PLOS ONE

RESEARCH ARTICLE

Applications of Spectral Gradient Algorithm for Solving Matrix $\ell_{2,1}$-Norm Minimization Problems in Machine Learning

Yunhai Xiao[1]*, Qiuyu Wang[2], Lihong Liu[3]

1 Institute of Applied Mathematics, School of Mathematics and Statistics, Henan University, Kaifeng, Henan Province, China, **2** School of Mathematics and Statistics, Henan University, Kaifeng, Henan Province, China, **3** School of Mathematics and Statistics, Henan University, Kaifeng, Henan Province, China

* yhxiao@henu.edu.cn

Citation: Xiao Y, Wang Q, Liu L (2016) Applications of Spectral Gradient Algorithm for Solving Matrix $\ell_{2,1}$-Norm Minimization Problems in Machine Learning. PLoS ONE 11(11): e0166169. doi:10.1371/journal.pone.0166169

Editor: Xiaolei Ma, Beihang University, CHINA

Received: June 19, 2016

Accepted: October 23, 2016

Published: November 18, 2016

Data Availability Statement: All relevant data is contained within the manuscript.

Funding: The research is supported by the Major State Basic Research Development Program of China (973 Program) (Grant No. 2015CB856003), the National Natural Science Foundation of China (Grant No. 11471101), and the Program for Science and Technology Innovation Talents in Universities of Henan Province (Grant No. 13HASTIT050).

Competing Interests: The authors have declared that no competing interests exist.

Abstract

The main purpose of this study is to propose, then analyze, and later test a spectral gradient algorithm for solving a convex minimization problem. The considered problem covers the matrix $\ell_{2,1}$-norm regularized least squares which is widely used in multi-task learning for capturing the joint feature among each task. To solve the problem, we firstly minimize a quadratic approximated model of the objective function to derive a search direction at current iteration. We show that this direction descends automatically and reduces to the original spectral gradient direction if the regularized term is removed. Secondly, we incorporate a nonmonotone line search along this direction to improve the algorithm's numerical performance. Furthermore, we show that the proposed algorithm converges to a critical point under some mild conditions. The attractive feature of the proposed algorithm is that it is easily performable and only requires the gradient of the smooth function and the objective function's values at each and every step. Finally, we operate some experiments on synthetic data, which verifies that the proposed algorithm works quite well and performs better than the compared ones.

1 Introduction

The tasks in medical diagnosis [1], text classification [2–5], biomedical informatics [6, 7] and other applications [8–12] are always related to each other. Hence, capturing the shared information among each task becomes the key issue to learn [13–15]. Given the training set of t tasks $A = [A_1; \ldots; A_t] \in \mathbb{R}^{m\times n}$ and $b = [b_1; \ldots; b_t]^\top \in \mathbb{R}^m$, where A_j is the data for the j-th task and b_j is the corresponding response. We let $x_j \in \mathbb{R}^n$ be the sparse feature for the j-th task, and let $X = [x_1, \ldots, x_t] \in \mathbb{R}^{n\times t}$ be the joint feature to be learned. In order to select features globally, it encourages several rows of X to be zeros and solves the following $\ell_{2,1}$-norm regularized least squares [16, 17]

$$\min_{X\in\mathbb{R}^{n\times t}} \frac{1}{2} \| AX - b \|_2^2 + \mu \| X \|_{2,1}, \tag{1}$$

where $\mu > 0$ is a weighting parameter, and $\|X\|_{2,1}$ is defined by the sum of the ℓ_2-norm of each row of a matrix. It is well known that the $\ell_{2,1}$-norm is used to encourage the multiple predictions from different tasks to share similar parameter sparsity patterns.

In the past few years, several algorithms have been proposed, analyzed, and tested to solve the nonsmooth convex minimization Problem (1). The algorithm in [18] transformed Eq (1) equivalently into a smooth convex optimization problem and minimized consequently by Nesterov's gradient method. The method in [16] reformulated Eq (1) as a constrained optimization problem and minimized alternately. The algorithm in [19] and its variant [20] reformulated the problem as an equivalent constrained minimization by introducing an auxiliary variable, and then minimized the corresponding augmented Lagrange function alternatively. Finally, for another accelerated proximal gradient version of the algorithm [19], one can refer to [21].

Unlike all the research activities which mainly concerned about Problem (1), in this paper, we focus on the following generalized nonsmooth convex optimization problem

$$\min_{X\in\mathbb{R}^{n\times t}} F(X) + \mu\| X \|_{2,1}, \tag{2}$$

where $F : \mathbb{R}^{n\times t} \to \mathbb{R}$ is continuously differentiable (may be non-convex) and bounded below. Clearly, Model (2) includes Eq (1) as a special case when F is a least square. As we all know, the spectral gradient method was originated by Barzilai and Borwein [22] for solving smooth unconstrained minimization problems, later was developed in [23–26], and then was extended to solve ℓ_1-regularized nonsmooth minimization [27]. However, its numerical performance in solving matrix $\ell_{2,1}$-norm involved nonsmooth minimization problems is still undiscovered. Therefore, extending the spectral gradient algorithm to solve Problem (2) may have significance both in theory and practice. The first contribution of this study lies in the design of the search direction at each iteration, which is derived by minimizing a quadratic approximated model of the objective function and at the same time making full use of the special structure of the $\ell_{2,1}$-norm. We also show that the generated direction descends automatically provided that the spectral coefficient is positive. The second contribution of the paper is the nonmonotone line search, which is used to improve the algorithm's performance. At each iteration, the algorithm requires the gradient of the smooth term and the value of the objective function, which means it has the ability to solve high dimensional problems. Finally, we do performance comparisons with a couple of solvers IAMD_MFL and SLEP, which illustrate that the proposed method is fast, efficient, and competitive.

The paper is organized as follows. In Section 2, we provide some notations and preliminaries, and construct the new algorithm together with its properties. In Section 3, we establish the global convergence of the algorithm. In Section 4, we report some numerical results and do some performance comparisons. Finally, we conclude our paper in Section 5.

