A Generalized Moreau Enhancement of $\ell_{2,1}$-norm and Its Application to Group Sparse Classification

Yang Chen, Masao Yamagishi, Isao Yamada
Dept. Information and Communications Engineering, Tokyo Institute of Technology, Tokyo, Japan
{chen, myamagi, isao}@sp.ict.e.titech.ac.jp

***Abstract*—In group sparse regularized least squares problem, the $\ell_{2,1}$-norm is widely used to approximate convexly the $\ell_{2,0}$ pseudo-norm, but it causes largely biased estimates which are not desired for many applications. In this paper, we propose a nonconvex group sparse regularizer which can be seen as a generalized Moreau enhancement of the $\ell_{2,1}$-norm. The proposed nonconvex regularizer promotes group sparsity more effectively than the $\ell_{2,1}$-norm and can achieve the overall convexity of the regularized least squares model. We also propose to apply this model to the group sparse classification problem. The proposed classifier can utilize the label information of training samples in terms of the grouping information with smaller bias than the $\ell_{2,1}$-norm, and thus is expected to improve the group sparse classification performance. Experimental results demonstrate that the proposed classifier certainly improves the performance of group sparse classification with $\ell_{2,1}$ regularizer, especially for unbalanced training data set.**

***Index Terms*—group sparsity, generalized Moreau enhancement, sparse representation based classification, proximal splitting algorithm**

I. Introduction

With the development of compressive sensing [1], [2], sparse representation has become an important tool for signal processing and machine learning [3]. It is based on the assumption that the signal of our interest can be represented as a linear combination of only a few columns in a dictionary matrix. In many applications, signals usually have specific sparsity structures. For example, in dynamic MRI [4], [5], DNA microarrays [6], hyperspectral unmixing [7], [8] and face recognition [9], the group sparsity structure has been exploited, that is, the ideal solution should have a natural grouping of its components, and the components within each group are likely to be either all zeros or all nonzeros [10]. Generally, grouping information is pre-defined based on prior knowledge of a specific problem.

Let $\boldsymbol{x} = [\boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{x}_2^{\mathrm{T}}, \cdots, \boldsymbol{x}_G^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^n$ represent the group structure of $\boldsymbol{x}$, where G is the number of groups. The group sparsity of $\boldsymbol{x}$ can be measured by the $\ell_{2,0}$ pseudo-norm, i.e., $\|\boldsymbol{x}\|_{2,0} = \|(\|\boldsymbol{x}_1\|_2, \|\boldsymbol{x}_2\|_2, \cdots, \|\boldsymbol{x}_G\|_2)\|_0$, where $\|\cdot\|_2$ is the Euclidean norm, and $\|\cdot\|_0$ is the ℓ_0 pseudo-norm which counts the number of nonzero entries in the vector. The underlying data can usually be represented approximately by a linear regression model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}+\boldsymbol{\varepsilon}$, where $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ are known, and $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ is an unknown noise vector. The group sparse regularized least squares problem can be modeled as

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_{2,0}, \tag{1}$$

where $\lambda > 0$ is the regularization parameter. Due to the fact that the combinatorial ℓ_0 minimization is an NP-hard problem [11], most of the studies choose $\ell_{2,1}$ as a convex approximation of $\ell_{2,0}$. More specifically, the widely used $\ell_{2,1}$-regularized least squares problem is defined as follows,

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_{2,1}, \tag{2}$$

where $\|\boldsymbol{x}\|_{2,1} = \sum_{i=1}^G \|\boldsymbol{x}_i\|_2$. This model is known as Group Lasso (least absolute shrinkage and selection operator) [12] in statistics. Although $\ell_{2,1}$ is the convex relaxation of $\ell_{2,0}$ and (2) can be efficiently solved through convex programming tools [13], it does not promote group sparsity as effective as $\ell_{2,0}$ mainly because of the large bias of $\ell_{2,1}$. In consideration of the good performance of nonconvex regularizers, some authors choose penalties such as group SCAD (smoothly clipped absolute deviation) [14], [15], group MCP (minimax concave penalty) [14], [15], $\ell_{p,q}$ ($\|\boldsymbol{x}\|_{p,q} := \left(\sum_{i=1}^G \|\boldsymbol{x}_i\|_p^q\right)^{1/q}$, $0 < q < 1 \leq p$) [16] and $\ell_{2,0}$ [17] for group sparse problems. However, these methods lose the overall convexity of the optimization problems and their algorithms have no guarantee of convergence to a global minimizer.

In this paper, in order to suppress the bias in the $\ell_{2,1}$-norm as well as to achieve the overall convexity of the group sparse regularized least squares model, we propose a generalized Moreau enhanced $\ell_{2,1}$ penalty based on linearly involved generalized-Moreau-enhanced (LiGME) model [18] (see Section II-A). We also propose to apply it to the group sparse classification (GSC) [19]–[22] which basically assigns the test sample to a class based on a group sparse representation with training samples (see Section II-B). By the proposed group sparse classifier, it is expected to improve the classification performance of the widely used $\ell_{2,1}$ penalty based classifiers while maintaining the overall convexity of the optimization model.

The reminder of this paper is as follows. Section II presents a brief review on LiGME and GSC. In Section III, the proposed group sparse representation method is introduced in details. Section IV discusses the application to group sparse classification problem. Conclusion is presented in Section V.

