GSOS: Gauss-Seidel Operator Splitting Algorithm for Multi-Term Nonsmooth Convex Composite Optimization

Li Shen [1] **Wei Liu** [1] **Ganzhao Yuan** [2] **Shiqian Ma** [3]

Abstract

In this paper, we propose a fast **G**auss-**S**eidel **O**perator **S**plitting (GSOS) algorithm for addressing multi-term nonsmooth convex composite optimization, which has wide applications in machine learning, signal processing and statistics. The proposed GSOS algorithm inherits the advantage of the Gauss-Seidel technique to accelerate the optimization procedure, and leverages the operator splitting technique to reduce the computational complexity. In addition, we develop a new technique to establish the global convergence of the GSOS algorithm. To be specific, we first reformulate the iterations of GSOS as a two-step iterations algorithm by employing the tool of operator optimization theory. Subsequently, we establish the convergence of GSOS based on the two-step iterations algorithm reformulation. At last, we apply the proposed GSOS algorithm to solve overlapping group Lasso and graph-guided fused Lasso problems. Numerical experiments show that our proposed GSOS algorithm is superior to the state-of-the-art algorithms in terms of both efficiency and effectiveness.

1. Introduction

In this paper, we focus on the multi-term nonsmooth convex composite optimization

$$\min_{x\in\mathcal{X}} f(x)+\sum_{i=1}^{n} g_i(x), \tag{1}$$

where $\mathcal{X}$ is a linear space, $g_i : \mathcal{X} \to (-\infty,+\infty]$ is a proper, lower semicontinuous convex function for all $i = 1,\cdots,n$, and $f : \mathcal{X} \to (-\infty,+\infty)$ is a continuous differentiable convex function with its gradient satisfying the inequality that

$$\frac{1}{L}\left\|\nabla f(x)-\nabla f(y)\right\|^2 \le \left\langle \nabla f(x)-\nabla f(y), x-y\right\rangle. \tag{2}$$

The above multi-term nonsmooth convex composite optimization problem (1) covers a large class of applications in machine learning such as simultaneous low-rank and sparsity (Richard et al., 2012; Zhou et al., 2013), overlapping group Lasso (Zhao et al., 2009; Jacob et al., 2009; Mairal et al., 2010), graph-guided fused Lasso (Chen et al., 2012; Kim & Xing, 2009), graph-guided logistic regression (Chen et al., 2011; Zhong & Kwok, 2014), variational image restoration (Combettes & Pesquet, 2011; Dupé et al., 2009; Pustelnik et al., 2011), and other types of structure regularization paradigms (Teo et al., 2010; 2007). By introducing the multi-term nonsmooth regularization term $\sum_{i=1}^{n} g_i(x)$ such as structured sparsity (Huang et al., 2011; Bach et al., 2012; Bach, 2010) and nonnegativity (Chen & Plemmons, 2015; Xu & Yin, 2013), more prior information can be included to enhance the accuracy of regularization models. However, due to the multi-term nonsmooth regularization term $\sum_{i=1}^{n} g_i(x)$, the optimization problem (1) is too complicated to be solved even for small n. For $n \le 2$, some existing popular first-order optimization methods are accelerated proximal gradient method (Beck & Teboulle, 2009; Nesterov, 2007), smoothing accelerated proximal gradient method (Nesterov, 2005a;b), three operator splitting method (Davis & Yin, 2015), and some primal-dual operator splitting methods such as majorized alternating direction method of multiplier (ADMM) (Cui et al., 2016; Lin et al., 2011), fast proximity method (Li & Zhang, 2016), and so on.

On the other hand, when $n \ge 3$, there also exist some algorithms for solving problem (1). A directly method for (1) is smoothing accelerated proximal gradient (S-APG) proposed by Nesterov (Nesterov, 2005a;b). Then, Yu (Yu, 2013) proposed a new approximation method called PA-APG for handling (1) by combining the proximal average approximation technique and Nesterov's acceleration technique, which has been enhanced very recently by Shen *et al.* (Shen et al., 2017). Their proposed method called APA-APG adopts an adaptive stepsize strategy. However,

[1]Tencent AI Lab, China [2]Sun Yat-sen University, China [3]The Chinese University of Hong Kong, China. Correspondence to: Li Shen <mathshenli@gmail.com>, Wei Liu <wliu@ee.columbia.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

the above mentioned methods S-APG, PA-APG and its enhanced version APA-APG all need a strict restriction on the nonsmooth functions $\{g_i(x)\}$ that each $g_i(x)$ must be Lipschitz continuous. In addition, some primal-dual parallel splitting methods (Briceno-Arias et al., 2011; Combettes & Pesquet, 2007; 2008; Condat, 2013; Vũ, 2013) generalized from traditional operator splitting, such as forward backward splitting method (Chen & Rockafellar, 1997) and Douglas Rachford splitting method (Eckstein & Bertsekas, 1992), can also solve the multi-term nonsmooth convex composite optimization problem (1). Different from prior work, Raguet *et al.* (Raguet et al., 2013) proposed an efficient primal operator splitting method called generalized forward backward splitting method using the classic forward backward splitting technique, which has shown the superiority over numerous existing primal-dual splitting methods (Monteiro & Svaiter, 2013; Combettes & Pesquet, 2012; Chambolle & Pock, 2011) in dealing with variational image restoration problems. All the above mentioned methods for problem (1) with $n \geq 3$ share a common feature that they all split the nonsmooth composite term $\sum_{i=1}^n g_i(x)$ in the Jacobi iteration manner, *i.e.*, parallelly. This is one of the main differences between existing splitting methods and our proposed method in this paper.

To split the nonsmooth composite term $\sum_{i=1}^n g_i(x)$ more efficiently, we propose a novel operator splitting algorithm to solve problem (1) by harnessing the advantage of Gauss-Seidel iterations, *i.e.*, the computation of the proximal mapping of the current function $g_i(x)$ uses the proximal mappings of $g_j(x)$ for all $j < i$ which have already been computed ahead. In addition, to further improve the algorithm's efficiency, we leverage the over-relaxation acceleration technique. What's more, we provide a new strategy that the over-relaxation stepsize can be determined adaptively, ensuring a larger value to accelerate the algorithm. The most important is that the convergence of our proposed GSOS algorithm is established by a newly developed analysis technique. In detail, given an invertible linear operator $\mathcal{R}$, we first argue that the optimal solution set $[\nabla f + \sum_{i=1}^n \partial g_i]^{-1}(0)$ of problem (1) can be recovered by the zero point set $\left[(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\,\partial g+\mathcal{A}\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_\mathcal{V}}\right]^{-1}(0)$. This is fulfilled through adopting the tool of operator optimization theory, in which the composite operator $\mathcal{S}_{\mathcal{R},\,\partial g+\mathcal{A}\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_\mathcal{V}}$ is generalized from the definition of the composite monotone operator $\mathcal{S}_{\lambda,\mathcal{A},\mathcal{B}}$ in (Eckstein & Bertsekas, 1992). Next, by unitizing the definition of the ϵ-enlargement of maximal monotone (Burachik et al., 1998; 1997; Burachik & Svaiter, 1999; Svaiter, 2000), we establish a key property for $\mathcal{S}_{\mathcal{R},\,\partial g+\mathcal{A}\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_\mathcal{V}}$, that is, $\mathrm{gph}\left(\mathcal{S}_{\mathcal{R},\,(\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A})^{[\epsilon]},\,\mathcal{N}_\mathcal{V}}\right) \subseteq \mathrm{gph}\left(\mathcal{R}^*[(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\,\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_\mathcal{V}}]^{[\epsilon]}\right)$. Based on this observation, we equivalently reformulate the GSOS algorithm as a two-step iterations algorithm. Then, the global convergence of the proposed GSOS algorithm is easily established based on this reformulation.