2 Algorithm

2.1 Notations and preliminaries

In the first place, we summarize the notations used in this paper. Matrices are written as uppercase letters. Vectors are described as lowercase letters. For the matrix X, its i-th row and j-th column are denoted by $X_{i,:}$ and $X_{:,j}$ respectively. The Frobenius norm and the $\ell_{2,1}$-norm of the matrix $X \in \mathbb{R}^{n\times t}$ are defined as, respectively,

$$\| X \|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{t} X_{i,j}^2}, \quad \text{and} \quad \| X \|_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{t} X_{i,j}^2} = \sum_{i=1}^{n}\| X_{i,:} \|_2.$$

For any two matrices $X, Y \in \mathbb{R}^{n\times t}$, we define $\langle X, Y\rangle = \mathrm{tr}(X^\top Y)$ (the standard trace inner product in $\mathbb{R}^t$), so that $\| X \|_F = \sqrt{\langle X, X\rangle}$. If $x \in \mathbb{R}^d$, we denote "Diag(x)" the diagonal matrix possessing the components of vector x on the diagonal. We define "$\top$" as the transpose of a vector or a matrix. For the sake of simplicity, we let $\Phi(X) = F(X) + \mu\|X\|_{2,1}$. Additional notations will be introduced when they occur.

We now quickly review the spectral gradient method for the unconstrained smooth minimization problem

$$\min f(x), \quad x \in \mathbb{R}^n,$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function. The spectral gradient method is defined by

$$x_{k+1} = x_k - \lambda_k^{-1}\nabla f(x_k),$$

where one of the choices of λ_k (named as spectral coefficient) is given by

$$\lambda_k = \frac{s_{k-1}^\top y_{k-1}}{\| s_{k-1} \|_2^2},$$

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = \nabla f(x_k) - \nabla f(x_{k-1})$. Obviously, if $s_{k-1}^\top y_{k-1} > 0$, i.e. $\lambda_k > 0$, the search direction $d_k := -\lambda_k^{-1}\nabla f(x_k)$ descends automatically at current point.

2.2 Algorithm

Now, we turn our attention to the original Model (2). Since the $\ell_{2,1}$-norm is nodifferentiable, we approximate the objective function by the following quadratic function Q_k:

$$\begin{aligned} Q_k(D): &= F(X_k + D) + \mu\| X_k + D \|_{2,1} \\ &\approx F(X_k) + \langle \nabla F(X_k), D\rangle + \frac{\Lambda_k}{2}\| D \|_F^2 + \mu\| X_k + D \|_{2,1}, \end{aligned} \tag{3}$$

where $\nabla F(X_k) \in \mathbb{R}^{n\times t}$ is the gradient of F at X_k; Λ_k is the so-called spectral coefficient which defined by

$$\Lambda_k = \frac{\langle S_{k-1}, Y_{k-1}\rangle}{\| S_{k-1} \|_F^2}, \tag{4}$$

where $S_{k-1} = X_k - X_{k-1}$ and $Y_{k-1} = \nabla F(X_k) - \nabla F(X_{k-1})$. Minimizing Eq (3) yields

$$\begin{aligned} &\arg\min_{D\in\mathbb{R}^{n\times t}} Q_k(D) \\ &= \arg\min_{D\in\mathbb{R}^{n\times t}} \langle \nabla F(X_k), D\rangle + \frac{\Lambda_k}{2}\| D \|_F^2 + \mu\| X_k + D \|_{2,1} \\ &= \arg\min_{D\in\mathbb{R}^{n\times t}} \frac{1}{\Lambda_k}\left(\langle \nabla F(X_k), D\rangle + \frac{\Lambda_k}{2}\| D \|_F^2 + \mu\| X_k + D \|_{2,1}\right) \\ &= \arg\min_{D\in\mathbb{R}^{n\times t}} \frac{1}{2}\| X_k + D - \left(X_k - \frac{1}{\Lambda_k}\nabla F(X_k)\right) \|_F^2 + \frac{\mu}{\Lambda_k}\| X_k + D \|_{2,1}. \end{aligned}$$

Denote $M_k = X_k + D$ and $N_k = X_k - \frac{1}{\Lambda_k}\nabla f(X_k)$. One can get

$$\arg\min_{D\in\mathbb{R}^{n\times t}} Q_k(D) = \arg\min_{D\in\mathbb{R}^{n\times t}} \sum_{i=1}^{n}\left(\frac{1}{2}\| (M_k)_{i,:} - (N_k)_{i,:} \|_2^2 + \frac{\mu}{\Lambda_k}\| (M_k)_{i,:} \|_2\right). \tag{5}$$

The favorable structure of Eq (5) make the i-th row of matrix M_k write explicitly as

$$(M_k)_{i,:} = \max\left\{\| (N_k)_{i,} \|_2 - \frac{\mu}{\Lambda_k}, 0\right\} \frac{(N_k)_{i,:}}{\| (N_k)_{i,:} \|_2},$$

where the convention $0 \cdot 0/0 = 0$ is followed. Hence, the search direction at current point can be expressed as

$$\begin{aligned}(D_k)_{i,:} &= -[(X_k)_{i,:} - (M_k)_{i,:}] \\ &= -\left[(X_k)_{i,:} - \max\left\{\| (N_k)_{i,} \|_2 - \frac{\mu}{\Lambda_k}, 0\right\} \frac{(N_k)_{i,:}}{\| (N_k)_{i,:} \|_2}\right] \\ &= -\left[(X_k)_{i,:} - \max\left\{\| \left(X_k - \frac{1}{\Lambda_k}\nabla F(X_k)\right)_{i,} \|_2 - \frac{\mu}{\Lambda_k}, 0\right\} \frac{\left(X_k - \frac{1}{\Lambda_k}\nabla F(X_k)\right)_{i,:}}{\| \left(X_k - \frac{1}{\Lambda_k}\nabla F(X_k)\right)_{i,:} \|_2}\right].\end{aligned} \tag{6}$$

Obviously, the Eq (6) reduces to $D_k = -\Lambda_k^{-1}\nabla F(x_k)$ at the case of $\mu = 0$, which means Eq (6) covers the traditional spectral gradient direction as a special case.

The following lemma verifies that D_k is a descent direction when the optimal solution is not achieved.