This work was supported in part by JSPS Grants-in-Aid (18K19804).

II. Review of previous work

A. Linearly involved generalized-Moreau-enhanced model

To promote sparsity or low-rankness more effectively than the convex envelopes of the direct discrete measures such as ℓ_0 and matrix rank, many efforts have been devoted to utilizing nonconvex regularizations while maintaining the overall convexity of the regularized least squares problems at the same time [18], [23]–[27]. Among them, linearly involved generalized-Moreau-enhanced (LiGME) model [18] provides a general framework to incorporate linear operators into a class of nonconvex penalties and builds parametric bridges between the direct discrete measures and their convex envelopes.

The LiGME model constructs nonconvex penalties for such regularized least squares while maintaining the convexity of the cost function. The model is given as the minimization of

$$J_{\Psi_{\boldsymbol{B}}\circ\mathfrak{L}}:\ \mathcal{X}\to\mathbb{R}:\ \boldsymbol{x}\mapsto\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_{\mathcal{Y}}^2+\lambda\Psi_{\boldsymbol{B}}\circ\mathfrak{L}(\boldsymbol{x}),\quad(3)$$

where $(\mathcal{X},\langle\cdot,\cdot\rangle_{\mathcal{X}},\|\cdot\|_{\mathcal{X}})$, $(\mathcal{Y},\langle\cdot,\cdot\rangle_{\mathcal{Y}},\|\cdot\|_{\mathcal{Y}})$, $(\mathcal{Z},\langle\cdot,\cdot\rangle_{\mathcal{Z}},\|\cdot\|_{\mathcal{Z}})$, $(\widetilde{\mathcal{Z}},\langle\cdot,\cdot\rangle_{\widetilde{\mathcal{Z}}},\|\cdot\|_{\widetilde{\mathcal{Z}}})$ are finite dimensional real Hilbert spaces, $\Psi\in\Gamma_0(\mathcal{Z})$ is coercive with $\mathrm{dom}\Psi=\mathcal{Z}$ [1], $\boldsymbol{B}\in\mathcal{B}(\mathcal{Z},\widetilde{\mathcal{Z}})$, $\mathfrak{L}\in\mathcal{B}(\mathcal{X},\mathcal{Z})$, $(\boldsymbol{A},\mathfrak{L},\lambda)\in\mathcal{B}(\mathcal{X},\mathcal{Y})\times\mathcal{B}(\mathcal{X},\mathcal{Z})\times\mathbb{R}_+$ and

$$\Psi_{\boldsymbol{B}}(\cdot):=\Psi(\cdot)-\min_{\boldsymbol{v}\in\mathcal{Z}}\left[\Psi(\boldsymbol{v})+\frac{1}{2}\|\boldsymbol{B}(\cdot-\boldsymbol{v})\|_{\widetilde{\mathcal{Z}}}^2\right].\quad(4)$$

We use the notations in [18] for specific $\Psi_{\boldsymbol{B}}(\cdot)$ in (4). For example, if $\Psi=\|\cdot\|_1$, $\Psi_{\boldsymbol{B}}(\cdot)$ is denoted by $(\|\cdot\|_1)_{\boldsymbol{B}}$. In this case, $J_{(\|\cdot\|_1)_{\boldsymbol{B}}\circ\mathrm{Id}}$ reproduces the model in [25], where Id is the identity operator. Although $\Psi_{\boldsymbol{B}}$ in (4) can be nonconvex for $\boldsymbol{B}\neq\boldsymbol{O}_{\mathcal{X}}$, the cost function $J_{\Psi_{\boldsymbol{B}}\circ\mathfrak{L}}$ in (3) is convex if

$$\boldsymbol{A}^*\boldsymbol{A}-\lambda\mathfrak{L}^*\boldsymbol{B}^*\boldsymbol{B}\mathfrak{L}\succeq\boldsymbol{O}_{\mathcal{X}},\quad(5)$$

where $\boldsymbol{A}^*$ denotes the adjoint of $\boldsymbol{A}$ and $\boldsymbol{O}_{\mathcal{X}}\in\mathcal{B}(\mathcal{X},\mathcal{X})$ is the zero operator. In particular, if Ψ is a certain norm over the vector space $\mathcal{Z}$, (5) becomes a necessary and sufficient condition of the overall convexity. In the case of $\mathfrak{L}=\mathrm{Id}$, the overall convexity condition is satisfied by $\boldsymbol{B}=\sqrt{\theta/\lambda}\boldsymbol{A}$ $(0\leq\theta\leq1)$ [25]. Furthermore, for any $\boldsymbol{B}$ satisfying the overall convexity condition (5), if $\Psi\in\Gamma_0(\mathcal{Z})$ is coercive, even symmetry and prox-friendly [2] with $\mathrm{dom}\Psi=\mathcal{Z}$, [18, Theorem 1] provides a proximal splitting algorithm of guaranteed convergence to a globally optimal solution of model (3).

