbf{z} - \mathbf{D}\mathbf{x}||_2^2 \}, \tag{7}$$

where ρ is either set as a large penalty value or takes a sequence of increasing values using the continuation scheme. It is clear that when ρ tends to infinity, the solution of (7) converges to (6) (Luenberger and Ye 2008).

We develop an alternating minimization algorithm solving (7), which consists of two steps. The first step calculates $\mathbf{x}$, with $\mathbf{z}$ fixed, via

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) + \frac{\rho}{2}||\mathbf{D}\mathbf{x} - \mathbf{z}_{k-1}||_2^2. \tag{8}$$

When f is the least square loss, $\mathbf{x}_k$ is the solution of the following linear equation system:

$$(\mathbf{A}^T\mathbf{A} + \rho\mathbf{D}^T\mathbf{D})\mathbf{x} = \mathbf{A}^T\mathbf{y} + \rho\mathbf{D}^T\mathbf{z}_{k-1}. \tag{9}$$

However, when f is other type of loss functions such as the logistic loss, it involves a nonlinear optimization. Also, when the dimensionality is high, solving (9) may be computational heavy. To address this issue, we resort to the linearization technique and solve the following surrogate optimization problem:

$$\begin{aligned} \mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^d} \langle \nabla f(\mathbf{x}_{k-1}), \mathbf{x} \rangle + \langle \rho\mathbf{D}^T(\mathbf{D}\mathbf{x}_{k-1} - \mathbf{z}_{k-1}), \mathbf{x} \rangle \\ + \frac{\eta}{2}||\mathbf{x} - \mathbf{x}_{k-1}||_2^2, \end{aligned} \tag{10}$$

where $\eta \geq (L_f + \rho||\mathbf{D}^T\mathbf{D}||_2)/2$ suggested by the analysis next. Particularly, when the Lipschitz constant is not known or computable for large scale problems, one may use the line search method to estimate η (Beck and Teboulle 2009). The above procedure reduces to a gradient descent step:

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \frac{1}{\eta}[\nabla f(\mathbf{x}_{k-1}) + \rho\mathbf{D}^T(\mathbf{D}\mathbf{x}_{k-1} - \mathbf{z}_{k-1})], \tag{11}$$

where $1/\eta$ plays the role of step size. The second step calculates $\mathbf{z}$, with $\mathbf{x}$ fixed, via

$$\mathbf{z}_k = \operatorname*{argmin}_{\mathbf{z} \in \mathbb{R}^p} g(\mathbf{z}) + \frac{\rho}{2}||\mathbf{z} - \mathbf{D}\mathbf{x}_k||_2^2, \tag{12}$$

which has closed form solution for typical regularizers.

Note that the above alternating scheme with update rules (11) and (12) consists of a gradient step and a proximal step, which bears a resemblance to the Forward-Backward splitting (FBS) algorithm. However, when dealing with the structured sparisty optimization problem (2), the traditional FBS method can not attain exact minimization of the proximal step. While our method can obtain an exact solution by introducing an auxilliary variable. Hence we call our method as the alternating Forward-Backward splitting (AFBS) algorithm summarized in Algorithm 1. Indeed, the AFBS algorithm is the coupling of the alternating minimization (also block coordinate descent) method with the FBS method. Inspired by FISTA (Beck and Teboulle 2009), we devise an accelerated variant of the AFBS algorithm summarized in Algorithm 2. The later experiments show that Algorithm 2 is more effective than Algorithm 1. The continuation scheme often can further accelerate the algorithm. So we update ρ by taking a sequence of increasing values, which is summarized in Algorithm 3.

Algorithm 1 The AFBS Algorithm

```
Input: $(\mathbf{x}_0, \mathbf{z}_0)$, $\mathbf{D}$, $\rho > 0$, $\eta \ge (L_f + \rho||\mathbf{D}^T\mathbf{D}||_2)/2$.
for $k = 1, 2, \dots$ do
    $\mathbf{x}_k = \mathbf{x}_{k-1} - \frac{1}{\eta}[\nabla f(\mathbf{x}_{k-1}) + \rho \mathbf{D}^T(\mathbf{D}\mathbf{x}_{k-1} - \mathbf{z}_{k-1})]$.
    $\mathbf{z}_k = \mathbf{prox}_g^{\rho}(\mathbf{D}\mathbf{x}_k)$.
end for
Output: $(\mathbf{x}, \mathbf{z})$.
```