Theorem 1 Suppose that $\Lambda_k > 0$ and D_k is determined by Eq (6). Then

$$\Phi(X_k + \theta D_k) \le \Phi(X_k) + \theta[\langle \nabla F(X_k), D_k\rangle + \mu\| X_k + D_k \|_{2,1} - \mu\| X_k \|_{2,1}] + o(\theta), \quad \theta \in (0, 1], \tag{7}$$

and

$$\langle \nabla F(X_k), D_k\rangle + \mu\| X_k + D_k \|_{2,1} - \mu\| X_k \|_{2,1} \le -\frac{\Lambda_k}{2}\| D_k \|_F^2 . \tag{8}$$

Proof. By the differentiability of F and the convexity of $\|X\|_{2,1}$, we have that for any $\theta \in (0, 1]$,

$$\begin{aligned}&\Phi(X_k + \theta D_k) - \Phi(X_k) \\ &= F(X_k + \theta D_k) + \mu\| X_k + \theta D_k \|_{2,1} - F(X_k) - \mu\| X_k \|_{2,1} \\ &= F(X_k + \theta D_k) - F(X_k) + \mu\| \theta(X_k + D_k) + (1-\theta)X_k \|_{2,1} - \mu\| X_k \|_{2,1} \\ &\le F(X_k + \theta D_k) - F(X_k) + \theta\mu\| X_k + D_k \|_{2,1} + (1-\theta)\mu\| X_k \|_{2,1} - \mu\| X_k \|_{2,1} \\ &= \theta\langle \nabla F(X_k), D_k\rangle + o(\theta) + \theta[\mu\| X_k + D_k \|_{2,1} - \mu\| X_k \|_{2,1}],\end{aligned}$$

which is exactly Eq (7). Noting that D_k is the minimizer of Eq (3) and $\theta \in (0, 1]$, by Eq (3) and the convexity of $\|X\|_{2,1}$, one can get

$$\begin{aligned}&\langle \nabla F(X_k), D_k\rangle + \frac{\Lambda_k}{2}\| D_k \|_F^2 + \mu\| X_k + D_k \|_{2,1} - \mu\| X_k \|_{2,1} \\ &\le \langle \nabla F(X_k), \theta D_k\rangle + \frac{\Lambda_k}{2}\| \theta D_k \|_F^2 + \mu\| X_k + \theta D_k \|_{2,1} - \mu\| X_k \|_{2,1} \\ &\le \langle \nabla F(X_k), \theta D_k\rangle + \frac{\Lambda_k\theta^2}{2}\| D_k \|_F^2 + \theta\mu\| X_k + D_k \|_{2,1} + \mu(1-\theta)\| X_k \|_{2,1} - \mu\| X_k \|_{2,1}.\end{aligned}$$

Hence,

$$(1-\theta)\langle \nabla F(X_k), D_k\rangle + \mu(1-\theta)\| X_k + D_k \|_{2,1} - \mu(1-\theta)\| X_k \|_{2,1} \leq -\frac{\Lambda_k}{2}(1-\theta^2)\| D_k \|_F^2,$$

i.e.,

$$\langle \nabla F(X_k), D_k\rangle + \mu\| X_k + D_k \|_{2,1} - \mu\| X_k \|_{2,1} \leq -\frac{\Lambda_k}{2}(1+\theta)\| D_k \|_F^2 .$$

Recalling $\theta \in (0, 1]$, the above inequality indicates Eq (8) is correct. ♯

To improve the algorithm's performance, we use the classical nonmonotone line search [28] to find a suitable stepsize along the direction. It is well known that this technique allows the functional values to increase occasionally in some iterations but decrease in the whole iterative process. Letting $\delta \in (0, 1)$, $\rho \in (0, 1)$ and $\tilde{m}$ be a given positive integer, we choose the smallest nonnegative integer j_k such that the stepsize $\alpha_k = \tilde{\alpha}\rho^{j_k}$ satisfies

$$\Phi(X_k + \alpha_k D_k) \leq \max_{0\leq j\leq m(k)} \Phi(X_{k-j}) + \delta\alpha_k\Delta_k, \tag{9}$$

where $0 \leq m(k) \leq \min\{m(k-1)+1, \tilde{m}\}$ $(m(0) = 0)$ and

$$\Delta_k = \langle \nabla F(X_k), D_k\rangle + \mu\| X_k + D_k \|_{2,1} - \mu\| X_k \|_{2,1}. \tag{10}$$

From Eq (8), it is clear that $\Delta_k \leq -\frac{\Lambda_k}{2}\| D_k \|_F^2 < 0$ whenever $D_k \neq 0$, which shows that Eq (9) is well-defined.

In summary, the full steps of the **N**onmonotone **S**pectral **G**radient algorithm for $L_{2,1}$-norm minimization (abbr. NSGL21) can be described as follows:

Algorithm 1 (NSGL21)

Step 0. Choose initial point X_0, constants $\mu > 0$, $\tilde{\alpha} > 0$, $\rho \in (0, 1)$, $\delta \in (0, 1)$ and positive integer $\tilde{m}$. Set $k := 0$.
Step 1. Stop if $\|D_k\|_F = 0$. Otherwise, continue.
Step 2. Compute D_k via Eq (6).
Step 3. Compute α_k via Eq (9).
Step 4. Let $X_{k+1} := X_k + \alpha_k d_k$.
Step 5. Let $k := k+1$. Go to Step 1.

As is stated in the proceeding section that the generated direction descend automatically whenever $\Lambda_k > 0$. To ensure $\Lambda_k > 0$, we choose a sufficiently small $\Lambda_{(\min)} > 0$ and a sufficiently large $\Lambda_{(\max)} > 0$, such that Λ_k is forced as

$$\Lambda_k := \min\{\Lambda_{(\max)}, \max\{\Lambda_k, \Lambda_{(\min)}\}\}.$$

This approach ensures that the hereditary descent property is guaranteed at each and every step.

Remark 1. The steps of the proposed algorithm is novel and different to other existing approaches. The well-known approach [18] reformulated Problem (2) as the following constrained smooth convex optimization problem

$$\min_{X\in\mathbb{R}^{n\times t},\xi\in\mathbb{R}^n} \left\{ F(X) + \mu\sum_{i=1}^{n}\xi_i \mid \| X_{i,:} \| \leq \xi_i \right\},$$

and then solved via the Nesterov's method. The method in [19] paid attention least square

Model (1) and used an auxiliary variable to transform the model equivalently as

$$\min_{X\in\mathbb{R}^{n\times t}}\left\{\frac{1}{2}\parallel Y\parallel_2^2+\mu\parallel X\parallel_{2,1}\mid AX-b=Y\right\}.$$

An alternating direction method of multiplier is used immediately to solve the resulting model and closed-form solution are derived at each subproblem. Clearly, our proposed algorithm is different from the above mentioned approaches in sense that we solve the original Model (2) directly without any transformation. ♯

3 Convergence analysis

This section is devoted to establishing the global convergence of algorithm NSGL21. For this purpose, we make the following assumption.