The closest algorithm to our proposed GSOS algorithm is the generalized forward backward splitting method proposed by Raguet *et al.* (Raguet et al., 2013). By carefully selecting the scaling matrix $\mathcal{H}$ in the forthcoming GSOS algorithm, it is easy to check that GSOS covers the generalized forward backward splitting method as a special case. Another highly related algorithm to our proposed GSOS algorithm is the matrix splitting method (Luo & Tseng, 1991; Yuan et al., 2016). Choosing the scaling matrix $\mathcal{H}$ suitably, the proposed GSOS algorithm can inherit the advantage of the matrix splitting technique which has shown the efficiency in (Yuan et al., 2016) for coping with a special class of coordinate separable composite optimization problems.

The rest of this paper is organized as follows. In Section 2, we first give the definitions of some useful notations which can make the paper much more readable. We also establish some lemmas and propositions based on monotone operator theory (Bauschke & Combettes, 2011), which are the key to the convergence of the GSOS algorithm. In Section 3, we present the proposed GSOS algorithm and then analyze its convergence and iteration complexity. In Section 4, we conduct numerical experiments on overlapping group Lasso and graph-guided fused Lasso problems to evaluate the efficacy of the GSOS algorithm. Finally, we draw conclusions in Section 5.

2. Preliminaries and Notations

Let $\mathcal{Y} = \prod_{i=1}^n \mathcal{X}_i$ be the product space of $\mathcal{X}_i$ with $\mathcal{X}_i = \mathcal{X}$ for all $i \in \{1, 2, \cdots, n\}$. Let $\mathcal{V}$ be a linear space and $\mathcal{V}^\perp$ be its complementary space with the following definitions

$$\mathcal{V} = \{y \in \mathcal{Y} \mid y_1 = \cdots = y_n\},\ \mathcal{V}^\perp = \{y \in \mathcal{Y} \mid \sum_i^n y_i = 0\}.$$

Let $\mathcal{I}_\mathcal{X} : \mathcal{X} \to \mathcal{X}$ be the identity map and $\mathcal{E}_\mathcal{Y} : \mathcal{X} \to \mathcal{Y}$ be a block linear operator defined as $\mathcal{E}_\mathcal{Y} = \begin{pmatrix} \mathcal{I}_\mathcal{X} & \cdots & \mathcal{I}_\mathcal{X} \end{pmatrix}^*$. Let $\mathcal{A} : \mathcal{Y} \to \mathcal{X}$ be a linear operator defined as $\mathcal{A}y = \frac{1}{n}\mathcal{E}_\mathcal{Y}^* y = \frac{1}{n}\sum_{i=1}^n y_i$. Hence, its adjoint operator $\mathcal{A}^* : \mathcal{X} \to \mathcal{Y}$ is defined as $\mathcal{A}^* x = \frac{1}{n}\mathcal{E}_\mathcal{Y} x$. Let $\mathcal{H}, \mathcal{R} : \mathcal{Y} \to \mathcal{Y}$ be block lower triangular linear invertible operators satisfying $(\mathcal{R}^*)^{-1} = \mathcal{H}$ and $\mathcal{H} + \mathcal{H}^* \succ 0$. Moreover, $\mathcal{H}$ is defined as

$$\begin{pmatrix} \mathcal{H}_{1,1} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \mathcal{H}_{n-1,1} & \cdots & \mathcal{H}_{n-1,n-1} & 0 \\ \mathcal{H}_{n,1} & \cdots & \mathcal{H}_{n-1,n} & \mathcal{H}_{n,n} \end{pmatrix}, \tag{3}$$

where $\mathcal{H}_{i,j} : \mathcal{X} \to \mathcal{X}$ is a linear operator for all $(i, j) \in \{1, \cdots, n\}$. It is worthwhile to emphasize that $\mathcal{H}_{i,i}$ is also possible to be a lower triangular linear operator satisfying

$\mathcal{H}_{i,i}+\mathcal{H}_{i,i}^*\succ 0$. Next, we abuse the notation $\|\cdot\|_{\mathcal{H}}$ which is induced by the inner product $\langle\cdot,\mathcal{H}\cdot\rangle$ satisfying

$$\begin{aligned}\|\cdot\|_{\mathcal{H}}:&=\sqrt{\langle\cdot,\mathcal{H}\cdot\rangle}=\sqrt{\langle\cdot,\mathcal{H}^*\cdot\rangle}\\&=\sqrt{\langle\cdot,\frac{\mathcal{H}+\mathcal{H}^*}{2}\cdot\rangle}=\|\cdot\|_{\frac{\mathcal{H}+\mathcal{H}^*}{2}}.\end{aligned}\tag{4}$$

In addition, we define the generalized proximal mapping of a proper, lower semicontinuous convex function $g_i(x)$ with respect to the invertible linear operator $\mathcal{H}_{i,i}$.

Definition 1 *For a given x, the proximal mapping denoted by $\mathrm{Prox}_{\mathcal{H}_{i,i}^{-1}g_i}(x)$ of a proper, lower semicontinuous convex function g_i with respect to an invertible linear operator $\mathcal{H}_{i,i}$ satisfying $H_{i,i}+\mathcal{H}_{i,i}^*\succ 0$ is defined to be the zero point of the following inclusion equation*

$$0\in\partial g_i(\cdot)+\mathcal{H}_{i,i}(\cdot-x).\tag{5}$$

Moreover, if $\mathcal{H}_{i,i}$ is symmetric, it can be reformulated as the following convex minimization

$$\mathrm{Prox}_{\mathcal{H}_{i,i}g_i}(x):=\arg\min_{y\in\mathcal{X}}g_i(y)+\frac{1}{2}\|y-x\|^2_{\mathcal{H}_{i,i}}.$$

Next, we recall the definition of ϵ-enlargement of monotone operators (Burachik et al., 1998; 1997; Burachik & Svaiter, 1999; Svaiter, 2000), which is an effective tool for establishing the convergence of the proposed GSOS algorithm.

Definition 2 *Given a maximal monotone operator $T:\mathbb{X}\Rightarrow\mathbb{X}$, the $\epsilon(\geq 0)$-enlargement of T is defined as the set $T^{[\epsilon]}(x):=\{v\in\mathbb{Y}\mid\langle w-v,z-x\rangle\geq-\epsilon$ for all $z\in\mathbb{X},w\in T(z)\}$.*

Recall that $f(x)$ is a gradient Lipschitz convex function satisfying inequality (2). There exits $0\preceq\Sigma\preceq\widehat{\Sigma}\preceq L\mathcal{I}$ such that the following two inequalities hold for any $x,x'\in\mathcal{X}$

$$f(x)\leq f(x')+\langle\nabla f(x'),x-x'\rangle+\frac{1}{2}\|x-x'\|^2_{\widehat{\Sigma}},\tag{6}$$

$$f(x)\geq f(x')+\langle\nabla f(x'),x-x'\rangle+\frac{1}{2}\|x-x'\|^2_{\Sigma}.\tag{7}$$

Actually, when $f(x)$ is a quadratic function, it holds $\Sigma=\widehat{\Sigma}$ directly in inequalities (6) and (7). The following lemma establishes the property of the enlargement of the composite operator $\mathcal{A}^*\circ\nabla f\circ\mathcal{A}$ with f satisfying inequalities (6)-(7) or (2), which is an essential ingredient for reformulating the GSOS algorithm as a two-step iterations algorithm.

Proposition 1 *Assume that f is a gradient Lipschitz continuous convex function satisfying inequality (2). For any $x_1,\,x_2\in\mathcal{Y}$, it holds that*

$$(\mathcal{A}^*\circ\nabla f\circ\mathcal{A})(x_2)\in(\mathcal{A}^*\circ\nabla f\circ\mathcal{A})^{[\epsilon]}(x_1)\tag{8}$$

with $\epsilon=\frac{L}{4}\|\mathcal{A}x_1-\mathcal{A}x_2\|^2$. In addition, if f further satisfies inequalities (6)-(7), it holds that

$$(\mathcal{A}^*\circ\nabla f\circ\mathcal{A})(x_2)\in(\mathcal{A}^*\circ\nabla f\circ\mathcal{A})^{[\epsilon]}(x_1)\tag{9}$$

with $\epsilon=\frac{1}{4}\|\mathcal{A}x_1-\mathcal{A}x_2\|^2_{2\widehat{\Sigma}-\Sigma}$.