Algorithm 2 The Accelerated AFBS Algorithm

```
Input: $(\mathbf{u}_1, \mathbf{v}_1) = (\mathbf{x}_0, \mathbf{z}_0)$, $\mathbf{D}$, $\rho > 0$, $\alpha_1 = 1$,
       $\eta \ge L_f + \rho(||\mathbf{D}^T\mathbf{D}||_2 + 1)$.
for $k = 1, 2, \dots$ do
    $\mathbf{x}_k = \mathbf{u}_k - \frac{1}{\eta}[\nabla f(\mathbf{u}_k) + \rho \mathbf{D}^T(\mathbf{D}\mathbf{u}_k - \mathbf{v}_k)]$.
    $\mathbf{z}_k = \mathbf{prox}_g^{\rho}(\mathbf{D}\mathbf{x}_k)$.
    $\alpha_{k+1} = \frac{1+\sqrt{1+4\alpha_k^2}}{2}$.
    $\mathbf{u}_{k+1} = \mathbf{x}_k + \frac{\alpha_k - 1}{\alpha_{k+1}}(\mathbf{x}_k - \mathbf{x}_{k-1})$.
    $\mathbf{v}_{k+1} = \mathbf{z}_k + \frac{\alpha_k - 1}{\alpha_{k+1}}(\mathbf{z}_k - \mathbf{z}_{k-1})$.
end for
Output: $(\mathbf{x}, \mathbf{z})$.
```

Algorithm 3 (Accelerated) AFBS with Continuation

```
Input: $\mathbf{D}$, $\rho > 0$, $\rho_{\max} > 0$, $(\mathbf{x}_0, \mathbf{z}_0)$.
while $\rho < \rho_{\max}$ do
    $(\mathbf{x}_m, \mathbf{z}_m) = \mathrm{argmin}_{\mathbf{x},\mathbf{z}} F(\mathbf{x}, \mathbf{z})$.
    $\rho = \alpha \cdot \rho, \ \ \alpha > 1$.
end while
Output: $(\mathbf{x}, \mathbf{z})$.
```

Remark The ADMM algorithm is another popular approach for solving the problem (2). Particularly, like AFBS, the linearized ADMM variant also consists of a gradient step and a proximal step for the primal variable updates (Ouyang et al. 2015). The key difference between AFBS and ADMM lies in that: AFBS is a primal method, while ADMM is a primal-dual method solving a saddle-point problem. In addition, the current linearized ADMM only consider the convex case. Our convergence analysis can be extended to establish similar convergence property for the nonconvex linearized ADMM.

Convergence Analysis

The convergence of the alternating minimization algorithm with update rules (8) and (12) is guaranteed by the convergence result for the Gauss-Seidel method, assuming that the minimum in each step is uniquely attained (Zangwill 1969). The rest of this section consider the convergence performance of the AFBS algorithm. The following theorem establishes the convergence property in terms of limit points.

Theorem 1 *Suppose that the assumptions **A1-3** hold. Let $\{(\mathbf{x}_k, \mathbf{z}_k)\}_{k\in\mathbb{N}}$ be the sequence generated by the AFBS algorithm. Then the following assertions hold.*

(i) The sequence $\{F(\mathbf{x}_k, \mathbf{z}_k)\}_{k\in\mathbb{N}}$ is nonincreasing and in particular

$$F(\mathbf{x}_k, \mathbf{z}_k) - F(\mathbf{x}_{k+1}, \mathbf{z}_{k+1}) \ge C_1||\mathbf{x}_{k+1} - \mathbf{x}_k||_2^2,$$

where $C_1 = \eta - (L_f + \rho||\mathbf{D}^T\mathbf{D}||_2)/2$.

(ii) $\sum_{k=0}^{\infty} ||\mathbf{x}_{k+1} - \mathbf{x}_k||_2^2 + ||\mathbf{z}_{k+1} - \mathbf{z}_k||_2^2 \le \infty$. In particular, $\lim_{k\to\infty} ||\mathbf{x}_{k+1} - \mathbf{x}_k||_2 + ||\mathbf{z}_{k+1} - \mathbf{z}_k||_2 = 0$.

(iii) Any limit point of $\{(\mathbf{x}_k, \mathbf{z}_k)\}_{k\in\mathbb{N}}$ is a critical point of F in (7).

Now we present the global convergence property of the AFBS algorithm. It should be noticed that the global convergence means that the sequence $\{(\mathbf{x}_k, \mathbf{z}_k)\}_{k\in\mathbb{N}}$ converges to a critical point of F in (7). Our global convergence analysis is an extension of (Attouch and Bolte 2009), which relies on the Kurdyka-Łojasiewicz inequality (refer to the supplemental material for details).