Assumption 1. The level set $\Omega=\{X: F(X)\leq F(X_0)\}$ is bounded.

Lemma 2. Suppose that the Assumption 1 holds and the sequence $\{X_k\}$ is generated by Algorithm 1. Then X_k is a stationary point of Problem (2) if and only if $D_k=0$.

Proof. In the case of $D_k\neq 0$, Lemma 1 shows that D_k is a descent direction, which implies that X_k is not a stationary point of F. On the other hand, since $D_k=0$ is the solution of Eq (5), for any $\xi D\in\mathbb{R}^{n\times t}$ with $\xi>0$ we have

$$\langle\nabla F(X_k),\xi D\rangle+\frac{\Lambda_k\xi^2}{2}\parallel D\parallel_F^2+\mu\parallel X_k+\xi D\parallel_{2,1}\geq\mu\parallel X_k\parallel_{2,1}. \tag{11}$$

Combining the fact $F(X_k+\xi D)-F(X_k)=\langle\nabla F(X_k),\xi D\rangle+o(\xi)$ with Eq (11), it yields

$$\begin{aligned}\Phi'(X_k;D) &= \lim_{\xi\downarrow 0}\frac{F(X_k+\xi D)-F(X_k)+\mu\parallel X_k+\xi D\parallel_{2,1}-\mu\parallel X_k\parallel_{2,1}}{\xi}\\ &= \lim_{\xi\downarrow 0}\frac{\xi\langle\nabla F(X_k),D\rangle+o(\xi)+\mu\parallel X_k+\xi D\parallel_{2,1}-\mu\parallel X_k\parallel_{2,1}}{\xi}\\ &\geq \lim_{\xi\downarrow 0}\frac{-\frac{\Lambda_k\xi^2}{2}\parallel D\parallel_F^2+o(\xi)}{\xi}\\ &= 0,\end{aligned}$$

which indicates that X_k is a stationary point of F. ♯

Lemma 3. Let $l(k)$ be an integer such that

$$k-m(k)\leq l(k)\leq k\quad\text{and}\quad\Phi(X_{l(k)})=\max_{0\leq j\leq m(k)}\Phi(X_{k-j}).$$

Then the sequence $\{\Phi(X_{l(k)})\}$ is nonincreasing and the search direction $D_{l(k)}$ satisfies

$$\lim_{k\to\infty}\alpha_{l(k)}\parallel D_{l(k)}\parallel_F^2=0. \tag{12}$$

Proof. It is not difficult to see that $\Phi(X_{l(k+1)})\leq\Phi(X_{l(k)})$, which indicates that the maximum value of the objective function is nonincreasing at each iteration. Moreover, by Eq (9), we have that for all $k>\tilde{m}$,

$$\begin{aligned}\Phi(X_{l(k)}) &= \Phi(X_{l(k)-1}+\alpha_{l(k)-1}D_{l(k)-1})\\ &\leq \max_{0\leq j\leq m(l(k)-1)}\Phi(X_{l(k)-1-j})+\delta\alpha_{l(k)-1}\Delta_{l(k)-1}\\ &= \Phi(X_{l(l(k)-1)})+\delta\alpha_{l(k)-1}\Delta_{l(k)-1}.\end{aligned}$$

By Assumption 1, the sequence $\{\Phi(X_{l(k)})\}$ admits a limit as $k \to \infty$. Hence, it follows that

$$\lim_{k\to\infty} \alpha_{l(k)} \Delta_{l(k)} = 0. \tag{13}$$

On the other hand, by the definition of Δ_k in Eq (10) and the inequality Eq (8), it is easy to deduce that

$$\Delta_{l(k)} \leq -\frac{\Lambda_{(\min)}}{2} \| D_{l(k)} \|_F^2 < 0.$$

Combining with Eq (13), one get

$$\lim_{k\to\infty} \alpha_{l(k)} \| D_{l(k)} \|_F^2 = 0,$$

which indicates the desirable result Eq (12). ♯

Theorem 1. Let the sequence $\{X_k\}$ and $\{D_k\}$ be generated by Algorithm 1. Then, there exists a subsequence $k \in \mathcal{K}$ such that

$$\lim_{k\to\infty, k\in\mathcal{K}} \| D_k \|_F = 0. \tag{14}$$

Proof. Let $\bar{X}$ be a limit point of $\{X_k\}$, and $\{X_k\}_{\mathcal{K}_1}$ be a subsequence of $\{X_k\}$ converging to $\bar{X}$. Then by Eq (12) either $(\| \bar{D} \|_F :=) \lim_{k\to\infty, k\in\mathcal{K}_1} \| D_k \|_F = 0$, or there exists a subsequence $\{X_k\}_{\mathcal{K}}$ ($\mathcal{K} \subset \mathcal{K}_1$) such that

$$\lim_{k\to\infty, k\in\mathcal{K}} D_k \neq 0 \quad \text{and} \quad \lim_{k\to\infty, k\in\mathcal{K}} \alpha_k = 0. \tag{15}$$

In this condition, we assume that there exists a constant $\epsilon > 0$ such that

$$\| D_k \|_F \geq \epsilon, \quad \forall\ k \in \mathcal{K}. \tag{16}$$

Since α_k is the first value to satisfy Eq (9), it follows from Step 3 in Algorithm 1 that there exists an index $\bar{k}$ such that, for all $k \geq \bar{k}$ and $k \in \mathcal{K}$,

$$\Phi\left(X_k + \frac{\alpha_k}{\rho} d_k\right) > \max_{0\leq j\leq m(k)} \Phi(X_{k-j}) + \delta \frac{\alpha_k}{\rho} \Delta_k \geq \Phi(X_k) + \delta \frac{\alpha_k}{\rho} \Delta_k. \tag{17}$$