B. Group sparse classification

For a classification problem, suppose that there are G classes of subjects, and let $\boldsymbol{A}=[\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_G]\in\mathbb{R}^{m\times n}$ be the matrix of n columns of training samples, where $\boldsymbol{A}_i=[\boldsymbol{a}_{i1},\boldsymbol{a}_{i2},\cdots,\boldsymbol{a}_{in_i}]\in\mathbb{R}^{m\times n_i}$ is the subset of the training samples from class i, $\boldsymbol{a}_{ij}$ represents the j-th training sample from the i-th class, n_i is the number of training samples in class i, and $n=\sum_{i=1}^{G}n_i$ is the number of training samples.

Sparse representation based classification (SRC) is first proposed by Wright et al. [28] for face recognition. It represents the test sample as a sparse linear combination of all training samples, and then classifies by the obtained sparse coefficients. Given a test sample $\boldsymbol{y}\in\mathbb{R}^m$, it is represented by $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x}\in\mathbb{R}^n$ is the desired sparse coefficient vector. The naive SRC is modeled as minimization of $\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}(\cdot)\|_2^2+\lambda\|\cdot\|_0$. However, this problem is NP-hard [11]. Most researchers use $\|\cdot\|_1$, in place of $\|\cdot\|_0$ in the model, since it is the largest convex minorant of $\|\cdot\|_0$ around zero. The relaxed model is the well-known Lasso [29].

In fact, for the classification task, the test sample is desired to be represented with training samples from as few classes as possible rather than few samples achievable by SRC. For pursuing the sparsity in the group form of more discriminative information than SRC, group sparse classification (GSC) approach is proposed [19], [20]. The naive GSC is to

$$\text{find }\hat{\boldsymbol{x}}:=\arg\min_{\boldsymbol{x}\in\mathbb{R}^n}\ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2+\lambda\|\boldsymbol{x}\|_{2,0}.\quad(6)$$

After obtaining the solution $\hat{\boldsymbol{x}}$, we can find the class of test sample $\boldsymbol{y}$ as the class that best approximates $\boldsymbol{y}$ by its training data. More precisely, $\boldsymbol{y}$ is assigned to class

$$i^\star=\arg\min_i\|\boldsymbol{y}-\boldsymbol{A}_i\hat{\boldsymbol{x}}_i\|_2.\quad(7)$$

Exactly for the same reason about convexity in SRC, $\|\cdot\|_{2,1}$ has been used as an approximation of $\|\cdot\|_{2,0}$ in (6) [19], [20], then the relaxed model becomes Group Lasso in (2). However, $\ell_{2,1}$ penalty not only suppresses the number of selected classes (nonzero groups in $\boldsymbol{x}$), but also suppresses significant nonzero coefficients within classes. The later may lead to biased estimates for high-amplitude elements and adversely affect the performance. In the application of face recognition, for tighter approximation of $\ell_{2,0}$, [21] employs a nonconvex surrogate function $\ell_{2,q}$ $(0<q<1)$ while [22] utilizes a MCP induced group sparse penalty at a cost of losing the convexity of the minimization problem. How to design a penalty that induces less bias and approximates $\ell_{2,0}$ better than $\ell_{2,1}$ while ensuring the overall convexity becomes a challenge.

In the next section, we present a less biased group sparsity inducing function and propose a convex model that can be used for classification problems to addresses the challenge.

III. Generalized Moreau enhancement of $\ell_{2,1}$-norm for group sparse representation

Let $\boldsymbol{B}\in\mathbb{R}^{l\times n}$, $\boldsymbol{x}=[\boldsymbol{x}_1^{\mathrm{T}},\boldsymbol{x}_2^{\mathrm{T}},\cdots,\boldsymbol{x}_G^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^n$ and $\boldsymbol{v}=[\boldsymbol{v}_1^{\mathrm{T}},\boldsymbol{v}_2^{\mathrm{T}},\cdots,\boldsymbol{v}_G^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^n$, where $\boldsymbol{x}_i,\boldsymbol{v}_i\in\mathbb{R}^{n_i}$ $(i=1,2,\cdots,G)$. Under the framework of LiGME, we can define our group sparse regularizer $(\|\cdot\|_{2,1})_{\boldsymbol{B}}:\mathbb{R}^n\to\mathbb{R}$ as follows,

$$(\|\cdot\|_{2,1})_{\boldsymbol{B}}(\boldsymbol{x})=\sum_{i=1}^{G}\|\boldsymbol{x}_i\|_2-\min_{\boldsymbol{v}\in\mathbb{R}^n}\left\{\sum_{i=1}^{G}\|\boldsymbol{v}_i\|_2+\frac{1}{2}\|\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{v})\|_2^2\right\}.\quad(8)$$

[1] $\Gamma_0(\mathcal{Z})$ is the set of proper lower semicontinuous convex function from $\mathcal{Z}$ to $(-\infty,\infty]$; a function $g:\mathcal{Z}\to(-\infty,\infty]$ is called coercive if $g(\boldsymbol{x})\to\infty$ $(\|\boldsymbol{x}\|_{\mathcal{Z}}\to\infty)$; $\mathrm{dom}\Psi$ denotes the domain of function Ψ.

[2] Even symmetry means $\Psi\circ(-\mathrm{Id})=\Psi$; prox-friendly means $\mathrm{Prox}_{\gamma\Psi}:\mathcal{Z}\to\mathcal{Z}:\boldsymbol{x}\mapsto\arg\min_{\boldsymbol{v}\in\mathcal{Z}}\{\Psi(\boldsymbol{v})+\frac{1}{2\gamma}\|\boldsymbol{v}-\boldsymbol{x}\|_{\mathcal{Z}}^2\}$ is computable $(\forall\gamma\in\mathbb{R}_{++})$.

