ITERATIVELY REWEIGHTED GROUP LASSO BASED ON LOG-COMPOSITE REGULARIZATION*

CHENGYU KE †, MIJU AHN†, SUNYOUNG SHIN ‡, AND YIFEI LOU‡

Abstract. The paper considers supervised learning problems of labeled data with grouped input features. The groups are non-overlapped such that the model coefficients corresponding to the input features form disjoint groups. The coefficients have group sparsity structure in the sense that coefficients corresponding to each group shall be simultaneously either zero or nonzero. To make effective use of such group sparsity structure given a priori, we introduce a novel log-composite regularizer, which can be minimized by an iterative algorithm. In particular, our algorithm iteratively solves for a traditional group LASSO problem that involves summing up the l_2 norm of each group until convergent. By updating group weights, our approach enforces a group of smaller coefficients from the previous iterate to be more likely to set to zero, compared to the group LASSO. Theoretical results include a minimizing property of the proposed model as well as the convergence of the iterative algorithm to a stationary solution under mild conditions. We conduct extensive experiments on synthetic and real datasets, indicating that our method yields superior performance over the state-of-the-art methods in linear regression and binary classification.

Key words. Group sparsity, nonconvex regularization, iteratively reweighted algorithm, directional stationarity, variable selection

AMS subject classifications. 62J07, 65C60, 65K05, 90C26, 92C60, 92D10

1. Introduction.

Supervised learning is an effective machine learning technique to harness the power of big data, where input features are used to predict output (response) values. A supervised learning model, such as regression and classification, is formulated by minimizing a cost function that associates the response with the input features via model coefficients. To facilitate feature (variable) selection in supervised learning, prior information and/or reasonable assumptions can be taken into account as a regularization. Since a large proportion of input features generated from big data correspond to zero coefficients, one of the most popular assumptions is sparsity, meaning a coefficient vector has only a few nonzero elements. To enforce sparsity, the l_1 norm of the coefficient vector has been extensively studied in compressed sensing, statistics, and operations research [7, 14]. For example, the l_1-regularized least squares functions, named least absolute shrinkage and selection operator (LASSO) [45], was proposed to promote zero coefficients in statistic models. The regularization was later extended to sparse signal recovery in the seminal works of [8, 10, 15, 24]. Many variations of LASSO have been studied in a statistical context, including [35, 43, 50, 59], to name a few. Despite the popularity, it has been pointed out in [16] that LASSO performs biased estimation towards coefficients with larger magnitude. To mitigate the undesirable effect, regularizers such as smoothly clipped absolute deviation (SCAD) [16] and minimax concave penalty (MCP) [54] were introduced with rigorous theoretical analyses. Other regularization methods include [30, 31, 33, 38, 46, 52], empirically showing great potential in promoting sparse solutions.

For effective recovery of group structured sparsity, existing methods capitalize on

*Submitted to the editors June XX, 2020.

Funding: The work of Ahn and Ke was partially supported by NSF under grant IIS 1948341. Lou was partially supported by NSF grant CAREER 1846690.

†Department of Engineering Management, Information, and Systems, Southern Methodist University, Dallas, TX 75205, USA. (cke@smu.edu, mijua@smu.edu).

‡Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080, USA. (sunyoung.shin@utdallas.edu, yifei.lou@utdallas.edu).

group patterns in the coefficients formulating a penalty function [22, 36, 39, 47, 53]. There are two types of group sparsity studied in the literature: intra-group sparsity and inter-group sparsity. Intra-group sparsity, which is also known as "within-group sparsity," means that coefficients belonging to the same group are sparse. Inter-group sparsity, or "across-group sparsity," refers to the case when only a few groups have nonzero coefficients. The traditional group LASSO [3, 32, 53] enforces the inter-group sparsity via minimization of a convex $l_{2,1}$ function to measure the total magnitude given by all groups, while one often uses the l_1 norm to promote the intra-group sparsity [57, 58].