Theorem 2 *Suppose that the assumptions **A1-3** hold, and furthermore that the objective function F in (7) satisfies the Kurdyka-Łojasiewicz property. Then the following assertions hold.*

(i) The sequence $\{(\mathbf{x}_k, \mathbf{z}_k)\}_{k\in\mathbb{N}}$ has finite length. That is,

$$\sum_{k=0}^{\infty} ||\mathbf{x}_{k+1} - \mathbf{x}_k||_2 + ||\mathbf{z}_{k+1} - \mathbf{z}_k||_2 < \infty.$$

(ii) The sequence $\{(\mathbf{x}_k, \mathbf{z}_k)\}_{k\in\mathbb{N}}$ converges to a critical point of F.

Theorem 3 *The sequence $\{(\mathbf{x}_k, \mathbf{z}_k)\}_{k\in\mathbb{N}}$ converges to a critical point $(\mathbf{x}_*, \mathbf{z}_*)$ of F in (7) with at least the sublinear convergence rate. Specifically, there exists $C > 0$ such that*

$$||(\mathbf{x}_k, \mathbf{z}_k) - (\mathbf{x}_*, \mathbf{z}_*)||_2 \le C\, k^{-\frac{1-\theta}{2\theta-1}}$$

where $\theta \in (\frac{1}{2}, 1)$.

Remark It is well established that subanalytic functions satisfy the Kurdyka-Łojasiewicz property (Bolte, Daniilidis, and Lewis 2007), which includes real analytic and semi-algebraic functions as typical examples. Moreover, the sum of a real analytic function and a subanalytic function is subanalytic (Bochnak, Coste, and Roy 1998). Thus, it admits the Kurdyka-Łojasiewicz property. The functions involved in this paper are all subanalytic. For example, the least square and logistic loss are real analytic; the ℓ_0 or ℓ_1 norm, capped-ℓ_1, MCP and SCAD are all semi-algebraic. Please refer to the supplemental material for more details and the proofs of Theorem 1,2,3.

Experiments

In this section, we demonstrate the efficiency and efficacy of the proposed method on the overlapping group lasso and the graph-guided fused logistic regression. Experiments are performed on a workstation with Intel Xeon E5-2690 × 2 CPU and 128GB memory.

Overlapping Group Lasso

We apply the ℓ_1, ℓ_0 norm, and capped-ℓ_1 penalty function as regularizer to the overlappling group lasso. This leads to the following optimization problems:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}||\mathbf{y}-\mathbf{A}\mathbf{x}||_2^2 + \lambda\sum_{k=1}^{K}||\mathbf{x}_{\mathrm{g}_k}||_2, \tag{13}$$

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}||\mathbf{y}-\mathbf{A}\mathbf{x}||_2^2 + \lambda\sum_{k=1}^{K}\min(||\mathbf{x}_{\mathrm{g}_k}||_2,\theta), \tag{14}$$

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}||\mathbf{y}-\mathbf{A}\mathbf{x}||_2^2 + \lambda\sum_{k=1}^{K}\mathbf{I}(||\mathbf{x}_{\mathrm{g}_k}||_2 \neq 0). \tag{15}$$

We generate the simulated datasets according to $\mathbf{y} = \mathbf{A}\mathbf{x}+\mathbf{w}$, where $\mathbf{w} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$ is the random noise. We set the noise level $\sigma = 1e-3$ for all experiments next. Each element of the design matrix $\mathbf{A}$ is drawn i.i.d. from the normal distribution with normalized columns. The K overlapping groups are defined as:

$$\{1,\ldots,50\},\{41,\ldots,90\},\ldots,\{d-49,\ldots,d\},$$

where $d = 40K + 10$. We generate the ground truth parameter $\mathbf{x}^*$ with each entry sampled i.i.d. from a standard Gaussian distribution. We randomly select $K/2$ predefined groups and set the remaining entries of $\mathbf{x}^*$ to be zeros.

In the first experiment, we compare the AFBS and accelerated AFBS algorithms with several state-of-the-art methods for solving the convex optimization problem (13). The following algorithms will be compared in the experiment:

1. AFBS: The AFBS algorithm with continuation.
2. AFBS_ACC: The accelerated AFBS algorithm with continuation.
3. Picard: A general algorithm based on FISTA (Beck and Teboulle 2009) solves the convex optimization problem (2) (Argyriou et al. 2011). The proximal step is solved by a fixed point method. We use the code from http://ttic.uchicago.edu/ argyriou/code/index.html.