Remark 1 *Two comments are made for Proposition 1:*

(1) *This proposition gives two types of estimations for ϵ in $\left(\mathcal{A}^*\circ\nabla f\circ\mathcal{A}\right)^{[\epsilon]}$ in (8) and (9). When f is a quadratic function, it is easy to check that*

$$\frac{1}{4}\|\mathcal{A}x_1-\mathcal{A}x_2\|^2_{2\widehat{\Sigma}-\Sigma}\leq\frac{L}{4}\|\mathcal{A}x_1-\mathcal{A}x_2\|^2$$

due to $\widehat{\Sigma}=\Sigma\preceq L\mathcal{I}$. When f is a general gradient Lipschitz continuous function, we do not know which estimation for ϵ is tighter in (8) and (9).

(2) *The second part of this proposition can be regarded as an intensified version of Lemma 2.2 in (Svaiter, 2014) for a specified composite operator $\mathcal{A}^*\circ\nabla f\circ\mathcal{A}$. The first part of the proposition coincides with the results by applying Lemma 2.2 in (Svaiter, 2014) for $\mathcal{A}^*\circ\nabla f\circ\mathcal{A}$.*

Next, we generalize the notation $\mathcal{S}_{\lambda,\mathcal{T}_1,\mathcal{T}_2}$ in (Eckstein & Bertsekas, 1992) for a given $\lambda>0$ and two maximal monotone operators $\mathcal{T}_1,\mathcal{T}_2$ as $\mathcal{S}_{\mathcal{R},\mathcal{T}_1,\mathcal{T}_2}$ for a given invertible linear operator $\mathcal{R}$ defined as

$$\begin{aligned}&\mathrm{gph}\,\mathcal{S}_{\mathcal{R},\mathcal{T}_1,\mathcal{T}_2}\\:=&\Big\{(x_1+\mathcal{R}y_2,x_2-x_1)\mid y_1\in\mathcal{T}_1(x_1),\\&\qquad y_2\in\mathcal{T}_2(x_2),x_1+\mathcal{R}^*y_1=x_2-\mathcal{R}^*y_2\Big\}.\end{aligned}\tag{10}$$

By (Eckstein & Bertsekas, 1992), we know that $\mathcal{S}_{\lambda,\mathcal{T}_1,\mathcal{T}_2}$ is maximal monotone if $\mathcal{T}_1$ and $\mathcal{T}_2$ are both maximal monotone. However, its generalized operator $\mathcal{S}_{\mathcal{R},\mathcal{T}_1,\mathcal{T}_2}$ is not monotone unless the invertible linear operator $\mathcal{R}$ reduces to be a constant. Very interesting, it can be shown that its composition with $(\mathcal{R}^*)^{-1}$, *i.e.*, $(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\mathcal{T}_1,\mathcal{T}_2}$, is maximal monotone for any invertible linear operator $\mathcal{R}$.

Lemma 1 *For any given invertible linear operator $\mathcal{R}$, operator $(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\mathcal{T}_1,\mathcal{T}_2}$ is maximal monotone if $\mathcal{T}_1$ and $\mathcal{T}_2$ are both maximal monotone operators.*

Setting $\mathcal{T}_1=\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A}$, $\mathcal{T}_2=\mathcal{N}_{\mathcal{V}}$, we obtain $\mathcal{S}_{\mathcal{R},\,\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_{\mathcal{V}}}$, which is defined as

$$\begin{aligned}&\mathrm{gph}\left(\mathcal{S}_{\mathcal{R},\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_{\mathcal{V}}}\right)\\:=&\Big\{(x_1+\mathcal{R}y_2,x_2-x_1)\mid y_1\in(\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A})(x_1),\\&\qquad y_2\in\mathcal{N}_{\mathcal{V}}(x_2),x_1+\mathcal{R}^*y_1=x_2-\mathcal{R}^*y_2\Big\}.\end{aligned}\tag{11}$$

By Lemma 1, we know that $(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\,\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_\mathcal{V}}$ is maximal monotone due to the maximal monotonicity of $\partial g + \mathcal{A}^* \circ \nabla f \circ \mathcal{A}$ and $\mathcal{N}_\mathcal{V}$. Hence, given a constant $\epsilon \geq 0$, the enlargement $[(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\,\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_\mathcal{V}}]^{[\epsilon]}$ is well defined. In addition, based on the definition of $\mathcal{S}_{\mathcal{R},\mathcal{T}_1,\mathcal{T}_2}$ again, we set $\mathcal{T}_1 = \partial g + (\mathcal{A}^* \circ \nabla f \circ \mathcal{A})^{[\epsilon]}$, or $\mathcal{T}_1 = (\partial g + \mathcal{A}^* \circ \nabla f \circ \mathcal{A})^{[\epsilon]}$ and $\mathcal{T}_2 = \mathcal{N}_\mathcal{V}$ in (10). Then we have the definition of $\mathcal{S}_{\mathcal{R},\partial g+(\mathcal{A}^*\circ\nabla f\circ\mathcal{A})^{[\epsilon]},\mathcal{N}_\mathcal{V}}$ or $\mathcal{S}_{\mathcal{R},(\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A})^{[\epsilon]},\mathcal{N}_\mathcal{V}}$ for any given invertible linear operator $\mathcal{R}$ and constant $\epsilon \geq 0$ as follows

$$\begin{aligned}
&\mathrm{gph}\left(\mathcal{S}_{\mathcal{R},\partial g+(\mathcal{A}^*\circ\nabla f\circ\mathcal{A})^{[\epsilon]},\mathcal{N}_\mathcal{V}}\right) \qquad (12)\\
&:= \Big\{(x_1+\mathcal{R}y_2, x_2-x_1) | y_1 \in (\partial g+(\mathcal{A}^*\circ\nabla f\circ\mathcal{A})^{[\epsilon]})(x_1),\\
&\qquad\qquad y_2 \in \mathcal{N}_\mathcal{V}(x_2), x_1+\mathcal{R}^*y_1 = x_2-\mathcal{R}^*y_2\Big\},\\
&\mathrm{gph}\left(\mathcal{S}_{\mathcal{R},(\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A})^{[\epsilon]},\mathcal{N}_\mathcal{V}}\right) \qquad (13)\\
&:= \Big\{(x_1+\mathcal{R}y_2, x_2-x_1) | y_1 \in (\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A})^{[\epsilon]})(x_1),\\
&\qquad\qquad y_2 \in \mathcal{N}_\mathcal{V}(x_2), x_1+\mathcal{R}^*y_1 = x_2-\mathcal{R}^*y_2\Big\}.
\end{aligned}$$

In the proposition below, we will establish the relationships among the above mentioned three operators $\mathcal{S}_{\mathcal{R},\partial g+(\mathcal{A}^*\circ\nabla f\circ\mathcal{A})^{[\epsilon]},\mathcal{N}_\mathcal{V}}$, $\mathcal{S}_{\mathcal{R},(\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A})^{[\epsilon]},\mathcal{N}_\mathcal{V}}$ and $[(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\,\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_\mathcal{V}}]^{[\epsilon]}$.

Proposition 2 *Given a constant $\epsilon \geq 0$ and an invertible linear operator $\mathcal{R}$, it holds that*

$$\begin{aligned}
&\mathrm{gph}\left(\mathcal{S}_{\mathcal{R},\,\partial g+(\mathcal{A}^*\circ\nabla f\circ\mathcal{A})^{[\epsilon]},\,\mathcal{N}_\mathcal{V}}\right)\\
&\subseteq \mathrm{gph}\left(\mathcal{S}_{\mathcal{R},\,(\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A})^{[\epsilon]},\,\mathcal{N}_\mathcal{V}}\right)\\
&\subseteq \mathrm{gph}\left(\mathcal{R}^*[(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\,\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_\mathcal{V}}]^{[\epsilon]}\right).
\end{aligned}$$

In the following, we establish the relationship between the optimal solution set $[\nabla f + \sum_{i=1}^n \partial g_i]^{-1}(0)$ of problem (1) and $\left[(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\,\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_\mathcal{V}}\right]^{-1}(0)$, which means that we can recover the solution of problem (1) through $\left[(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\,\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_\mathcal{V}}\right]^{-1}(0)$.