Many algorithms have been introduced for group LASSO problems including coordinate descent [26, 39], a second-order cone program [28], a semismooth Newton method [56], a subspace acceleration method [12], and the alternating direction method of multipliers (ADMM) [4, 13] along with sophisticated ways of updating step sizes [29, 37]. However, an equal amount of penalization to all coefficients imposed by the group l_2 norm may result in erroneous group selection. In addition, group LASSO inherits the biased issue [16] from LASSO. More recently, nonconvex group regularizers such as group SCAD [47] and group MCP [22] have been introduced for group selection problems based on individual selection methods [16, 54]. Computational difficulties associated with nonconvexity of the latter regularizers were addressed by several developed algorithms [5, 51] with convergence analysis conducted in [48].

In many scientific and engineering applications, the across-group sparsity assumption is reasonable, and prior knowledge on grouping information contributes a better estimation of the model coefficients. Such prior information decomposes input features into non-overlapping groups, enforcing model coefficients corresponding to each group altogether zero or nonzero. Only a few groups are expected to be nonzero, which means that the input features of only those groups are useful in the outcome prediction. For example, patients' numerous medical conditions, which are used for predicting their disease outcomes with a supervised learning model [44], are grouped prior to the estimation by their known relationship. The medical conditions for each given group are simultaneously used in the model or not used at all, thus the multiple coefficients corresponding to each group are entirely nonzero or zero. Consequently, one requires the sparsity of the coefficients corresponding to each group to be enforced altogether [53]. Another commonly found example is when a small proportion of inputs are associated with an output in a nonlinear manner. One uses a basis expansion of each input to generate polynomial features forming an input group. Simultaneous regularization on the polynomial coefficients of each given group is desirable to identify a small number of nonzero groups that are useful to learn the association between the inputs and the output.

We propose a log-composite regularizer and an iteratively reweighted algorithm that aim to enforce the across-group sparsity in the model coefficients. For minimizing the nonconvex log-composite regularizer, we extend an iteratively reweighted l_1 minimization algorithm [9] for individual variable selection into group selection. One advantage of the proposed approach is its easy computation in that each iteration is quickly made by existing methods for group LASSO. We further show that the iterates of our algorithm converge to a stationary solution of the nonconvex formulation. Similar algorithms have been considered in [26, 39, 49]; however, many existing approaches are rather heuristic without explicitly having an objective function to minimize. Wipf and Nagarajan [49] considered a regularizer function mainly for individual sparsity. Zhao et al. [57, 58] discussed both within-group and across-group sparsity, but the reweighted scheme was adopted only on the within-group sparsity.

In summary, the major contributions of our work are four-fold,

(1) We introduce a novel log-composite function that aims to promote group structured sparsity of the model coefficients;
(2) To solve the nonconvex optimization problem, we present an efficient algorithm that iteratively minimizes a convex group LASSO problem weighted by the previous iterate;
(3) We provide convergence analysis of the proposed algorithm, showing that it reaches a stationary solution of the nonconvex formulation;
(4) We conduct extensive experiments, showcasing the superior performance of the proposed approach over the state-of-the-art methods in linear regression and binary classification problems.

The rest of the paper is organized as follows. In Section 2, we introduce a log-composite regularized formulation to recover across-group sparse patterns and investigate a minimizing property of a stationary solution. In Section 3, we present an iteratively reweighted algorithm to solve the proposed model and analyze the convergence of the algorithm. Section 4 presents in-depth numerical experiments for regression and classification with group structures on synthetic datasets and real datasets from genomics and public health studies. Finally, Section 5 concludes the paper.

2. Log-composite Regularization Method.

Given a group structure with subsets of the model coefficients forming distinct groups, we aim to simultaneously assign either zero or nonzero values to all members of a group. In this section, we first introduce a new regularizer for group variable selection, followed by comparison to existing ones and an optimality condition of stationary solutions.

Notation. We define some notations used throughout this paper. A (training) dataset consists of (i) an input data matrix $A \in \mathbb{R}^{n\times d}$ where each row of the matrix, denoted by $\mathbf{a}_i \in \mathbb{R}^d$, contains measurements of the i-th observation; and (ii) a response vector $\mathbf{b} \in \mathbb{R}^n$ where each component $b_i \in \mathbb{R}$ is a response obtained from the i-th observation, e.g., a continuous value or a discrete label. We denote $\mathbf{x} \in \mathbb{R}^d$ as a vector of coefficients or model coefficients to be trained.