Proposition 1. *(The group sparse penalty term $\lambda\left(\|\cdot\|_{2,1}\right)_{\boldsymbol{B}}$ can bridge the gap between $\|\cdot\|_{2,0}$ and $\|\cdot\|_{2,1}$.) Let $\boldsymbol{B}=\frac{1}{\sqrt{\gamma}}\boldsymbol{I}_n$ for $\gamma\in\mathbb{R}_{++}$ and $\lambda=\frac{2}{\gamma}$, where $\boldsymbol{I}_n$ is the n-dimensional identity matrix. Then, for any $\boldsymbol{x}\in\mathbb{R}^n$,*

$$\lim_{\gamma\downarrow 0}\frac{2}{\gamma}\left(\|\cdot\|_{2,1}\right)_{\frac{1}{\sqrt{\gamma}}\boldsymbol{I}_n}(\boldsymbol{x})=\|\boldsymbol{x}\|_{2,0}. \tag{9}$$

Together with the fact that $\left(\|\cdot\|_{2,1}\right)_{\boldsymbol{O}_n}(\boldsymbol{x})=\|\boldsymbol{x}\|_{2,1}$, the penalty term $\lambda\left(\|\cdot\|_{2,1}\right)_{\boldsymbol{B}}$ can serve as a parametric bridge between $\|\cdot\|_{2,0}$ and $\|\cdot\|_{2,1}$.

Proof. The penalty term $\lambda\left(\|\cdot\|_{2,1}\right)_{\boldsymbol{B}}$ reproduces

$$\begin{aligned}&\frac{2}{\gamma}\left(\|\cdot\|_{2,1}\right)_{\frac{1}{\sqrt{\gamma}}\boldsymbol{I}_n}:\mathbb{R}^n\to\mathbb{R}\\ &\qquad\qquad :[\boldsymbol{x}_1^{\mathrm{T}},\boldsymbol{x}_2^{\mathrm{T}},\cdots,\boldsymbol{x}_G^{\mathrm{T}}]^{\mathrm{T}}\mapsto\sum_{i=1}^{G}\frac{2}{\gamma}\psi_i(\boldsymbol{x}_i),\end{aligned} \tag{10}$$

where $\psi_i(\boldsymbol{x}_i)=\|\boldsymbol{x}_i\|_2-\min_{\boldsymbol{v}_i\in\mathbb{R}^{n_i}}\left\{\|\boldsymbol{x}_i\|_2+\frac{1}{2\gamma}\|\boldsymbol{x}_i-\boldsymbol{v}_i\|_2^2\right\}$, $i=1,\cdots,G$. By [30, Example 24.20], there is

$$\frac{2}{\gamma}\psi_i(\boldsymbol{x}_i)=\begin{cases}\frac{2}{\gamma}\|\boldsymbol{x}_i\|_2-\frac{1}{\gamma^2}\|\boldsymbol{x}_i\|_2^2, & \text{if } \|\boldsymbol{x}_i\|_2\le\gamma,\\ 1, & \text{otherwise,}\end{cases} \tag{11}$$

which satisfies

$$\lim_{\gamma\downarrow 0}\frac{2}{\gamma}\psi_i(\boldsymbol{x}_i)=\begin{cases}0, & \text{if } \|\boldsymbol{x}_i\|_2=0,\\ 1, & \text{otherwise.}\end{cases} \tag{12}$$

Therefore, $\lim_{\gamma\downarrow 0}\lambda\left(\|\cdot\|_{2,1}\right)_{\frac{1}{\sqrt{\gamma}}\boldsymbol{I}_n}=\lim_{\gamma\downarrow 0}\sum_{i=1}^{G}\lambda\psi_i(\boldsymbol{x}_i)=\sum_{i=1}^{G}|\|\boldsymbol{x}_i\|_2|_0=\|\boldsymbol{x}\|_{2,0}$, where $|t|_0=0$ if $t=0$ and $|t|_0=1$ otherwise. □

Remark 1. *Since the $\ell_{2,1}$-norm is in fact the combination of an ℓ_2-norm within class and an ℓ_1-norm across classes, there are three approaches to get a generalized Moreau enhanced group sparse penalty. In addition to (8), the other two approaches are: (i) applying LiGME strategy only on the inner ℓ_2 , while keeping the outer ℓ_1 unchanged; (ii) keeping the inner ℓ_2 maintained while only performing generalized Moreau enhanced operation on the outer ℓ_1. All these three approaches can bridge the gap between $\|\cdot\|_{2,0}$ and $\|\cdot\|_{2,1}$, as they are equal when $\boldsymbol{B}$ is a scalar multiple of the identity matrix. Here we omit detailed analysis. A penalty with $\boldsymbol{B}=\boldsymbol{I}_n$ is used in [31] for multiple measurement vector problem.*

For a group sparse regularized least squares problem, the optimization model with penalty (8) to estimate the vector $\boldsymbol{x}$ with group sparsity structure can be formulated as