Lemma 2 *Let linear operators $\mathcal{H}$ and $\mathcal{R}$ satisfy $(\mathcal{R}^*)^{-1} = \mathcal{H}$ and $\mathcal{H}$ satisfy (3). Denote $\Omega = \left[(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\,(\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A}),\,\mathcal{N}_\mathcal{V}}\right]^{-1}(0)$. It holds that*

$$\left[\nabla f + \sum_{i=1}^n \partial g_i\right]^{-1}(0) = \left(\mathcal{E}_\mathcal{Y}^T\mathcal{H}^*\mathcal{E}_\mathcal{Y}\right)^{-1}\mathcal{E}_\mathcal{Y}^T\mathcal{H}^*(\Omega).$$

3. GSOS Algorithm

In this section, we first propose the Gauss-Seidel operator splitting algorithm for solving the multi-term nonsmooth convex composite problem (1). Then, based on the preliminaries in Section 2, we establish the convergence and iteration complexity of the GSOS algorithm.

Algorithm 1 GSOS Algorithm

Parameters: Choose $\overline{\sigma} \in (0,1)$, a linear operator $\mathcal{H}$ satisfying (3) and a starting point $z^0 \in \mathcal{Z}$. Set $\theta^{\mathrm{fix1}} \in \big(-1,\ \overline{\theta}_1\big]$ and $\theta^{\mathrm{fix2}} \in \big(-1,\ \overline{\theta}_2\big]$, where $\overline{\theta}_1$ and $\overline{\theta}_2$ are defined via equations (14a) and (14b), respectively.
for $k = 0, 1, 2, \cdots, K$ **do**
  $x^k := \mathcal{E}_\mathcal{Y}\left(\mathcal{E}_\mathcal{Y}^T\mathcal{H}\mathcal{E}_\mathcal{Y}\right)^{-1}\mathcal{E}_\mathcal{Y}^T\mathcal{H}z^k$;
  for $i = 1, 2\cdots, n$ **do**
    $y_i^k := \mathrm{Prox}_{\mathcal{H}_{i,i}^{-1}g_i}\Big(\mathcal{H}_{i,i}^{-1}[\sum_{j=1}^i \mathcal{H}_{i,j}(2x_j^k - z_j^k) - \frac{1}{n}\nabla f(\frac{1}{n}\sum_{i=1}^n x_i^k) - \sum_{j=1}^{i-1}\mathcal{H}_{i,j}y_j^k]\Big)$;
  end for
  set $\theta_k^{\mathrm{adap1}}$ as (14c) and $\theta_k^{\mathrm{adap2}}$ as (14d);
  set $\theta_k \in [\theta^{\mathrm{fix1}}, \theta_k^{\mathrm{adap1}}] \cup [\theta^{\mathrm{fix2}}, \theta_k^{\mathrm{adap2}}]$;
  $z^{k+1} := z^k + (1+\theta_k)(y^k - x^k)$;
end for
return $\omega^K := \left(\mathcal{E}_\mathcal{Y}^T\mathcal{H}^*\mathcal{E}_\mathcal{Y}\right)^{-1}\mathcal{E}_\mathcal{Y}^T\mathcal{H}^*z^K$.

In Algorithm 1, parameters $\overline{\theta}_1, \overline{\theta}_1$, $\theta_k^{\mathrm{adap1}}, \theta_k^{\mathrm{adap1}}$ are defined as

$$\begin{cases}
\overline{\theta}_1 = \max\Big\{\theta | (\theta-\overline{\sigma})(\mathcal{H}+\mathcal{H}^*) + L\mathcal{A}^*\mathcal{A} \preceq 0\Big\}; & (14a)\\
\overline{\theta}_2 = \max\Big\{\theta \mid (\theta-\overline{\sigma})(\mathcal{H}+\mathcal{H}^*) & (14b)\\
\qquad\qquad +\mathcal{A}^*(2\widehat{\Sigma}-\Sigma)\mathcal{A} \preceq 0\Big\};\\
\theta_k^{\mathrm{adap1}} = \overline{\sigma} - \dfrac{L\|\mathcal{A}(x^k-y^k)\|^2}{\|x^k-y^k\|^2_{\mathcal{H}+\mathcal{H}^*}}; & (14c)\\
\theta_k^{\mathrm{adap2}} = \overline{\sigma} - \dfrac{\|\mathcal{A}(x^k-y^k)\|^2_{2\widehat{\Sigma}-\Sigma}}{\|x^k-y^k\|^2_{\mathcal{H}+\mathcal{H}^*}}. & (14d)
\end{cases}$$

Remark 2 *We make some comments on GSOS below.*

(1) *For the updating step of x^k, we obtain $x^k = \mathcal{E}_\mathcal{Y}\big(\sum_{i,j=1}^K \mathcal{H}_{ij}\big)^{-1}\sum_{j=1}^K\sum_{i=j}^K \mathcal{H}_{ij}z_j^k$ by using the notations $\mathcal{H}$ and $\mathcal{E}_\mathcal{Y}$. Similarly, we have $\omega^k = \big(\sum_{i,j=1}^K \mathcal{H}_{ij}\big)^{-1}\sum_{j=1}^K\sum_{i=i}^j \mathcal{H}_{ji}^*z_j^k$. Hence, we need to compute the inverse of $\sum_{i,j=1}^n \mathcal{H}_{i,j}$. However, if $\mathcal{H}_{i,j}$ is a lower triangular matrix operator, x^k and ω^k can be obtained easily.*

(2) *By the definitions of $\mathrm{Prox}_{\mathcal{H}_{i,i}^{-1}g_i}$ and y^k, we need to solve the following inclusion equation*

$$G_i^k \in \mathcal{H}_{i,i}y_i^k + \partial g_i(y_i^k),$$

where $G_i^k = \mathcal{H}_{i,i}^{-1}\big[\sum_{j=1}^i \mathcal{H}_{i,j}(2x_j^k - z_j^k) - \frac{1}{n}\nabla f(\frac{1}{n}\sum_{i=1}^n x_i^k) - \sum_{j=1}^{i-1}\mathcal{H}_{i,j}y_j^k\big]$. Usually, it is easy to choose a suitable $\mathcal{H}_{i,i}$ such that the solution of the above inclusion equation has a closed form.

(3) *θ_k is the over-relaxation stepsize for accelerating the GSOS algorithm. If the computations of θ_k^{adap1} and θ_k^{adap2} are time consuming, we can set $\theta_k = \max\{\theta^{\text{fix1}}, \theta^{\text{fix2}}\}$.*

(4) *When $\mathcal{H}$ is a diagonal matrix, i.e., $\mathcal{H}_{i,j} = 0$ and $\mathcal{H}_{i,i} = a_i\mathcal{I}$ with some nonnegative constant a_i, and the over relaxation stepsize θ_k is fixed to a smaller region, the GSOS algorithm reduces to the generalized forward backward splitting method in (Raguet et al., 2013).*

In the following, we reformulate the GSOS algorithm as a two-step iterations algorithm by utilizing monotone optimization theory established in Section 2, which is the key to the convergence of the GSOS algorithm.