To define a group structure, we assume that each variable x_j, for $j = 1, \ldots, d$, belongs to a group among a set of m pre-defined groups. For $k = 1, \ldots, m$, we define $\mathcal{G}_k \subseteq \{1, \ldots, d\}$ as an index set of the coefficients that belong to the k-th group. We assume each coefficient x_j only belongs to one group, i.e., $\mathcal{G}_k \cap \mathcal{G}_l = \emptyset$ for any $k \neq l$, and each $\mathcal{G}_k$ is a nonempty set such that $\cup_{k=1}^m \mathcal{G}_k = \{1, \ldots, d\}$. The cardinality of the set $\mathcal{G}_k$ is denoted by $|\mathcal{G}_k|$. Without loss of generality, we re-order $\mathbf{x}$ such that $\mathbf{x} = (\mathbf{x}_{\mathcal{G}_1}^T, \ldots, \mathbf{x}_{\mathcal{G}_m}^T)^T$, where $\mathbf{x}_{\mathcal{G}_k}$ is a subvector of $\mathbf{x}$ that consists of all the members of $\mathcal{G}_k$ for $k = 1, \ldots, m$.

2.1. Formulation.

Given the group membership of the variables, we consider an optimization problem defined by

$$\min_{\mathbf{x}} \; F_\lambda(\mathbf{x}) \triangleq L(\mathbf{x}) + \lambda \sum_{k=1}^{m} \sqrt{|\mathcal{G}_k|}\; J_\varepsilon(\mathbf{x}_{\mathcal{G}_k}), \tag{2.1}$$

where $L(\cdot)$ is a loss function, $J_\varepsilon(\cdot)$ is a group-specific regularization term with a hyperparameter ε, and λ is a hyperparameter to balance the two criteria. The model complexity is determined by the sum of the group-specific regularization terms $J_\varepsilon(\mathbf{x}_{\mathcal{G}_k})$, weighted by the cardinality of each group, denoted by $|\mathcal{G}_k|$. Given the data points

observed $\{(\mathbf{a}_i, b_i)\}_{i=1}^n$, or so-called training data, the loss function is defined by

$$L(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}; \mathbf{a}_i, b_i), \tag{2.2}$$

which measures how well the model is fitted to the observed data. The form of the loss function is typically determined by the response vector observed. If each b_i takes a real value, then the least squares loss, denoted by L^{ls}, is often used. The following problem is referred to as ordinary least squares (OLS),

$$\min_{\mathbf{x}} \; L^{\text{ls}}(\mathbf{x}) \triangleq \frac{1}{2n}||A\mathbf{x} - \mathbf{b}||_2^2. \tag{2.3}$$

If b_i is binary, then the logistic loss, denoted by L^{logit}, can be selected for binary classification. The logistic regression can be expressed as

$$\min_{\mathbf{x}} \; L^{\text{logit}}(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp(-b_i\, \mathbf{a}_i^T\mathbf{x})\right), \tag{2.4}$$

where $b_i \in \{-1, +1\}$ for $i = 1, \ldots, n$.

Motivated by the logarithmic function applied on every component for individual variable selection [9], we propose a log-composite function for group variable selection, named group LOG:

$$J_\varepsilon(\mathbf{x}_{\mathcal{G}_k}) \triangleq \log\left(\sqrt{\|\mathbf{x}_{\mathcal{G}_k}\|_2^2 + \varepsilon} + \|\mathbf{x}_{\mathcal{G}_k}\|_2\right), \tag{2.5}$$

where ε is a positive value. We choose such regularization (2.5) by further taking into an algorithmic aspect, which will be elaborated in Section 3. Note that we use the same form for each group and omit the group membership $\mathcal{G}_k$ in $\mathbf{x}$ when the context is clear. The roles of ε can be interpreted from several perspectives. On one hand, it gives a lower bound of $\log(\varepsilon^{\frac{1}{2}})$ on J_ε, preventing its output reaching to negative infinity. On the other hand, it provides numerical stability of our algorithm for the case when a group only consists of zero components. The log-composite function has several properties that promote sparsity in the obtained solution, which were first introduced in [16]. Specifically, the function is continuous, differentiable everywhere except at the origin, and the curvature of the function flattens as the input value increases. Such properties are known to be essential to achieve sparsity and unbiasedness of the solution. Moreover, the function is neither convex nor concave on $\mathbb{R}^p$ for $p \geq 2$. We provide a 2-dimensional example.