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\ f(\boldsymbol{x}):=\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2+\lambda\left(\|\cdot\|_{2,1}\right)_{\boldsymbol{B}}(\boldsymbol{x}). \tag{13}$$

Note that the proposed model (13) reproduces Group Lasso model (2) in the case of $\boldsymbol{B}=\boldsymbol{O}_n$. By (5), the cost function $f(\boldsymbol{x})$ in (13) is convex if $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}-\lambda\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\succeq\boldsymbol{O}_n$. Since the proximal splitting algorithm in [18, Theorem 1] can be implemented for a general $\boldsymbol{B}$ that meets the convexity condition, by checking

1) $\|\cdot\|_{2,1}\in\Gamma_0(\mathbb{R}^n)$ is coercive;
2) $\|\cdot\|_{2,1}$ is even symmetry, i.e. $\|\cdot\|_{2,1}\circ(-\mathrm{Id})=\|\cdot\|_{2,1}$;
3) $\|\cdot\|_{2,1}$ is prox-friendly, where the proximity operator is given by

$$\begin{aligned}\mathrm{Prox}_{\gamma\|\cdot\|_{2,1}}:\ &\mathbb{R}^n\to\mathbb{R}^n\\ &\boldsymbol{x}\mapsto\left\{\left(1-\frac{\gamma}{\max\{\|\boldsymbol{x}_i\|_2,\gamma\}}\right)\boldsymbol{x}_i\right\}_{i=1}^{G};\end{aligned} \tag{14}$$

4) $\mathrm{dom}\left(\|\cdot\|_{2,1}\right)=\mathbb{R}^n$,

it can be utilized to solve optimization problem (13). Our method can be applied to many different applications that conform to group sparsity structure. We give the algorithm for the application of classification in the next section.

IV. Application to Group Sparse Classification

A. Proposed algorithm for group sparse classification

As discussed in Section II-B, $\ell_{2,1}$-norm is the most frequently adopted regularizer in GSC methods. In this section, we argue that its tendency to yield biased estimates for high-amplitude coefficients might lead to undesirable results. A toy example in Fig. 1 illustrates the potential risk of the $\ell_{2,1}$ regularizer, where four training samples (i.e., $\boldsymbol{a}_{1,1}$, $\boldsymbol{a}_{1,2}$, $\boldsymbol{a}_{2,1}$, $\boldsymbol{a}_{3,1}$) from three different classes (represented by triangle, diamond and circle respectively) and a query sample $\boldsymbol{y}$ (represented by star) are shown. The query sample can be well represented by the samples from triangle, i.e., $\boldsymbol{y}=0.95\boldsymbol{a}_{1,1}+0.04\boldsymbol{a}_{1,2}$. It can also be well represented by a combination of one sample from diamond and one sample from circle as $\boldsymbol{y}=0.50\boldsymbol{a}_{2,1}+0.40\boldsymbol{a}_{3,1}$.

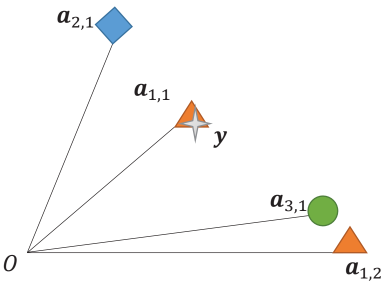

Fig. 1. A toy example of two different representations. A query image $\boldsymbol{y}$ (star) can be well represented by samples $\boldsymbol{a}_{1,1}$ and $\boldsymbol{a}_{1,2}$ of one class (triangle). It can also be well represented by samples $\boldsymbol{a}_{2,1}$ and $\boldsymbol{a}_{3,1}$ of another two classes (diamond and circle respectively).

Table I gives the value of different penalties (regularization parameter equals 1) by the above two representations. As we can see, ℓ_0 penalty cannot distinguish between the two representations in this case whereas $\ell_{2,0}$ penalty certainly chooses the first one, which demonstrates the advantage of GSC over SRC. However, $\ell_{2,1}$ penalty chooses the second one, the representation by samples from two classes instead of a single class, which is undesirable for GSC framework. This is due to that $\ell_{2,1}$ penalty refuses the large coefficient in the first representation, but accepts the two small coefficients in the second one.

TABLE I
PENALTY TERM VALUES OF REPRESENTATIONS IN FIG. 1

Penalty	ℓ_0	ℓ_1	$\ell_{2,0}$	$\ell_{2,1}$	$\ell_{2,1/2}$	$(\|\cdot\|_{2,1})_{\boldsymbol{I}_2}$
First	2	0.99	**1**	0.951	**0.951**	**0.50**
Second	2	0.90	2	0.90	1.794	0.695

Although $\ell_{2,0}$ and its approximation $\ell_{2,1/2}$ succeed in this example, they lead to nonconvex optimization. Thus we consider the proposed penalty in (8). For example, $(\|\cdot\|_{2,1})_{\frac{1}{\sqrt{\gamma}}\boldsymbol{I}_2}$ will prefer the first representation in Fig. 1 as long as $\gamma \leq 4.85$. This also reflects its role as a bridge between $\ell_{2,1}$ and $\ell_{2,0}$.