Proposition 3 *Let $g : \mathcal{Y} \to (-\infty, +\infty]$ be the function defined as $g(x) = \sum_{i=1}^n g_i(x_i)$. Assume that the sequences (x^k, y^k) and z^k are generated by Algorithm 1 with $\overline{\sigma} \in (0, 1)$. Let $v^k = (\mathcal{R}^*)^{-1}(x^k - y^k)$ and $\overline{z}^k = y^k + \mathcal{R}(\mathcal{R}^*)^{-1}(z^k - x^k)$. Then, for all $k \in \mathbb{N}$, there exists $\epsilon_k \geq 0$ such that the iterations in Algorithm 1 can be reformulated as the following two-step iterations algorithm:*

$$\begin{cases} v^k \in [(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\,\partial g+(\mathcal{A}^*\circ\nabla f\circ\mathcal{A}),\,\mathcal{N}_{\mathcal{V}}}]^{[\epsilon_k]}(\overline{z}^k), & \text{(15a)} \\ \theta_k\|\mathcal{R}^*v^k\|^2_{\mathcal{R}^{-1}} + \|\mathcal{R}^*v^k + \overline{z}^k - z^k\|^2_{\mathcal{R}^{-1}} \\ \qquad +2\epsilon_k \leq \overline{\sigma}\|\overline{z}^k - z^k\|^2_{\mathcal{R}^{-1}}, & \text{(15b)} \end{cases}$$

and $z^{k+1} = z^k - (1+\theta_k)\mathcal{R}^*v^k$.

Remark 3 *Based on Proposition 3, the GSOS algorithm can be regarded as an inexact over-relaxed metric proximal point algorithm for the composite inclusion*

$$0 \in (\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_{\mathcal{V}}}(z).$$

By Proposition 3 and Lemma 2, we can establish the convergence of the GSOS algorithm based on the relationship between the two zero point sets $[\nabla f + \sum_{i=1}^n \partial g_i]^{-1}(0)$ and Ω.

Theorem 1 *Let $\{(x^k, y^k, z^k)\}$ be the sequence generated by Algorithm 1. We have:*

(i) *for any $z^* \in [(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_{\mathcal{V}}}]^{-1}(0)$, it holds that*

$$\|z^{k+1} - z^*\|^2_{\mathcal{R}^{-1}} \leq \|z^k - z^*\|^2_{\mathcal{R}^{-1}} - (1-\overline{\sigma})(1+\theta_k)\|x^k - y^k\|^2_{\mathcal{R}^{-1}}; \quad \text{(16)}$$

(ii) *z^k converges to a point belonging to zero point set $[(\mathcal{R}^*)^{-1}\mathcal{S}_{\mathcal{R},\partial g+\mathcal{A}^*\circ\nabla f\circ\mathcal{A},\,\mathcal{N}_{\mathcal{V}}}]^{-1}(0)$ and ω^k converges to a point belonging to $[\nabla f + \sum_{i=1}^n \partial g_i]^{-1}(0)$, i.e., the optimal solution set of problem (1).*

Theorem 1 indicates that $\|x^k - y^k\|$ approaching to zero implies the convergence of the GSOS algorithm. In the theorem below, we measure the convergence rates of two sequences $\|x^k - y^k\|$ and $\|\omega^k - \omega^{k+1}\|$.

Theorem 2 *Let z^k be the sequence generated by the GSOS algorithm. Then, there exists $i \in \{1, 2, \cdots, k\}$ such that*

$$\|x^i - y^i\|^2 \leq O(\frac{1}{k}),\ \|\omega^{i+1} - \omega^i\|^2 \leq O(\frac{1}{k}).$$

Due to the space limit, all proofs of the propositions, lemmas and theorems are placed into the supplementary material.

4. Experiments

In this section, we apply the proposed algorithm to the overlapping group Lasso (Zhao et al., 2009; Jacob et al., 2009; Mairal et al., 2010) and graph-guided fused Lasso problems (Chen et al., 2012; Kim & Xing, 2009), which can be formulated as

$$\min \frac{1}{2}\|\mathcal{S}x - b\|^2 + \sum_{i=1}^K g_i(x). \quad \text{(17)}$$

For overlapping group Lasso problem (21), $g_i(x) = \nu\alpha_i\|x_{\mathcal{G}_i}\|$ and K denotes the number of groups. For graph-guided Lasso problem (25), $g_i(x) = \nu\alpha_{ij}\|x_i - x_j\|$ and K denotes the number of edges in the graph edge set E.

We describe the detailed techniques in the experimental implementation for (17). Given $a > \frac{1}{2}$ and a positive definite operator $\mathcal{D}$ satisfying $\mathcal{D} \succeq S^TS$, we set

$$\mathcal{H}_{i,j} = \begin{cases} \frac{1}{K^2}\mathcal{D}, & i \geq j \in \{1, 2, \cdots, K\}; \\ \frac{a}{K^2}\mathcal{D}, & i = j \in \{1, 2, \cdots, K\}. \end{cases} \quad \text{(18)}$$

Hence, it easy to check that $\mathcal{H} + \mathcal{H}^* = \mathcal{A}^*\mathcal{D}\mathcal{A} + \frac{2a-1}{K^2}\text{Diag}(\mathcal{E}_{\mathcal{Y}}\mathcal{D}) \succ 0$. Due to the smooth term in overlapping group Lasso (21) is quadratic, the two estimations $\overline{\theta}_2$ and θ_k^{adap2} in (14b) and (14d) are preferred to be used. By specific $\mathcal{H}$, we obtain $\sum_{i,j=1}^K \mathcal{H}_{i,j} = \frac{K(K-1)+2\alpha K}{2K^2}\mathcal{D}$ and $\sum_{j=1}^K\sum_{i=j}^K \mathcal{H}_{i,j}z_j^k = \frac{\mathcal{D}}{K^2}\sum_{j=1}^K(a+K-j)z_j^k$, which further imply $x^k = \big(\sum_{i,j=1}^K \mathcal{H}_{i,j}\big)^{-1}\sum_{j=1}^K\sum_{i=j}^K \mathcal{H}_{i,j}z_j^k = \frac{2\sum_{j=1}^K(a+K-j)z_j^k}{K(K-1)+2aK}$. Moreover, by the positive definiteness of $\mathcal{H}_{i,i}$ and $\mathcal{D}$, it holds that $\sum_{j=1}^n\sum_{i=i}^j \mathcal{H}_{j,i}^*z_j^k = \frac{\mathcal{D}}{K^2}\sum_{j=1}^K(a+j-1)z_j^k$. Hence, we attain $\omega^k = \frac{2\sum_{j=1}^K(a+j-1)z_j^k}{K(K-1)+2aK}$. In addition, by the definition of $\mathcal{H}$, we reformulate the estimation (14b) for θ_k as the following form:

$$\overline{\theta} = \max\Big\{\theta \mid \mathcal{E}_{\mathcal{Y}}\big[(\overline{\sigma}-\theta)\mathcal{D} - S^TS\big]\mathcal{E}_{\mathcal{Y}}^* + (2a-1)\big(\overline{\sigma}-\theta\big)\text{Diag}(\mathcal{E}_{\mathcal{Y}}\mathcal{D}) \succeq 0\Big\}.$$

Due to $a \geq \frac{1}{2}$ and the positive definiteness of $\mathcal{D}$, a sufficient condition satisfying the constraint in the above set is $\left\{(\overline{\sigma}-\theta)\mathcal{D} - S^TS \succeq 0, \theta \leq \overline{\sigma}\right\}$. Hence, we have an alternative estimation for $\overline{\theta}$ as

$$\overline{\theta} = \max\left\{\theta \mid (\overline{\sigma}-\theta)\mathcal{D} - S^TS \succeq 0, \theta \leq \overline{\sigma}\right\}. \quad (19)$$

Similarly, the adaptive stepsize estimation (14d) is reformulated as

$$\theta_k^{\text{adap}} = \overline{\sigma} - \frac{\frac{1}{2K^2}\left\|\sum_{i=1}^{K}(x^k - y_i^k)\right\|^2_{S^TS}}{\sum_{j=1}^{K}\sum_{i=j}^{K}(x^k - y_i^k)^T\mathcal{H}_{ij}(x^k - y_j^k)}. \quad (20)$$

Therefore, the GSOS algorithm can be specified as the following form for solving problem (17).