Example. We choose $\varepsilon = 1$. Suppose $J_\varepsilon(\mathbf{x})$ is convex in $\mathbf{x} \in \mathbb{R}^2$, then we must have $0.5J_\varepsilon(\mathbf{x}) + 0.5J_\varepsilon(\mathbf{y}) \geq J_\varepsilon(\mathbf{z})$ for any triplet $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ satisfying $0.5\mathbf{x} + 0.5\mathbf{y} = \mathbf{z}$. The inequality is violated if we select $\mathbf{x} = (0, 0), \mathbf{y} = (0, 2)$ and $\mathbf{z} = (0, 1)$. Alternatively, we must have $0.5J_\varepsilon(\mathbf{x}) + 0.5J_\varepsilon(\mathbf{y}) \leq J_\varepsilon(\mathbf{z})$ if $J_\varepsilon(\mathbf{x})$ is concave. The inequality does not hold if we choose $\mathbf{x} = (2, 4), \mathbf{y} = (2, 0)$ and $\mathbf{z} = (2, 2)$.

Note that the function J_ε is concave in $|x|$ for $x \in \mathbb{R}$, yet the above example involving points in the nonnegative orthant illustrates the function does not have such monotonic behavior in the higher dimension. The property described is also found in some existing functions for individual variable selection, e.g., transformed l_1 penalty [33, 55]; when a scalar input is replaced with the l_2 norm of coefficients belonging

Regularizer	Definition
Group LASSO [53]	$\sum_{k=1}^{m} \sqrt{\lvert\mathcal{G}_k\rvert}\ \lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2$
Group SCAD [47] ($\gamma > 2, \eta > 0$)	$\sum_{k=1}^{m} \sqrt{\lvert\mathcal{G}_k\rvert} \begin{cases} \eta\lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2, & \lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2 \le \eta \\ \dfrac{2\gamma\eta\lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2 - \lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2^2 - \eta^2}{2(\gamma-1)}, & \eta < \lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2 \le \gamma\eta \\ \dfrac{(\gamma+1)\eta^2}{2}, & \gamma\eta < \lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2 \end{cases}$
Group MCP [5] ($\gamma > 2, \eta > 0$)	$\sum_{k=1}^{m} \sqrt{\lvert\mathcal{G}_k\rvert} \begin{cases} \eta\lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2 - \dfrac{\lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2^2}{2\gamma}, & \lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2 \le \gamma\eta \\ \dfrac{1}{2}\gamma\eta^2, & \lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2 > \gamma\eta \end{cases}$
Group LOG ($\varepsilon > 0$)	$\sum_{k=1}^{m} \sqrt{\lvert\mathcal{G}_k\rvert}\ \log\left(\sqrt{\lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2^2 + \varepsilon} + \lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2\right)$

TABLE 1
A list of regularization functions to enhance group sparsity.

to the same group, the penalty function has a landscape of mixed curvatures. To the best of our knowledge, such property has not been discussed in the literature for group sparsity.

We list some regularizers for group variable selection in Table 1. Group LASSO [3, 53] uses the l_2 norm of coefficients of a given group, i.e., $\lVert\mathbf{x}_{\mathcal{G}_k}\rVert_2$, to promote group-wise sparse structure. Although group LASSO is computationally favorable due to its convexity, it imposes undesirable biased penalization towards groups having larger magnitude, which can be mitigated by nonconvex penalty functions. Figure 1 depicts the landscapes of the log-composite function J_ε in a 2-dimensional space with different values of ε. The function steeply increases near the origin, then flattens its curvature as the input increases. We also provide graphs of group LASSO and group MCP to compare their curvatures to those of the group LOG.