Since GSC relies on the group structure of training samples, when the training set is unbalanced, the influence caused by bias of $\ell_{2,1}$ penalty would be amplified. For example, a test sample $\boldsymbol{y}$ can be well represented by a combination of all n_i samples from class i, i.e., $\boldsymbol{y} = \boldsymbol{A}_i\boldsymbol{x}_i$ and $\|\boldsymbol{x}_i\|_1 = 1$, where $\boldsymbol{x}_i \in \mathbb{R}^{n_i}$. To simplify the expression, suppose that the number of samples in this class is doubled by duplication, and then the training set of class i becomes $\widetilde{\boldsymbol{A}_i} = [\boldsymbol{A}_i, \boldsymbol{A}_i] \in \mathbb{R}^{m\times 2n_i}$. Obviously, $\boldsymbol{y}$ can also be well represented by $\boldsymbol{y} = \widetilde{\boldsymbol{A}_i}\widetilde{\boldsymbol{x}}_i$, where $\widetilde{\boldsymbol{x}}_i = [\eta\boldsymbol{x}_i^{\mathrm{T}}, (1-\eta)\boldsymbol{x}_i^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2n_i}$ $(0 \leq \eta \leq 1)$ and $\|\widetilde{\boldsymbol{x}}_i\|_1 = 1$. However, $\|\boldsymbol{x}_i\|_2^2 - \|\widetilde{\boldsymbol{x}}_i\|_2^2 = 2\eta(1-\eta)\|\boldsymbol{x}_i\|_2^2 \geq 0$. That is, $\ell_{2,1}$ penalty value of the first representation (before duplication) is greater than that of the second one (after duplication). This indicates that the group size will affect the value of $\ell_{2,1}$ penalty. Therefore, when training set is unbalanced, $\ell_{2,1}$ penalty is unfair for classes of different sizes. The representation by groups with few samples is more likely to have a large $\ell_{2,1}$ penalty value. As $\ell_{2,1}$ penalty tends to refuse large coefficients, this bias can easily cause misclassification, especially when the correct class has relatively few samples. Note that $\ell_{2,0}$ penalty does not have such unfairness, it is independent of group size.

We expect to improve the performance of Group Lasso on unbalanced training sets by replacing the $\ell_{2,1}$ penalty with $(\|\cdot\|_{2,1})_{\boldsymbol{B}}$, that is, by using model (13). Considering that the proximal splitting algorithm in [18] can be utilized for any $\boldsymbol{B}$ satisfying $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} - \lambda\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} \succeq \boldsymbol{O}_n$, such as $\boldsymbol{B} = \sqrt{\theta/\lambda}\boldsymbol{A}$ $(0 \leq \theta \leq 1)$, we apply it to model (13) for classification problem (note that the algorithm in [25] cannot be applied to the case of $\theta = 1$, even though it satisfies the overall convexity condition). Our algorithm is guaranteed to converge to a globally optimal solution of the problem (13) under overall convexity condition. It is summarized in Algorithm 1. Compared with proximal gradient method for Group Lasso model (2) [32], Algorithm 1 requires at each update only one additional computation for $\mathrm{Prox}_{\gamma\|\cdot\|_{2,1}}$ in (14).

B. Experiments

In order to investigate the influence by bias of Group Lasso on the classification problem of unbalanced training set, and to see the performance improvement by the proposed method in such case, we conducted the experiments on a relatively simple dataset. The USPS handwritten digit database [33] has 11,000 samples of digits "0" through "9" (1,100 samples per class). The dimension of each sample is 16×16. In our classification experiments, the number of training samples for each class are not necessarily equal, which varies from 5 to 50 (the size of test set is fixed to 50 images per class).

Algorithm 1 The proposed group sparsity enhanced classification algorithm

```
Input: A matrix of training samples $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_G] \in \mathbb{R}^{d\times n}$ grouped by
    class information, and a test sample vector $\boldsymbol{y} \in \mathbb{R}^n$;
    1. Initialization: Let $(\boldsymbol{x}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{w}^{(0)}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$;
    Choose $(\sigma, \tau, \kappa) \in \mathbb{R}_{++} \times \mathbb{R}_{++} \times (1, +\infty)$ satisfying [3]
        $\sigma\boldsymbol{I}_n - \frac{\kappa}{2}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} \succeq \boldsymbol{O}_n \text{ and } \tau \geq \left(\frac{\kappa}{2} + \frac{2}{\kappa}\right)\lambda\|\boldsymbol{B}\|_{\text{spec}}^2,$
    where $\|\cdot\|_{\text{spec}}$ is the spectral norm calculating the largest singular value of a matrix.
    2. For $k = 0, 1, 2, \cdots$, compute
        $\boldsymbol{x}^{(k+1)} = \Big[\boldsymbol{I}_n - \frac{1}{\sigma}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} - \lambda\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})\Big]\boldsymbol{x}^{(k)} - \frac{\lambda}{\sigma}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{u}^{(k)} - \frac{\lambda}{\sigma}\boldsymbol{w}^{(k)} + \frac{1}{\sigma}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y},$
        $\boldsymbol{u}^{(k+1)} = \mathrm{Prox}_{\frac{\lambda}{\tau}\|\cdot\|_{2,1}}\Big[\frac{2\lambda}{\tau}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}^{(k+1)} - \frac{\lambda}{\tau}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}^{(k)} + (\boldsymbol{I}_n - \frac{\lambda}{\tau}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})\boldsymbol{u}^{(k)}\Big],$
        $\boldsymbol{w}^{(k+1)} = \Big(\mathrm{Id} - \mathrm{Prox}_{\|\cdot\|_{2,1}}\Big)\Big(2\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} + \boldsymbol{w}^{(k)}\Big),$
    until the stopping criterion is fulfilled.
    3. Compute the class label $i^\star$ of $\boldsymbol{y}$ by
        $i^\star = \arg\min_i \|\boldsymbol{y} - \boldsymbol{A}_i\boldsymbol{x}_i^{(k+1)}\|_2.$
Output: The class label $i^\star$ corresponding to $\boldsymbol{y}$.
```