Algorithm 2 GSOS Algorithm for Solving Problem (17)

Parameters: Choose $\overline{\sigma} \in (0,1)$, positive definite operators $\mathcal{D}$ and $\mathcal{H}_{i,j}$ satisfying (18), and a starting point $z^0 \in \mathcal{Z}$. Set $\overline{\theta}$ as (19) and $\theta^{\text{fix}} \in \left(-1, \overline{\theta}_1\right]$.
for $k = 0, 1, 2, \cdots$, **do**
  $x^k := \frac{2\sum_{j=1}^{K}(\alpha+K-j)z_j^k}{K(K-1)+2\alpha K}$;
  for $i = 1, 2\cdots, K$ **do**
    $y_i^k := \text{Prox}_{\mathcal{H}_{i,i}^{-1}g_i}\big(\mathcal{H}_{i,i}^{-1}[\sum_{j=1}^{i}\mathcal{H}_{i,j}(2x^k - z_j^k) - \frac{1}{K}S^T(Sx^k - b) - \sum_{j=1}^{i-1}\mathcal{H}_{i,j}y_j^k]\big)$;
  end for
  set $\theta_k \in [\theta_k^{\text{fix}}, \theta_k^{\text{adap}}]$, where θ_k^{adap} is defined via (20);
  for $j = 1, 2\cdots, K$ **do**
    $z_j^{k+1} := z_j^k + (1+\theta_k)(y_j^k - x^k)$;
  end for
end for
return $\omega^N = \frac{2\sum_{j=1}^{K}(\alpha+j-1)z_j^N}{K(K-1)+2\alpha K}$.

In this paper, we compare the proposed GSOS algorithm with four state-of-the-art algorithms below.

- GFB (Raguet et al., 2013): **G**eneralized **F**orward **B**ackward (GFB) splitting algorithm is a primal first-order operator splitting algorithm for solving (1) proposed by Raguet *et al.* (Raguet et al., 2013), which has been shown to outperform other competing algorithms such as (Monteiro & Svaiter, 2013; Combettes & Pesquet, 2012; Chambolle & Pock, 2011) for variational image restoration.
- PDM (Condat, 2013): A first-order **P**rimal-**D**ual splitting **M**ethod (PDM) (Condat, 2013) for solving jointly the primal and dual formulations of large-scale convex minimization problems involving Lipschitz, proximal and linear composite terms.
- PA-APG (Yu, 2013): **P**roximal **A**verage approximated **A**ccelerated **P**roximal **G**radient (PA-APG) algorithm (Yu, 2013) is a primal first-order method, which utilizes the proximal average technique (Bauschke et al., 2008) to separate the multi-term nonsmooth function in (1). It has been shown to outperform the smoothing accelerated proximal gradient method (Nesterov, 2005b;a).
- APA-APG (Shen et al., 2017): An enhanced version of PA-APG, which incorporates the **A**daptive **P**roximal **A**verage approximation technique with the **A**ccelerated **P**roximal **G**radient (APA-APG) method to improve the efficiency of the optimization procedure.

It is worthwhile to emphasize that PA-APG and APA-APG algorithms can only be applied to a specific class of problems (1), in which the multi-term nonsmooth regularization is Lipschitz continuous. Since the nonsmooth regularization terms in overlapping group Lasso and graph-guided fused Lasso are all exactly Lipschitz continuous, the two efficient solvers PG-APG (Yu, 2013) and its enhanced version APA-APG (Shen et al., 2017) are also compared with the GSOS algorithm to illustrate the efficacy of GSOS. In the implementation, the approximation parameter for PA-APG is set as $1.0e-5$.

4.1. Overlapping Group Lasso

In this subsection, we apply the proposed GSOS algorithm to the overlapping group Lasso problem, which takes the following formal definition:

$$\min \frac{1}{2}\|Sx - b\|^2 + \nu\sum_{i=1}^{K}\alpha_i\|x_{\mathcal{G}_i}\|, \quad (21)$$

where $S \in \mathbb{R}^{n\times d}$ is the sampling matrix, b is the noisy observation vector, $\mathcal{G} = \{\mathcal{G}_1, \cdots, \mathcal{G}_K\}$ denotes the set of overlapping groups ($\mathcal{G}_i \subset \{1, \cdots, d\}$ satisfying $\bigcup_{i=1}^{K}\mathcal{G}_i = \{1, \cdots, d\}$ and $\mathcal{G}_i \bigcap \mathcal{G}_j \neq \emptyset$ for some i, j), $x_{\mathcal{G}_i} \in \mathbb{R}^d$ is a duplication of x with $x_{\{1,\cdots,d\}\setminus\mathcal{G}_i} = 0$, α_i is the weight for the i-th group, and ν is the regularization parameter controlling group sparsity.

During the implementation of Algorithm 2, we need to calculate the generalized proximal mapping of $\|x_{\mathcal{G}_i}\|$ in the updating step of y_i^k. By the positive definiteness of $\mathcal{H}_{i,i}$, the calculation of y_i^k in Algorithm 2 is equivalent to solving the following problem:

$$y_i^k := \arg\min_x \frac{1}{2}\|x - b^k\|^2_{\mathcal{H}_{i,i}} + \nu\alpha_i\|x_{\mathcal{G}_i}\|,$$

where $b^k = \mathcal{H}_{i,i}^{-1}\big[\sum_{j=1}^{i}\mathcal{H}_{i,j}(2x^k - z_j^k) - \frac{1}{K}S^T(Sx^k - b) - \sum_{j=1}^{i-1}\mathcal{H}_{i,j}y_j^k\big]$. In the proposition below, given c, diag-

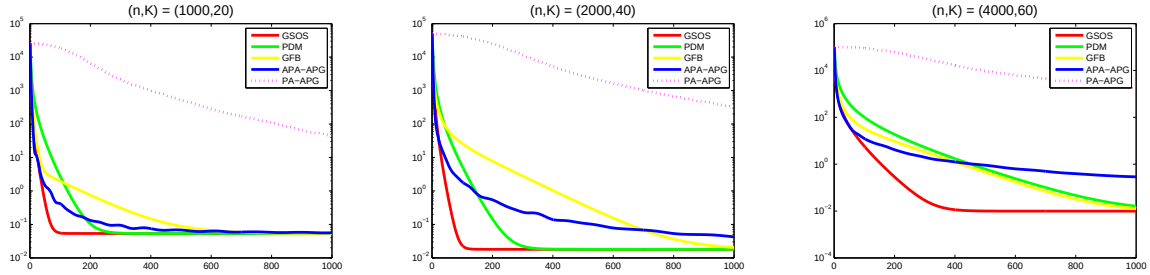

Figure 1. Objective value vs. iteration on overlapping group Lasso.

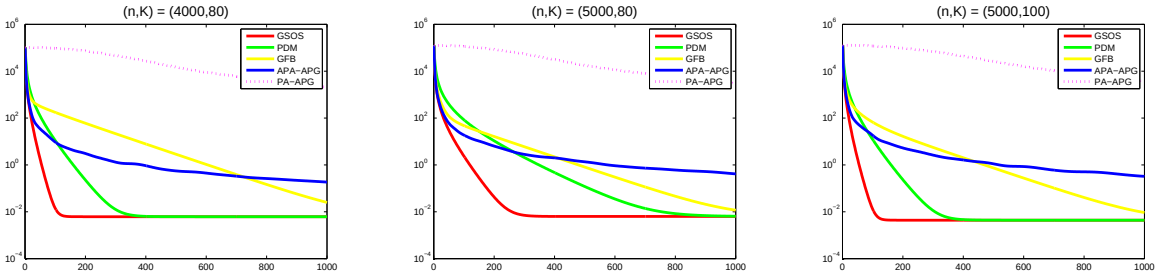

Figure 2. Objective value vs. iteration on overlapping group Lasso.

onal positive definite operatpr $\mathcal{H}_{i,i}$ and group $\mathcal{G}$, we solve

$$x^* := \arg\min_x \frac{1}{2}\|x-c\|^2_{\mathcal{H}_{i,i}} + \nu\|x_{\mathcal{G}}\|. \tag{22}$$

When $\mathcal{H}_{i,i}$ is identity matrix $\mathcal{I}$, (22) has the closed-form solution

$$x^* = \begin{cases} x^*_{\mathcal{G}}, & i \in \mathcal{G}, \\ c_i, & \text{else}, \end{cases}$$

where $x^*_{\mathcal{G}} = \begin{cases} (1-\nu/\|c_{\mathcal{G}}\|)c_{\mathcal{G}}, & \|c_{\mathcal{G}}\| \geq t; \\ 0, & \text{else}. \end{cases}$