Since F is continuously differentiable, by the mean-value theorem on F, we can find that there exists a constant $\theta_k \in (0, 1)$, such that

$$F\left(X_k + \frac{\alpha_k}{\rho} D_k\right) - F(X_k) = \frac{\alpha_k}{\rho} \langle \nabla F\left(X_k + \theta_k \frac{\alpha_k}{\rho} D_k\right), D_k \rangle.$$

Combining with Eq (17), we have

$$\langle \nabla F\left(X_k + \theta_k \frac{\alpha_k}{\rho} D_k\right), D_k \rangle + \frac{\mu \| X_k + \frac{\alpha_k}{\rho} D_k \|_{2,1} - \mu \| X_k \|_{2,1}}{\alpha_k/\rho} > \delta \Delta_k. \tag{18}$$

Since $\alpha_k \to 0$ in Eq (15), we have $\alpha_k < \rho$ as $k \to \infty$. It is not difficult to show that

$$\frac{\mu \| X_k + \frac{\alpha_k}{\rho} D_k \|_{2,1} - \mu \| X_k \|_{2,1}}{\alpha_k/\rho} - \left[\mu \| X_k + D_k \|_{2,1} - \mu \| X_k \|_{2,1}\right] \leq 0. \tag{19}$$

Subtracting left side of Eq (18) by Δ_k and noting the definition of Δ_k, it is distinct that

$$\begin{aligned}
&\langle \nabla f\left(X_k+\theta_k\frac{\alpha_k}{\rho}D_k\right), D_k\rangle + \frac{\mu\| X_k+\frac{\alpha_k}{\rho}D_k\|_{2,1}-\mu\| X_k\|_{2,1}}{\alpha_k/\rho} - \Delta_k \\
&= \langle \nabla f\left(X_k+\theta_k\frac{\alpha_k}{\rho}D_k\right), D_k\rangle - \langle \nabla f(X_k), D_k\rangle \\
&+ \left[\frac{\mu\| X_k+\frac{\alpha_k}{\rho}D_k\|_{2,1}-\mu\| X_k\|_{2,1}}{\alpha_k/\rho} - \left(\mu\| X_k+D_k\|_{2,1}-\mu\| X_k\|_{2,1}\right)\right].
\end{aligned}$$

Noting Eq (19), thus Eq (18) shows that

$$\begin{aligned}
&\langle \nabla F\left(X_k+\theta_k\frac{\alpha_k}{\rho}D_k\right), D_k\rangle - \langle \nabla F(X_k), D_k\rangle \\
&> -(1-\delta)\Delta_k \\
&\geq (1-\delta)\frac{\Lambda_{(\min)}}{2}\| D_k\|_F^2 .
\end{aligned} \tag{20}$$

Taking the limit as $k\in\mathcal{K}, k\to\infty$ in the both sides of Eq (20) and using the smoothness of F, we obtain

$$0=\langle \nabla F(\bar{X}), \bar{D}\rangle - \langle \nabla F(\bar{X}), \bar{D}\rangle \geq (1-\delta)\frac{\Lambda_{(\min)}}{2}\| \bar{D}\|_F^2,$$

which implies $\|D_k\|_F\to 0$ as $k\in\mathcal{K}, k\to\infty$. This yields a contradiction because Eq (16) indicates that $\|D_k\|_F$ is bounded. ♯

4 Numerical experiments

In this section, we present numerical results to illustrate the feasibility and efficiency of the algorithm NSGL21. In particular, we also test against the recent solvers IADM_MFL and SLEP for performance comparison. In running SLEP (Sparse Learning with Efficient Projections), we use the code at http://www.public.asu.edu/~jye02/Software/SLEP/index.htm in its Matlab package, and choose `mFlag = 1` and `lFlag = 1` for using an adaptive line search. All experiments are carried out under Windows 7 and Matlab v7.8 (2009a) running on a Lenovo laptop with an Intel Pentium CPU at 2.5 GHz and 4 GB of memory.

As [16], in the first test, $\bar{X}_{:j}$ is generated from a 5-dimensional Gaussian distribution with zero-mean and con-variance diag{1, 0.64, 0.49, 0.36, 0.25}. Regarding each $\bar{X}_{:j}$, we keep adding up to 20 irrelevant dimensions which are exactly zeros. The training and test data A_j is Gaussian matrices and their response data b_j is generated by

$$b_j = A_j\bar{X}_{:j}+\omega,$$

where ω is zero-mean Gaussian noise with standard deviation $1.e-2$. We start NSGL21 from zero point and terminate the iterative process when

$$\| D_k\|_F < tol, \tag{21}$$

where $tol>0$ is a tolerance. The quality of the solution X^* is measured by the relative error to