We compared the proposed method with Lasso (SRC scheme) [28] and Group Lasso (GSC scheme) [20]. In order to achieve the overall convexity, we set $\boldsymbol{B} = \sqrt{\theta/\lambda}\boldsymbol{A}$ and fix $\theta = 0.9$ for proposed method. We set $\kappa = 1.1$, $\sigma = \|(\kappa/2)\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} + \lambda\boldsymbol{I}\|_{\text{spec}} + (\kappa - 1)$ and $\tau = (\kappa/2 + 2/\kappa)\lambda\|\boldsymbol{B}\|_{\text{spec}}^2 + (\kappa - 1)$. Lasso is implemented following [28], and Group Lasso is implemented by Algorithm 1 since the proposed model (13) with $\boldsymbol{B} = \boldsymbol{O}_n$ reproduces Group Lasso by setting $\theta = 0$. The initial estimate is set as $(\boldsymbol{x}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{w}^{(0)}) = (\boldsymbol{0}_n, \boldsymbol{0}_n, \boldsymbol{0}_n)$, and the stopping criterion is set to either $\|(\boldsymbol{x}^{(k)}, \boldsymbol{u}^{(k)}, \boldsymbol{w}^{(k)}) - (\boldsymbol{x}^{(k+1)}, \boldsymbol{u}^{(k+1)}, \boldsymbol{w}^{(k+1)})\|_2 < 10^{-4}$ or steps reaching 10,000.

Fig. 2 shows an example of unbalanced training set (digits "0" through "4" have 5 samples per class and "5" through "9" have 25 samples per class). The input is an image of digit "0" which was misclassified into digital "6" by Group Lasso while classified correctly by proposed method. In Fig. 2, the obtained coefficient vectors by Group lasso and proposed method (both with $\lambda = 4$) are illustrated respectively, and some samples corresponding to nonzero coefficients are also displayed. It can be seen that in Group Lasso, the samples from digit "6" made the greatest contribution to the representation, and samples from "5" and "0" also made small contribution. In our method, samples from the correct class "0" made the biggest contribution and led to correct result. This is reasonable, because our method did not excessively suppress the high value coefficients, whereas $\ell_{2,1}$ suppressed them too much. This large bias made the coefficients of the correct class cannot be large enough, and thus can easily lead to misclassification.

Table II summarizes the recognition accuracy of different methods (training set setting: digits "0" through "4" have β samples per class and "5" through "9" have α samples

[3] For example, any $\kappa > 1$, $\sigma = \|(\kappa/2)\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} + \lambda\boldsymbol{I}\|_{\text{spec}} + (\kappa - 1)$ and $\tau = (\kappa/2 + 2/\kappa)\lambda\|\boldsymbol{B}\|_{\text{spec}}^2 + (\kappa - 1)$ can satisfy (1).

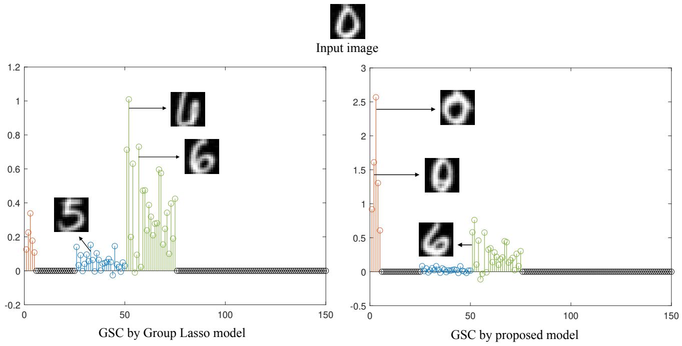

Fig. 2. Estimated sparse coefficients $\hat{\boldsymbol{x}}$.

TABLE II
RECOGNITION RESULTS ON THE USPS DATABASE

Method	Training set size ($\alpha = \max_i\{n_i\}$, $\beta = \min_i\{n_i\}$)						
	α	10		25		50	
	β	5	10	5	25	25	50
Lasso		82.2%	83.4%	79.6%	88.2%	86.2%	92.0%
Group Lasso		81.4%	86.6%	73.6%	91.4%	88.4%	93.2%
Proposed		82.6%	87.8%	80.8%	92.2%	90.6%	93.4%

per class). We see that the Group Lasso model degrades for unbalanced training set as expected, and the proposed method outperforms Group Lasso especially in such case.