Assumptions on the loss function. We introduce a set of assumptions to be imposed on the loss function, denoted by $L : \mathbb{R}^d \to \mathbb{R}$ in (2.2),

A1. There exists a scalar $\sigma \ge 0$ such that

$$L(\mathbf{y}) - L(\mathbf{x}) - \langle \nabla L(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle \ge \frac{\sigma}{2}\lVert\mathbf{x} - \mathbf{y}\rVert_2^2 \quad \forall\, \mathbf{x}, \mathbf{y}.$$

A2. L is coercive, i.e., $\lim_{\lVert\mathbf{x}\rVert_2 \to \infty} L(\mathbf{x}) = +\infty$.

A3. L has a lower bound, denoted by $v \in \mathbb{R}$, such that $\inf_{\mathbf{x}} L(\mathbf{x}) \ge v$.

The assumption A1 implies that L is differentiable and convex (strongly convex for $\sigma > 0$). Although the assumption of strong convexity might be restrictive, we impose the condition to show the case when a global solution can be reached. One may impose a weaker assumption by localizing around the stationary solutions, referred to as restricted strong convexity in the literature [21]. However, this type of condition is hardly verifiable in practice, especially when comparing empirical solutions to the ground-truth coefficients that are not accessible; also see [1]. The assumptions A2 and A3 are used to show convergence of the algorithm in Section 3.2.

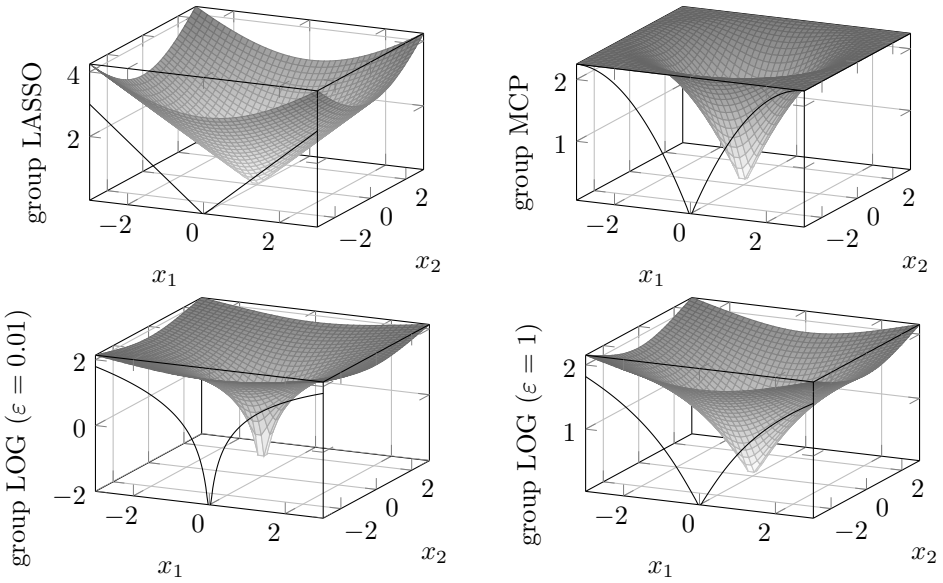

FIG. 1. *Plots of group sparsity regularizers where x_1 and x_2 form a group: (top-left) group LASSO; (top-right) group MCP defined by $\gamma = 2$ and $\eta = 1.5$. On the bottom, two plots of group LOG $J_\varepsilon(\mathbf{x})$ defined by different choices of $\varepsilon > 0$: (bottom-left) $\varepsilon = 0.01$, and (bottom-right) $\varepsilon = 1$.*

2.2. Property of stationary solutions. It is important to characterize an optimality property that a stationary solution of a nonconvex program can obtain. Here, we analyze the nonconvex and nondifferentiable problem (2.1) through directional stationary solutions [34], a specific kind of stationary solutions defined by the directional derivative. The directional derivative of F_λ at the point $\widehat{\mathbf{x}} \in \mathbb{R}^d$ in the direction $\mathbf{d} \in \mathbb{R}^d$ is defined by

$$F'_\lambda(\widehat{\mathbf{x}};\, \mathbf{d}) \triangleq \lim_{\tau \to 0^+} \frac{F_\lambda(\widehat{\mathbf{x}} + \tau \mathbf{d}) - F_\lambda(\widehat{\mathbf{x}})}{\tau}. \tag{2.6}$$