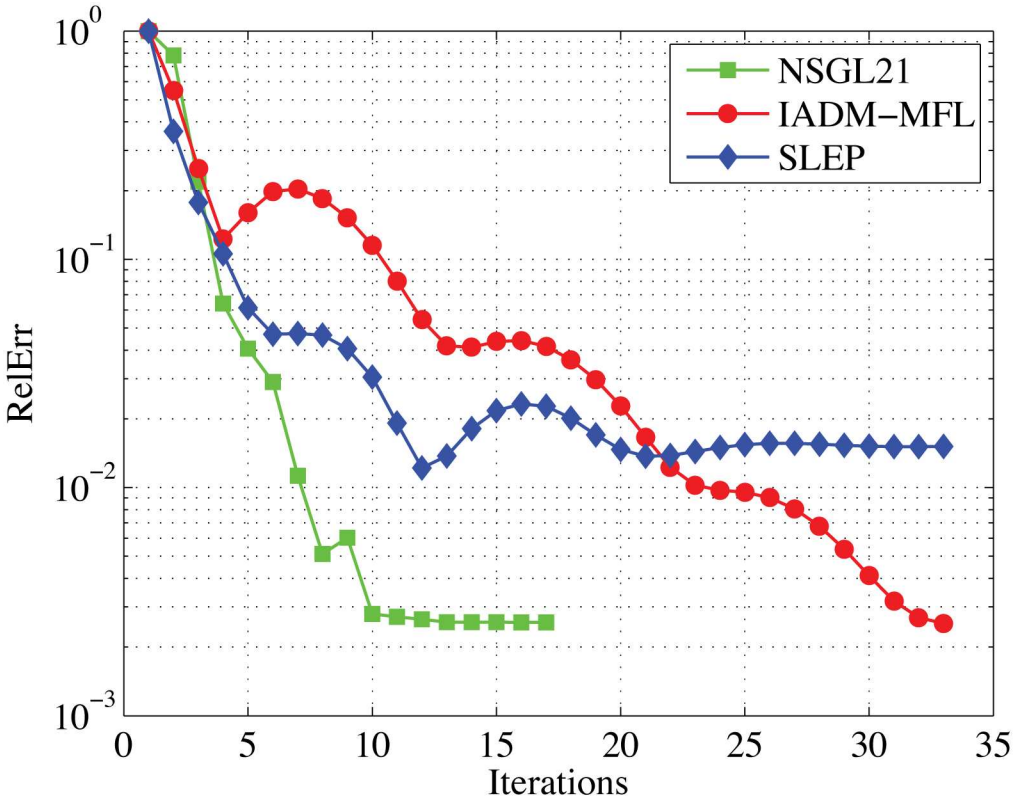

Fig 1. Comparison results of NSGL21, IADM MFL, and SLEP. The x-axes represents the number of iterations and the y-axes represents the relative error.

doi:10.1371/journal.pone.0166169.g001

$\bar{X}$, i.e.,

$$\text{RelErr} = \frac{\| X^* - \bar{X} \|_F}{\| \bar{X} \|_F}.$$

In this test, we take $\tilde{\alpha} = 1$, $\mu = 1e-2$, $t = 200$, $n = 15$, $tol = 1e-3$, $\Lambda_{(\min)} = 10^{-20}$, $\Lambda_{(\max)} = 10^{20}$, and $m_j = 100$ for all $j = 1, 2, \ldots, t$. Moreover, to compare the performance of these algorithms in a fair way, we run each code from zero point, use all the default parameter values, and observe their convergence behavior in obtaining similar accurate solutions. To specifically illustrate the performance of each algorithm, we draw a couple of figures to show their convergence behaviors with respect to the relative error and computing time proceed in Figs 1 and 2.

Observing Figs 1 and 2, we clearly know that IADM_MFL and NSGL21 produced faithful results expect for SLEP. We have tried to run SLEP with more iterations in our experiments' preparation, but it cannot achieve progress any more. Meanwhile, NSGL21 requires less number of iterations than IADM_MFL to achieve the similar quality of solutions. In both plots, we see that the green line lies at the bottom of each plot in most cases, which indicates that NSGL21 is superior to the other two solvers.

The simple test is not enough to verify that NSGL21 is the winner. To further illustrate the benefit of NSGL21, we give some insights to the behavior of NSGL21 with different dimensions and different number of tasks. The results are listed in Table 1, which contains the

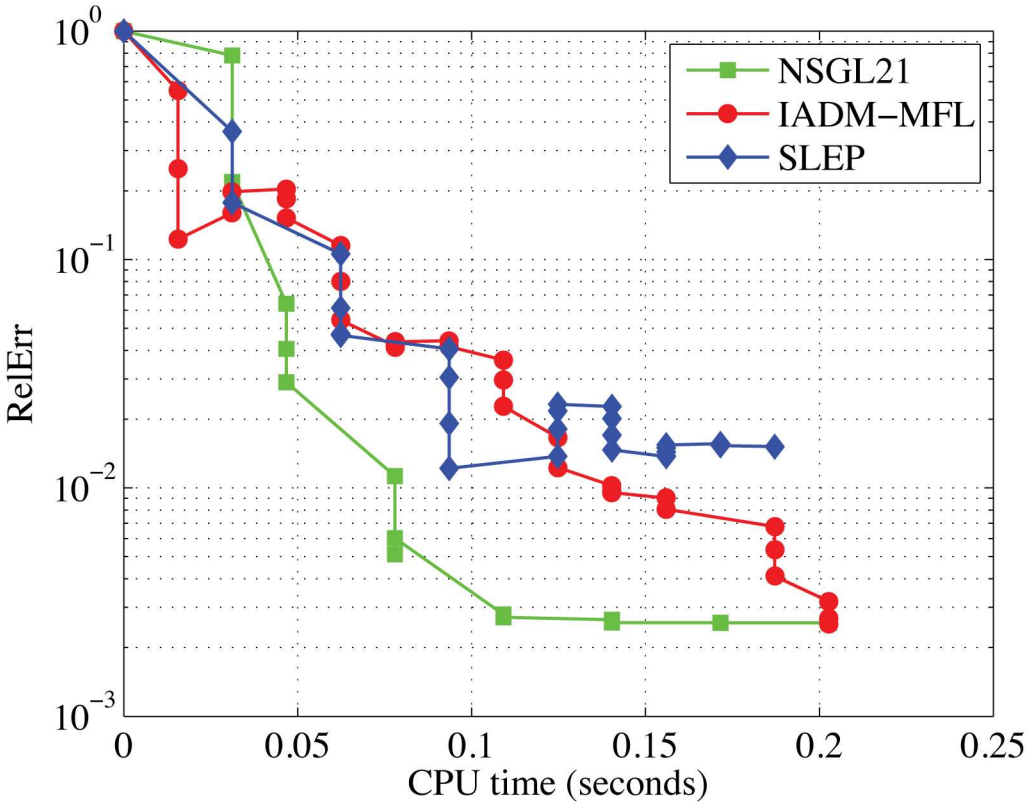

Fig 2. Comparison results of NSGL21, IADM MFL, and SLEP. The x-axes represents the CPU time in seconds and the y-axes represents the relative error.

doi:10.1371/journal.pone.0166169.g002

number of iterations (Iter), the CPU time in seconds (Time), the relative errors (RelErr), and the final functional values (Fun).

From Table 1, we clearly observe that each algorithm requires more computing time with the increase of the problems' dimensions and the number of tasks. Meanwhile, the number of iterations required by NSGL21 and IADM_MFL increases slightly at the higher dimensions case. We also observe that, for all the tested problems, both NSGL21 and IADM_MFL are terminated abnormally in producing similar quality solutions in sense of comparable relative errors and final function values. However, SLEP cannot generate acceptable solutions although more iterations are permitted in experiments' preparation. Hence, we conclude that NSGL21 and IADM_MFL perform better than SLEP. Now, we turn our attention to the performance comparison of solvers IADM_MFL and NSGL21. For getting similar quality of solutions, we take notice that NSGL21 is faster than IADM_MFL and saves at least 50% number of iterations. It is reasonable to make an conclusion that NSGL21 is the winner among the compared solvers.