V. Conclusion

We proposed a generalized Moreau enhancement of $\ell_{2,1}$-norm based on LiGME framework. This nonconvex penalty promotes group sparsity more effectively than $\ell_{2,1}$ with smaller bias while maintaining the overall convexity of the regression model. The proposed model can be utilized in many applications and we applied it to classification. Our model makes use of the grouping structure by class information and suppresses the tendency of bias estimation for high-amplitude coefficients. Experimental results showed that the proposed method is effective and competitive for image classification.

References

[1] D. L. Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[3] S. Theodoridis, *Machine learning: a Bayesian and optimization perspective*. Academic press, 2015.

[4] M. Usman, C. Prieto, T. Schaeffter, and P. Batchelor, "k-t group sparse: A method for accelerating dynamic mri," *Magnetic Resonance in Medicine*, vol. 66, no. 4, pp. 1163–1176, 2011.

[5] Y.-B. Lee, J. Lee, S. Tak, K. Lee, D. L. Na, S. W. Seo, Y. Jeong, J. C. Ye, A. D. N. Initiative *et al.*, "Sparse spm: Group sparse-dictionary learning in spm framework for resting-state functional connectivity mri analysis," *Neuroimage*, vol. 125, pp. 1032–1045, 2016.

[6] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi, "Recovering sparse signals using sparse measurement matrices in compressed dna microarrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 275–285, 2008.

[7] X. Wang, Y. Zhong, L. Zhang, and Y. Xu, "Spatial group sparsity regularized nonnegative matrix factorization for hyperspectral unmixing," *IEEE Trans. Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6287–6304, 2017.

[8] L. Drumetz, T. R. Meyer, J. Chanussot, A. L. Bertozzi, and C. Jutten, "Hyperspectral image unmixing with endmember bundles and group sparsity inducing mixed norms," *IEEE Trans. Image Processing*, vol. 28, no. 7, pp. 3435–3450, 2019.

[9] S. Tan, X. Sun, W. Chan, L. Qu, and L. Shao, "Robust face recognition with kernelized locality-sensitive group sparsity representation," *IEEE Trans. Image Processing*, vol. 26, no. 10, pp. 4661–4668, 2017.

[10] W. Deng, W. Yin, and Y. Zhang, "Group sparse optimization by alternating direction method," in *Wavelets and Sparsity XV*, vol. 8858. International Society for Optics and Photonics, 2013, p. 88580R.

[11] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[12] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[13] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University press, 2004.

[14] J. Huang, P. Breheny, and S. Ma, "A selective review of group selection in high-dimensional models," *Statistical Science*, vol. 27, no. 4, 2012.

[15] P. Breheny and J. Huang, "Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors," *Statistics and Computing*, vol. 25, no. 2, pp. 173–187, 2015.

[16] Y. Hu, C. Li, K. Meng, J. Qin, and X. Yang, "Group sparse optimization via $\ell_{p,q}$ regularization," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 960–1011, 2017.

[17] Y. Jiao, B. Jin, and X. Lu, "Group sparse recovery via the $\ell^0(\ell^2)$ penalty: Theory and algorithm," *IEEE Trans. Signal Processing*, vol. 65, no. 4, pp. 998–1012, 2017.

[18] J. Abe, M. Yamagishi, and I. Yamada, "Linearly involved generalized moreau enhanced models and their proximal splitting algorithm under overall convexity condition," *Inverse Problems*, vol. 36, p. 035012, 2020.

[19] A. Majumdar and R. K. Ward, "Classification via group sparsity promoting regularization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 861–864.

[20] J. Huang, F. Nie, H. Huang, and C. Ding, "Supervised and projected sparse coding for image classification," in *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[21] J. Zheng, P. Yang, S. Chen, G. Shen, and W. Wang, "Iterative reconstrained group sparse face recognition with adaptive weights learning," *IEEE Trans. Image Processing*, vol. 26, no. 5, pp. 2408–2423, 2017.

[22] C. Zhang, H. Li, C. Chen, Y. Qian, and X. Zhou, "Enhanced group sparse regularized nonconvex regression for face recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2020.

[23] A. Blake and A. Zisserman, *Visual reconstruction*. MIT press, 1987.

[24] M. Nikolova, M. K. Ng, and C.-P. Tam, "Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction," *IEEE Trans. Image Processing*, vol. 19, no. 12, pp. 3073–3088, 2010.

[25] I. Selesnick, "Sparse regularization via convex analysis," *IEEE Trans. Signal Processing*, vol. 65, no. 17, pp. 4481–4494, 2017.

[26] L. Yin, A. Parekh, and I. Selesnick, "Stable principal component pursuit via convex analysis," *IEEE Trans. Signal Processing*, vol. 67, no. 10, pp. 2595–2607, 2019.

[27] J. Abe, M. Yamagishi, and I. Yamada, "Convexity-edge-preserving signal recovery with linearly involved generalized minimax concave penalty function," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4918–4922.

[28] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2008.

[29] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[30] H. H. Bauschke, P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2017.

[31] K. Suzuki and M. Yukawa, "Robust recovery of jointly-sparse signals using minimax concave loss function," *IEEE Trans. Signal Processing*, vol. 69, pp. 669–681, 2021.

[32] Z. Qin, K. Scheinberg, and D. Goldfarb, "Efficient block-coordinate descent algorithms for the group lasso," *Mathematical Programming Computation*, vol. 5, no. 2, pp. 143–169, 2013.

[33] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550–554, 1994.