Proposition 4 *Let $(\mathcal{H}_{i,i})_{\mathcal{G}}$ be the subdiagonal matrix of $\mathcal{H}_{i,i}$ with the index set $\mathcal{G}$, and t^* be the optimal solution of the one-dimensional optimization problem*

$$\min_{t\geq 0}\left\{\frac{1}{2}\left\langle c_{\mathcal{G}}, \left[(\mathcal{H}_{i,i})^{-1}_{\mathcal{G}} + 2t\mathcal{I}\right]^{-1} c_{\mathcal{G}}\right\rangle + t\nu^2\right\}. \tag{23}$$

Hence, the optimal solution of (22) has the following form

$$x^* = \begin{cases} c_{\mathcal{G}} - \left[\mathcal{I} + 2t^*(\mathcal{H}_{i,i})_{\mathcal{G}}\right]^{-1} c_{\mathcal{G}}, & i \in \mathcal{G}; \\ c_i, & \textit{else}. \end{cases} \tag{24}$$

Like (Chen et al., 2012; Yu, 2013), the entries of sampling matrix $S \in \mathbb{R}^{n\times d}$ are sampled from an $i.i.d.$ normal distribution, and $x \in \mathbb{R}^d$ with $x_j = (-1)^j \exp^{-(j-1)/100}$ and $d = 90K + 10$. Let ξ be the noise sampled from the standard normal distribution, and the noisy observation satisfies $b = Sx + \xi$. In addition, we set $\nu = 1$ and $\alpha_i = \frac{1}{K^2}$ for each group $\mathcal{G}_i$ and the groups $\{\mathcal{G}_i\}$ are overlapped by 10 elements, that is

$$\left\{\begin{array}{cc} \mathcal{G}_1 = \{1,\cdots,100\} & \mathcal{G}_2 = \{91,\cdots,190\} \\ \cdots & \mathcal{G}_K = \{d-99,\cdots,d\} \end{array}\right\}.$$

The sampling size and the number of groups (n, K) are chosen from the following set

$$(n,K) \in \left\{\begin{array}{ccc} (1000,20), & (2000,40), & (4000,60), \\ (4000,80), & (5000,80), & (5000,100) \end{array}\right\}.$$

To further reduce the computations, in Algorithm 2 we set $\mathcal{H}_{i,i} = \|S^TS\|\mathcal{I}$ and the over-relaxation stepsize θ_k as $\overline{\theta}$ in (19). Hence, the compared five solvers GSOS, GFB, PDM, PA-APG and APA-APG have the same computational cost in each iteration. To be fair, all the compared algorithms start with the same initial point. The following six pictures in Figures 1 and 2 display the comparisons of the five solvers for a variety of (n, K). It is apparent that our proposed GSOS algorithm shows great superiorities over the other four solvers. The primal-dual solver PDM is slightly faster than the primal solver GFB. PA-APG is the slowest algorithm, because the prespecified proximal average approximation precision is $1.0e-5$ which leads to a very small stepsize. Also, APA-APG is much faster than the other four solvers at the first 50 iterations. However, it is slowed down since the stepsize used in AP-APG becomes smaller and smaller as the iterations go on.

4.2. Graph-Guided Fused Lasso

In this subsection, we perform experiments on graph-guided fused Lasso which is formulated as

$$\min \frac{1}{2}\|Sx-b\|^2 + \nu \sum_{(i,j)\in E} \alpha_{ij}|x_i - x_j|, \tag{25}$$

where $\alpha_{ij} \geq 0$ is the weight for the fused term $\|x_i - x_j\|$ for all $(i, j) \in E$ (E is the given graph edge set), and ν is

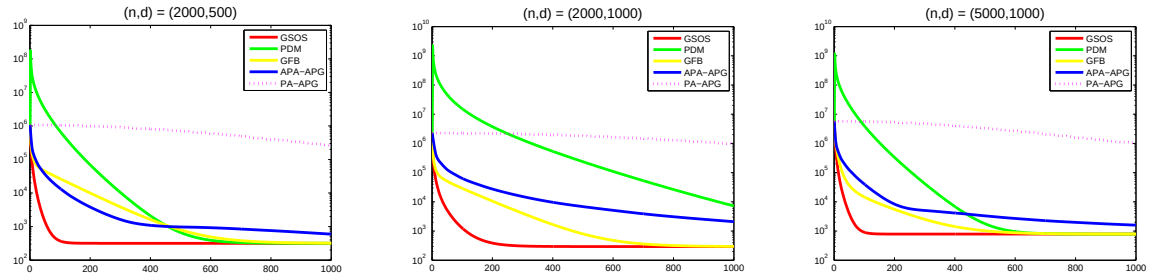

Figure 3. Objective value vs. iteration on graph-guided fused Lasso.

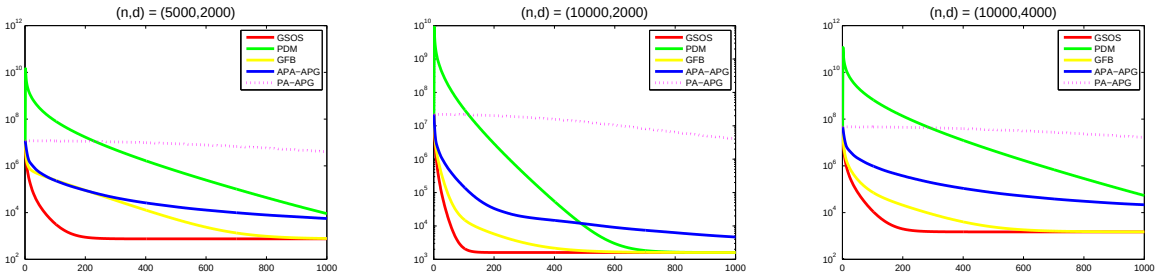

Figure 4. Objective value vs. iteration on graph-guided fused Lasso.

the regularization parameter.

In the implementation of Algorithm 2 for tackling graph-guided fused Lasso (25), we need to solve the following optimization in the updating step of y^k:

$$x^* := \arg\min_x \frac{1}{2}\|x-b\|^2_{\mathcal{H}_{i,i}} + \nu|x_i - x_j|, \qquad (26)$$

where $\mathcal{H}_{i,i}$ is a diagonal positive definite matrix, and b and ν are given constants. Let h_{ii} and h_{jj} be the i-th and j-th diagonal elements of $\mathcal{H}_{i,i}$, respectively.

Proposition 5 *The optimal solution of* (26) *takes the following closed-form:*

$$x^* = \begin{cases} b_l - h_{ll}^{-1}\lambda^*, & l = i, \\ b_l + h_{ll}^{-1}\lambda^*, & l = j, \\ b_l, & l \neq i, j, \end{cases} \qquad (27)$$

where λ^* is defined as

$$\lambda^* = \begin{cases} \frac{b_i - b_j}{h_{ii}^{-1}+h_{jj}^{-1}}, & \left|\frac{b_i-b_j}{h_{ii}^{-1}+h_{jj}^{-1}}\right| \leq \nu; \\ \mathrm{sign}\,(b_i - b_j)\,\nu, & \left|\frac{b_i-b_j}{h_{ii}^{-1}+h_{jj}^{-1}}\right| > \nu. \end{cases}$$

In the implementation, we use the similar parameter settings of S, ν as above. The dimension parameter pair (n, d) is chosen from the following set

$$(n,d) \in \left\{ \begin{array}{c} (2000,500),(2000,1000),(5000,1000), \\ (5000,2000),(10000,2000),(10000,4000) \end{array} \right\},$$

and the parameter $\alpha_i = 100/|E|^2$. Similarly, all the compared algorithms start with the same initial point. The following six pictures in Figures 3 and 4 display the comparisons of the five solvers for six kinds of choices of (n, d). It is obvious that the other four solvers GFB, PDM, AP-APG and APA-APG are not as efficient as the proposed GSOS algorithm, which demonstrates that the Gauss-Seidel technique is very useful for addressing nonsmooth optimization. It is worthwhile to point out that the primal solver GFB is faster than the primal-dual solver PDM on graph-guided fused Lasso. One possible reason is that the number of nonsmooth terms is too large, which will lead to a large quantity of dual variables introduced in PDM and hence slow down the updating of primal variables.