5 Conclusions

In this paper, we have proposed, then analyzed, and later tested a nonmonotone spectral gradient algorithm for solving $\ell_{2,1}$-norm regularized minimization problem. The type of this problem mainly appears in computer version, text classification and biomedical informatics. Due

Table 1. Comparison results of NSGL21 with IADM_MFL and SLEP.

		NSGL21				IADM_MFL				SLEP			
t	n	Iter	Time	Error	Fun	Iter	Time	Error	Fun	Iter	Time	Error	Fun
50	5	12	0.03	1.32e-3	0.49	23	0.05	3.49e-3	0.53	32	0.06	1.66e-2	2.27
50	10	12	0.03	1.95e-3	0.48	32	0.06	2.28e-3	0.49	29	0.06	1.67e-2	2.26
50	15	14	0.03	2.33e-3	0.47	34	0.08	2.53e-3	0.48	29	0.03	1.67e-2	2.26
50	20	15	0.03	3.00e-3	0.45	39	0.06	2.79e-3	0.46	30	0.06	1.66e-2	2.26
50	25	14	0.05	3.49e-3	0.44	42	0.06	2.87e-3	0.45	33	0.09	1.65e-2	2.25
100	5	11	0.06	1.39e-3	0.83	24	0.05	1.62e-3	0.84	32	0.09	1.51e-2	3.61
100	10	12	0.05	2.13e-3	0.81	29	0.06	2.25e-3	0.83	41	0.09	1.52e-2	3.66
100	15	17	0.06	2.49e-3	0.79	33	0.09	2.55e-3	0.82	32	0.12	1.49e-2	3.57
100	20	15	0.09	2.99e-3	0.75	38	0.11	2.37e-3	0.79	32	0.11	1.50e-2	3.59
100	25	19	0.12	3.43e-3	0.74	43	0.14	2.72e-3	0.80	28	0.16	1.55e-2	3.73
150	5	12	0.06	1.43e-3	1.14	24	0.08	1.81e-3	1.16	35	0.14	1.51e-2	5.19
150	10	14	0.09	1.98e-3	1.11	29	0.09	2.44e-3	1.15	33	0.16	1.49e-2	5.18
150	15	17	0.12	2.57e-3	1.08	34	0.17	2.91e-3	1.15	32	0.22	1.51e-2	5.20
150	20	15	0.16	3.04e-3	1.03	40	0.20	2.79e-3	1.11	35	0.17	1.50e-2	5.16
150	25	19	0.22	3.45e-3	0.99	45	0.23	3.00e-3	1.08	35	0.28	1.49e-2	5.14
200	5	12	0.12	1.41e-3	1.45	24	0.12	1.68e-3	1.46	45	0.12	1.53e-2	7.10
200	10	12	0.12	1.94e-3	1.41	29	0.19	2.09e-3	1.45	41	0.14	1.53e-2	7.10
200	15	17	0.19	2.57e-3	1.35	33	0.25	2.54e-3	1.41	33	0.25	1.51e-2	6.98
200	20	15	0.19	3.10e-3	1.32	38	0.25	3.09e-3	1.41	34	0.25	1.51e-2	6.95
200	25	19	0.28	3.52e-3	1.26	43	0.31	3.22e-3	1.35	27	0.28	1.57e-2	7.30
250	5	11	0.12	1.43e-3	1.74	24	0.17	1.58e-3	1.75	38	0.28	1.55e-2	8.80
250	10	14	0.25	2.01e-3	1.68	31	0.25	2.30e-3	1.74	37	0.31	1.55e-2	8.77
250	15	17	0.28	2.58e-3	1.61	36	0.28	3.00e-3	1.71	33	0.31	1.54e-2	8.70
250	20	15	0.31	3.02e-3	1.56	39	0.36	3.13e-3	1.66	34	0.37	1.53e-2	8.70
250	25	19	0.37	3.46e-3	1.50	46	0.45	3.63e-3	1.62	30	0.25	1.61e-2	9.26
300	5	12	0.22	1.40e-3	2.04	26	0.25	1.77e-3	2.07	35	0.28	1.54e-2	10.55
300	10	12	0.23	2.04e-3	1.96	30	0.27	2.31e-3	2.03	45	0.42	1.57e-2	10.77
300	15	17	0.37	2.52e-3	1.90	35	0.37	3.10e-3	2.03	35	0.39	1.53e-2	10.50
300	20	14	0.37	3.03e-3	1.83	41	0.51	3.52e-3	1.96	34	0.31	1.54e-2	10.52
300	25	20	0.58	3.52e-3	1.72	45	0.62	4.36e-3	1.91	29	0.44	1.62e-2	11.26

doi:10.1371/journal.pone.0166169.t001

to the nonsmoothness of the regularization term, the task of minimizing the problem is full of challenges. To the best of our knowledge, SLEP and IADM_MFL are the only available solvers of solving this problem. However, both solvers transferred equivalently to an equality-constrained minimization problem and then minimized alternatively. As we all know that the spectral gradient algorithm is very effective to solve smooth minimization problem. Hence, its performance in solving $\ell_{2,1}$-norm regularized problems is worthy of investigating. Certainly, it is the main motivation of our paper. At each iteration, the method proposed in this paper minimizes an approximal quadratic model of the objective function to produce a search direction. We showed that the generated direction descends automatically and the algorithm converges globally under some mild conditions. Additionally, the numerical experiments illustrate that the proposed algorithm is competitive with or even performs better than SLEP and IADM_MFL. Of course, this is the numerical contribution of our paper. We have said that the $\ell_{2,1}$-norm regularized minimization problem is partly arising in multi-task learning for

capturing joint feather between each task. However, we did not test its real performance by using real data, this should be our further task to investigate. Finally, we expect that the proposed method and its extensions could produce even applications for problems in relevant areas of the machine learning.

Acknowledgments

The research of Y. Xiao was supported by the Major State Basic Research Development Program of China (973 Program) (Grant No. 2015CB856003), the National Natural Science Foundation of China (Grant No. 11471101), and the Program for Science and Technology Innovation Talents in Universities of Henan Province (Grant No. 13HASTIT050).