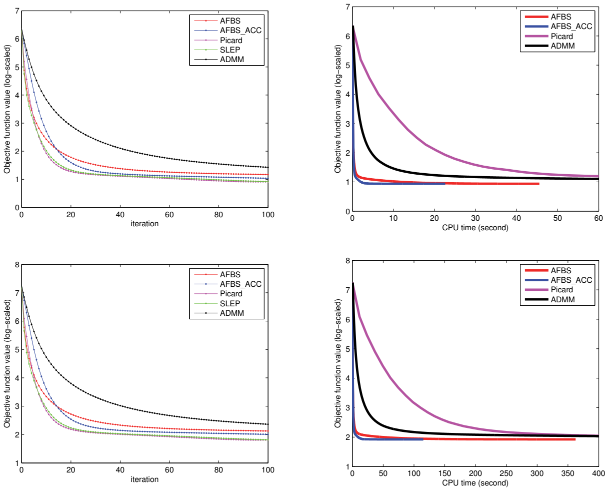

Figure 1: Convergence performance of objective function value with respect to iterations (left column) and running times (right column). First row: $n = 4000, K = 120, d = 4810$. Second row: $n = 10000, K = 300, d = 12010$.

4. SLEP: An algorithm based on FISTA solves the overlapping group lasso problem (13) (Yuan, Liu, and Ye 2013). The proximal step is computed by solving its dual problem. The code can be found in the SLEP package (Liu, Ji, and Ye 2009) (http://www.yelab.net/software/SLEP/).
5. ADMM: The alternating direction method of multipliers for solving the overlappling group lasso problem (13) (Deng, Yin, and Zhang 2011). We use the code from the YALL1 group package (http://yall1.blogs.rice.edu/).

All the algorithms above are implemented in MATLAB, except that the proximal map in SLEP is computed by the C++ code. Thus, we choose the Picard algorithm to be compared with other algorithms when evaluated with respect to the running time. To be fair, all methods start from zero initialization and terminate when the relative change of the objective function is less than $1e-4$. We initiate $\rho = 1$ in AFBS and AFBS_ACC and set the penalty parameter $\lambda = K/100$. Figure 1 illustrates the convergence behavior of these method (notice that just part of all iterations are illustrated in figure 1 left column, and the listed iterations of AFBS and AFBS_ACC belong to the first stage approximation with respect to ρ). Overall, the Picard and SLEP algorithms achieves the fastest convergence rate, since they are based on the optimal proximal gradient method. However, in terms of running time, the AFBS and AFBS_ACC algorithms are the most efficient ones. This is because of that our methods have lower per-iteration computational complexity, while the Picard and SLEP methods needs to solve more difficult proximal maps that lead to higher per-iteration computation computational complexity. Therefore, our methods achieve a better tradeoff between convergence rate and computational complexity and are more scalable for large size problems.

In the second experiment, we compare the AFBS and accelerated AFBS algorithms with a CCCP based algorithm in (Yuan, Liu, and Ye 2013) for solving (14). The CCCP based algorithm solve the nonconvex problem (14) by solving a sequence of convex overlapping group lasso problems (13). Here we use the SLEP algorithm to solve the inner convex problem. The thresholding parameter in the capped-ℓ_1 regularizer is set as $\theta = 0.1$. Figure 2 demonstrates that in terms of running time, the AFBS and AFBS_ACC algorithms are more efficient than the CCCP based algorithm for large size problems. In addition, we observe that AFBS and AFBS_ACC converge to a smaller objective function value than CCCP. It means that our methods achieve more accuracy approximation to the original problem (2).

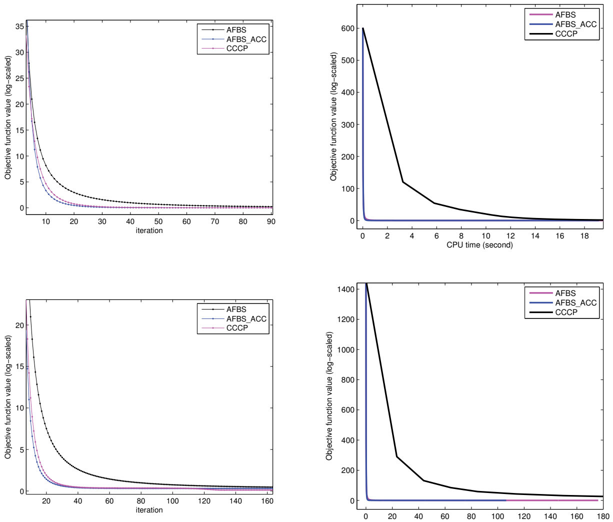

Figure 2: Convergence performance of objective function value with respect to iterations (left column) and running times (right column). First row: $n = 4000, K = 120, d = 4810$. Second row: $n = 10000, K = 300, d = 12010$.