5. Conclusions

In this paper, we proposed a novel first-order algorithm called GSOS for addressing multi-term nonsmooth convex composite optimization. This algorithm inherits the advantages of the Gauss-Seidel technique and the operator splitting technique, therefore being largely accelerated. We found that the GSOS algorithm includes the generalized forward backward splitting method (Raguet et al., 2013) as a special case. In addition, we developed a new technique to establish the global convergence and iteration complexity of the GSOS algorithm. Last, we applied the proposed GSOS algorithm to solve overlapping group Lasso and graph-guided fused Lasso problems, and compared it against several state-of-the-art algorithms. The experimental results show the great superiority of the GSOS algorithm in terms of both efficiency and effectiveness.

Acknowledgements

Yuan is supported by NSF-China (61402182).

References

Bach, F. R. Structured sparsity-inducing norms through submodular functions. *Advances in Neural Information Processing Systems*, pp. 118–126, 2010.

Bach, F. R., Jenatton, R., Mairal, J., and Obozinski, G. Structured sparsity through convex optimization. *Statistical Science*, 27(4):pgs. 450–468, 2012.

Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Space*. Springer New York, 2011.

Bauschke, H. H., Goebel, R., Lucet, Y., and Wang, X. The proximal average: basic theory. *SIAM Journal on Optimization*, 19(2):766–785, 2008.

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

Briceno-Arias, L. M., Combettes, P. L., Pesquet, J. C., and Pustelnik, N. Proximal algorithms for multicomponent image recovery problems. *Journal of Mathematical Imaging and Vision*, 41(1-2):3–22, 2011.

Burachik, R. S. and Svaiter, B. F. ε-enlargements of maximal monotone operators in banach spaces. *Set-Valued Analysis*, 7(2):117–132, 1999.

Burachik, R. S., Iusem, A. N., and Svaiter, B. F. Enlargement of monotone operators with applications to variational inequalities. *Set-Valued and Variational Analysis*, 5(2):159–180, 1997.

Burachik, R. S., Sagastizábal, C. A., and Svaiter, B. F. ε-enlargements of maximal monotone operators: Theory and applications. In *Reformulation: nonsmooth, piecewise smooth, semismooth and smoothing methods*, pp. 25–43. Springer, 1998.

Chambolle, A. and Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1): 120–145, 2011.

Chen, D. and Plemmons, R. J. *Nonnegativity constraints in numerical analysis*. 2015.

Chen, G. H. G. and Rockafellar, R. T. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

Chen, X., Lin, Q., Kim, S., Carbonell, J. G., and Xing, E. P. An efficient proximal gradient method for general structured sparse learning. *stat*, 1050:26, 2011.

Chen, X., Lin, Q., Kim, S., Carbonell, J. G., and Xing, E. P. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752, 2012.

Combettes, P. L. and Pesquet, J. C. A douglas–rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):564–574, 2007.

Combettes, P. L and Pesquet, J. C. A proximal decomposition method for solving convex variational inverse problems. *Inverse problems*, 24(6):065014, 2008.

Combettes, P. L. and Pesquet, J. C. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pp. 185–212. Springer, 2011.

Combettes, P. L. and Pesquet, J. C. Primal-dual splitting algorithm for solving inclusions with mixtures of composite, lipschitzian, and parallel-sum type monotone operators. *Set-Valued and variational analysis*, 20(2):307–330, 2012.

Condat, L. A primal-dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.

Cui, Y., Li, X., Sun, D. F., and Toh, K. C. On the convergence properties of a majorized alternating direction method of multipliers for linearly constrained convex optimization problems with coupled objective functions. *Journal of Optimization Theory & Applications*, 169(3): 1013–1041, 2016.

Davis, D. and Yin, W. A three-operator splitting scheme and its optimization applications. *Mathematics*, 19(3): 407–12, 2015.

Dupé, F. X., Fadili, J. M., and Starck, J. L. A proximal iteration for deconvolving poisson noisy images using sparse representations. *IEEE Transactions on Image Processing*, 18(2):310–321, 2009.

Eckstein, J. and Bertsekas, D. P. On the douglasłrachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992.

Huang, J., Zhang, T., and Metaxas, D. Learning with structured sparsity. *Journal of Machine Learning Research*, 12(Nov):3371–3412, 2011.

Jacob, L., Obozinski, G., and Vert, J. P. Group lasso with overlap and graph lasso. In *International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June*, pp. 433–440, 2009.

Kim, S. and Xing, E. P. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet*, 5(8):e1000587, 2009.

Li, Q. and Zhang, N. Fast proximity-gradient algorithms for structured convex optimization problems . *Applied & Computational Harmonic Analysis*, 41(2):491–517, 2016.

Lin, Z., Liu, R., and Su, Z. Linearized alternating direction method with adaptive penalty for low-rank representation. *Advances in Neural Information Processing Systems*, pp. 612–620, 2011.

Luo, Z. Q. and Tseng, P. On the convergence of a matrix splitting algorithm for the symmetric monotone linear complementarity problem. *SIAM Journal on Control and Optimization*, 29(5):1037–1060, 1991.

Mairal, J., Jenatton, R., Bach, F. R., and Obozinski, G. R. Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems*, pp. 1558–1566, 2010.

Monteiro, R. D. C. and Svaiter, B. F. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, 23(1):475–507, 2013.

Nesterov, Y. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, 16 (1):235–249, 2005a.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005b.

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(2007076): 125–161, 2007.

Pustelnik, N., Chaux, C., and Pesquet, J. C. Parallel proximal algorithm for image restoration using hybrid regularization. *IEEE Transactions on Image Processing*, 20 (9):2450–2462, 2011.

Raguet, H., Fadili, J., and Peyré, G. A generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, 6(3):1199–1226, 2013.

Richard, E., Savalle, P. A., and Vayatis, N. Estimation of simultaneously sparse and low rank matrices. *arXiv preprint arXiv:1206.6474*, 2012.

Shen, L., Liu, W., Huang, J., Jiang, Y. G., and Ma, S. Adaptive proximal average approximation for composite convex minimization. In *AAAI*, 2017.

Svaiter, B. F. A family of enlargements of maximal monotone operators. *Set-Valued Analysis*, 8(4):311–328, 2000.

Svaiter, B. F. A class of fejér convergent algorithms, approximate resolvents and the hybrid proximal-extragradient method. *Journal of Optimization Theory and Applications*, 162(1):133–153, 2014.

Teo, C. H., Smola, A., Vishwanathan, S., and Le, Q. V. A scalable modular convex solver for regularized risk minimization. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 727–736. ACM, 2007.

Teo, C. H., Vishwanthan, S., Smola, A. J., and Le, Q. V. Bundle methods for regularized risk minimization. *Journal of Machine Learning Research*, 11(Jan):311–365, 2010.

Vũ, B. C. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

Xu, Y. and Yin, W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *Siam Journal on Imaging Sciences*, 6(3):1758–1789, 2013.

Yu, Y. Better approximation and faster algorithm using the proximal average. *Advances in Neural Information Processing Systems*, pp. 458–466, 2013.

Yuan, G., Zheng, W. S., and Ghanem, B. A matrix splitting method for composite function minimization. *arXiv preprint arXiv:1612.02317*, 2016.

Zhao, P., Rocha, G., and Yu, B. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pp. 3468–3497, 2009.

Zhong, W. and Kwok, J. T. Y. Accelerated stochastic gradient method for composite regularization. In *AISTATS*, pp. 1086–1094, 2014.

Zhou, K., Zha, H., and Song, L. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *AISTATS*, volume 31, pp. 641–649, 2013.