Author Contributions

Conceptualization: YX.

Data curation: QW.

Formal analysis: YX LL.

Methodology: YX LL.

Project administration: YX.

Software: QW.

Supervision: YX.

Validation: YX.

Writing – original draft: LL.

Writing – review & editing: YX.

References

1. Bi J, Xiong X, Yu S, Dundar M, Rao B, An improved multi-task learning approach with applications in medical diagnosis. In European Conference on Machine Learning, 2008.
2. Zhang J, Ghahramani Z, Yang Y, Flexible latent variable models for multi-task learning. Maching Learning, 3: (2008), 221–242. doi: 10.1007/s10994-008-5050-1
3. Zheng Y, Jeon B, Xu D, Wu QMJ, Zhang H. Image segmentation by generalized hierarchical fuzzy C-means algorithm Journal of Intelligent and Fuzzy Systems, 28 (2015), 961–973. doi: 10.3233/IFS-141378
4. Obozinski G, Taskar B, Jordan MI. Joint covariate selection for grouped classification. Technical report, Statistics Department, UC Berkeley, 2007.
5. Obozinski G, Taskar B, Jordan MI. Joint covariate selection and joint subspace selection for multiple classification problems. Statistics and Computing, 20 (2010), 231–252. doi: 10.1007/s11222-008-9111-x
6. Fu Z, Wu X, Guan C, Sun X, Ren K. Towards Efficient Multi-keyword Fuzzy Search over Encrypted Outsourced Data with Accuracy Improvement. IEEE Transactions on Information Forensics and Security, doi: 10.1109/TIFS.2016.2596138
7. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics, 23 (2007), 2507–2517. doi: 10.1093/bioinformatics/btm344 PMID: 17720704
8. Fu Z, Ren K, Shu J, Sun X, and Huang F. Enabling Personalized Search over Encrypted Outsourced Data with Efficiency Improvement. IEEE Transactions on Parallel and Distributed Systems, 2015. doi: 10.1109/TPDS.2015.2506573
9. Gu B, Sheng VS. A Robust Regularization Path Algorithm for v-Support Vector Classification IEEE Transactions on Neural Networks and Learning Systems, 2016 doi: 10.1109/TNNLS.2016.2527796 PMID: 26929067

10. Xia Z, Wang X, Sun X, Wang B. Steganalysis of least significant bit matching using multi-order differences. Security and Communocatopm Networks, 7 (2014), 1283–1291. doi: 10.1002/sec.864
11. Chen B, Shu H, Coatrieux G, Chen G, Sun X, Coatrieux JL. Color image analysis by quaternion-type moments. Journal of Mathematical Imaging and Vision, 51 (2015), 124–144. doi: 10.1007/s10851-014-0511-6
12. Xia Z, Wang X, Zhang L, Qin Z, Sun X, Ren K. A Privacy-preserving and Copy-deterrence Content-based Image Retrieval Scheme in Cloud Computing IEEE Transactions on Information Forensics and Security, 2016, doi: 10.1109/TIFS.2016.2590944
13. Ando RK, Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data Journal of Machine Learning Research, 6 (2005), 1817–1853.
14. Bakker B. Heskes T. Task clustering and gating for Bayesian multi-task learning Journal of Machine Learning Research, 4 (2003), 83–99.
15. Evgeniou T, Micchelli CA, Pontil M. Learning multiple tasks with kernel methods. Journal of Machine Learning Research, 6 (2005), 615–637.
16. Argyriou A, Evgeniou T, Massimiliano P. Convex multi-convex feature learning. Machine Learning, 73 (2008), 243–272. doi: 10.1007/s10994-007-5040-8
17. Obozinski G, Taskar B, Jordan MI. Multi-task feature selection Technical Report, UC Berkeley, 2006.
18. Liu J, Ji S, Ye J. Multi-task feather learning via efficient $\ell_{2,1}$-norm minimization. in Comference on Uncertainty in Artificial Intelligence, 2009.
19. Xiao Y, Wu SY, He BS. A proximal alternating direction method for $\ell_{2,1}$-norm least squares problem in multi-task feature learning. Journal of Industrial and Management Optimization, 8 (2012), 1057–1069. doi: 10.3934/jimo.2012.8.1057
20. Deng W, Yin W, Zhang Y. Group sparse optimization by alternating direction method Technical Report TR11-06, Rice University, 2011. available at http://www.caam.rice.edu/~zhang/reports/tr1106.pdf
21. Hu Y, Wei Z, Yuan G. Inexact accelerated proximal gradient algorithms for matrix $\ell_{2,1}$-Norm minimization problem in multi-task feature learning. Statistics, Optimization & Information Computing, 2 (2014), 352–367. doi: 10.19139/soic.v2i4.106
22. Barzilai J, Borwein JM. Two point step size gradient method. IMA Journal of Numerical Analysis, 8 (1988), 141–148. doi: 10.1093/imanum/8.1.141
23. Birgin EG, Martínez JM, Raydan M. Nonmonotone spectral projected gradient methods on convex sets. SIAM Journal on Optimization, 10 (2000), 1196–1121. doi: 10.1137/S1052623497330963
24. Raydan M. On the Barzilai and Borwein choice of steplength for the gradient methodz. IMA Journal of Numerical Analysis, 13 (1993), 321–326. doi: 10.1093/imanum/13.3.321
25. Raydan M. The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. SIAM Journal on Optimization, 7 (1997), 26–33. doi: 10.1137/S1052623494266365
26. Cheng W, Li DH. A derivative-free nonmonotone line search and its application to the spectral residual method. IMA Journal of Numerical Analysis, 29 (2009), 814–825. doi: 10.1093/imanum/drn019
27. Xiao Y, Wu SY, Qi L. Nonmonotone Barzilai-Borwein Gradient Algorithm for ℓ_1-Regularized Nonsmooth Minimization in Compressive Sensingzz. Journal of Scientific Computing, 61 (2014), 17–41. doi: 10.1007/s10915-013-9815-8
28. Grippo L, Lampariello F, Lucidi S. A nonmonotone line search technique for Newton's methodz. SIAM Journal on Numerical Analysis, 23 (1986), 707–716. doi: 10.1137/0723046