In the third experiment, we show the sparse pattern recovery performance of the models (13), (14), and (15). We solve (13) by SLEP, and solve (14) and (15) by AFBS. We use two criteria to evaluate the recovery performance: variable selection error (VSE) and group selection error (GSE). And the VSE and GSE are the proportion of wrongly selected variables and groups respectively in the estimator $\hat{\mathbf{x}}$ based on the true $\mathbf{x}^*$. Figure 3 reports the result with 10 repeatitions. Clearly, the nonconvex approach outperforms its convex counterpart.

Graph-guided Fused Logistic Regression

We apply the ℓ_1,ℓ_0 norm, and capped-ℓ_1 penalty to the graph-guided fused logistic regression. It leads to the following problems:

$$\min_{\mathbf{x}} f(\mathbf{x}) + \lambda \sum_{(i,j)\in E} |x_i - x_j|, \qquad (16)$$

$$\min_{\mathbf{x}} f(\mathbf{x}) + \lambda \sum_{(i,j)\in E} \min(|x_i - x_j|, \theta), \qquad (17)$$

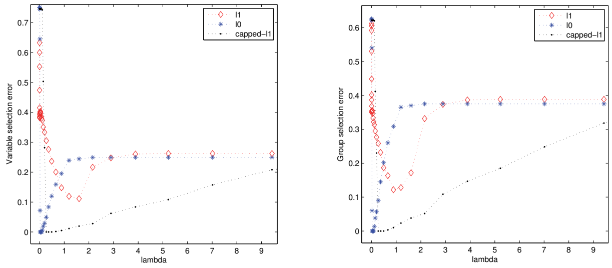

Figure 3: Sparse patten recovery performance of the models (13), (14), and (15). $n = 500, K = 60$, $\theta = 0.1$, lambda= logspace$(-3, 2, 40)$.

$$\min_{\mathbf{x}} f(\mathbf{x}) + \lambda \sum_{(i,j)\in E} \mathbf{I}(|x_i - x_j| \neq 0), \qquad (18)$$

where f is the logistic loss and E is the set of edges for the graph. We construct the graph $\mathcal{G} := \{V, E\}$ by using the sparse inverse covariance selection on the training data set (Banerjee, Ghaoui, and d'Aspremont 2008).

We conduct experiment on the *20newsgroup* [†] data set. The 20newsgroup data set contains $16,242$ samples with 100 binary features (words). There are four classes of the 20newsgroup data set: *computer*, *recreation*, *science*, *talks*. We divide the classification task into four one *v.s.* all classifiers. The samples are randomly divided into three subsets: 1% as the training data, 70% as the testing data, and the rest as the validation data. Our main goal here is to demonstrate that the nonconvex approach performs better than its convex counterpart. Thus we choose the convex method as the baseline. Table 2 shows the classification accuracy of the AFBS algorithm on the 20newsgroup dataset with 10 repetitions. We can see that the ℓ_0 norm based approach achieves the best performance.

Table 2: Classification accuracy (%) with graph-guided fused logistic regression on the 20newsgroup dataset. "ggflr-ℓ_0" denotes the proposed graph-guided fused logistic regression with the ℓ_0 norm. "ggflr-capped" is with the capped-ℓ_1 regularizer. "ggflr-ℓ_1" is with the ℓ_1 norm.

data set	ggflr-ℓ_1	ggflr-capped	ggflr-ℓ_0
com. *vs* rest	82.32(±0.013)	84.83(±0.014)	**84.93**(±0.013)
rec. *vs* rest	86.34(±0.017)	87.35(±0.013)	**90.07**(±0.009)
sci. *vs* rest	79.53(±0.016)	83.02(±0.01)	**85.58**(±0.005)
talk. *vs* rest	83.91(±0.02)	85.17(±0.016)	**86.47**(±0.01)

Conclusion

In this paper, we have proposed an alternating Forward-Backward splitting method for solving structured sparsity optimization problems. The AFBS method alternates between a gradient step and an easily solvable proximal step, and thus enjoys low per-iteration computational complexity.

[†] http://www.cs.nyu.edu/∼roweis/data.html

Furthermore, we have established the global convergence of the AFBS method. We also have devised an accelerated variant of the AFBS method, which has better empirical performance. Moreover, we have proposed the ℓ_0 norm penalized overlapping group lasso and graph-guided fused logistic regression for the first time, which can be solved by the AFBS method and its accelerated variant.

Acknowledgments

Shubao Zhang and Hui Qian acknowledge support from the National Natural Science Foundations of China (No. 61472347). Xiaojin Gong acknowledges support from the State High-Tech Development Plan (863 Program) of China (2014AA09A510).

References

Argyriou, A.; Micchelli, C. A.; Pontil, M.; Shen, L.; and Xu, Y. 2011. Efficient first order methods for linear composite regularizers. *arXiv:1104.1436v1*.

Attouch, H., and Bolte, J. 2009. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming* 116(1-2, Ser. B):5–16.

Banerjee, O.; Ghaoui, L. E.; and d'Aspremont, A. 2008. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research* 9:485–516.

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

Bochnak, J.; Coste, M.; and Roy, M. F. 1998. *Real algebraic geometry*. Springer.

Bolte, J.; Daniilidis, A.; and Lewis, A. 2007. The lojasiewicz inequality for nonsmooth sub-analytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization* 17(4):1205–1223.

Bolte, J.; Sabach, S.; and Teboulle, M. 2014. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming* 146(A-B):459–494.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends of Machine Learning* 3(1):1–122.

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

Combettes, P., and Pesquet, J. 2011. Proximal splitting methods in signal processing. In *Fixed-point Algorithms for Inverse Problems in Science and Engineering*. Springer.

Daubechies, I.; Devore, R.; Fornasier, M.; and Güntürk, C. S. 2010. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics* 63(1):1–38.

Deng, W.; Yin, W. T.; and Zhang, Y. 2011. Group sparse optimization by alternating direction method. *Rice CAAM Report TR11-06*.

Fan, J., and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456):1348–1360.

Gasso, G.; Rakotomamonjy, A.; and Canu, S. 2009. Recovering sparse signals with a certain family of non-convex penalties and dc programming. In *IEEE Transactions on Signal Proessing*, volume 57, 4686–4698.

Gong, P. H.; Zhang, C. S.; Zhao, Z. S.; Huang, J. Z.; and Ye, J. P. 2013. A general iterative shrinkage and thresholding algorithm for nonconvex regularized optimization problems. In *Proceedings of 30th International Conference on Machine Learning*.

Hastie, T. J.; Tibshirani, R. J.; and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, second edition.

Huang, J.; Wei, F.; and Ma, S. 2012. Semiparametric regression pursuit. *Statist Sinica* 22(1):1403–1426.

Hunter, D., and Li, R. 2005. Variable selection using MM algorithms. *The Annals of Statistics* 33(4):1617–1642.

Liu, J.; Ji, S.; and Ye, J. 2009. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University.

Luenberger, D. G., and Ye, Y. 2008. *Linear and Nonlinear Programming*. New York: Springer.

Nesterov, Y. 2007. Gradient methods for minimizing composite functions. *Mathematical Programming*.

Ouyang, Y. Y.; Chen, Y. M.; Lan, G. H.; and Jr., E. P. 2015. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences* 8(1):644–681.

Tibshirani, R. J., and Taylor, J. 2011. The solution path of the generalized lasso. *The Annals of Statistics* 39(3):1335–1371.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58:267–288.

Wang, L.; Chen, G.; and Li, H. 2007. Group scad regression analysis for microarray time course gene expression data. *Bioinformatics* 23:1486–1494.

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

Yuan, M., and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*.

Yuan, L.; Liu, J.; and Ye, J. P. 2013. Efficient methods for overlapping group lasso. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(9).

Yuille, A. L., and Rangarajan, A. 2001. The concave-convex procedure (cccp). In *Advances in Neural Information Processing Systems*.

Zangwill, W. I. 1969. *Nonlinear programming: a unified approach*. Englewood Cliffs, N. J.: Prentice-Hall.

Zhang, S. B.; Qian, H.; Chen, W.; and Zhang, Z. H. 2013. A concave conjugate approach for nonconvex penalized regression with the mcp penalty. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Zhang, C.-H. 2010a. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38:894–942.

Zhang, T. 2010b. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research* 11:1081–1107.

Zhao, P.; Rocha, G.; and Yu, B. 2009. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics* 37(6A):3468–3497.