We point out that if there is a zero-group, e.g., $\mathbf{x}_{\mathcal{G}_k} = 0$, then the log-composite function is not differentiable in the ordinary sense, yet is directionally differentiable. Hence, the directional derivative of $F_\lambda(\cdot)$ is well-defined under the assumption A1. Formally, we define $\widehat{\mathbf{x}}$ as a directional stationary solution of the problem (2.1) if

$$F'_\lambda(\widehat{\mathbf{x}};\, \mathbf{x} - \widehat{\mathbf{x}}) \geq 0, \quad \forall \mathbf{x} \in \mathbb{R}^d. \tag{2.7}$$

Hereafter, a stationary solution refers to any point that satisfies (2.7).

We investigate a minimizing property that a stationary solution of (2.1) may have under some conditions. We present Lemma 2.1, which serves as a building block to show the property in Theorem 2.2. For simpler presentation, the following discussion uses the notation of $\mathbf{x}$ that holds for each $\mathbf{x}_{\mathcal{G}_k}$, as $k = 1, \cdots, m$.

LEMMA 2.1. *Given $\beta \geq \dfrac{2}{3\sqrt{3}\,\varepsilon}$, the function $G_\beta(\mathbf{x}) \triangleq \dfrac{\beta}{2}\|\mathbf{x}\|_2^2 + J_\varepsilon(\mathbf{x})$ is convex and hence we have*

$$J_\varepsilon(\mathbf{y}) - J_\varepsilon(\mathbf{x}) \geq J'_\varepsilon(\mathbf{x}; \mathbf{y} - \mathbf{x}) - \frac{\beta}{2}\|\mathbf{x} - \mathbf{y}\|_2^2, \quad \forall\, \mathbf{x},\, \mathbf{y}. \tag{2.8}$$

Proof. Observe that $G_\beta(\mathbf{x}) = G_1 \circ G_2(\mathbf{x})$, where $G_2(\mathbf{x}) \triangleq \|\mathbf{x}\|_2$ and $G_1(u) \triangleq \frac{\beta}{2}u^2 + \log\big(\sqrt{u^2+\varepsilon}+u\big)$ is a univariate function defined in the domain $u \geq 0$. Since G_2 is convex, it suffices to show that G_1 is non-decreasing and convex for all $u \geq 0$. For this purpose, we compute the first and second derivatives of G_1, given by

$$G_1'(u) = \beta u + \frac{1}{\sqrt{u^2+\varepsilon}} \quad \text{and} \quad G_1''(u) = \beta - \frac{u}{(u^2+\varepsilon)^{3/2}}.$$

For any $u \geq 0, G_1'(u) \geq 0$. If $\beta \geq \sup_{u\geq 0} \frac{u}{(u^2+\varepsilon)^{3/2}} = \frac{2}{3\sqrt{3}\,\varepsilon}$, we have $G_1''(u) \geq 0$, which validates its convexity. Therefore the composite function G_β is convex, and by definition, we have $G_\beta(\mathbf{y}) \geq G_\beta(\mathbf{x}) + G_\beta'(\mathbf{x};\mathbf{y}-\mathbf{x})$ which implies that

$$\frac{\beta}{2}\|\mathbf{y}\|_2^2 + J_\epsilon(\mathbf{y}) \geq \frac{\beta}{2}\|\mathbf{x}\|_2^2 + J_\epsilon(\mathbf{x}) + \langle \beta\mathbf{x}, \mathbf{y}-\mathbf{x}\rangle + J_\varepsilon'(\mathbf{x};\mathbf{y}-\mathbf{x}).$$

Some simple manipulations lead to the desired inequality (2.8). ∎

THEOREM 2.2. *Let Assumption A1 hold with $\sigma > 0$. If $\lambda \geq 0$ and satisfies $3\sqrt{3}\,(\sigma\,\varepsilon) \geq 2\,\lambda \max_{1\leq k\leq m}\sqrt{|\mathcal{G}_k|}$, then any stationary solution of (2.1) is a global minimizer.*

Proof. Denote $\widehat{\mathbf{x}}$ as a stationary point of (2.1) and β be a constant such that $\beta \geq \frac{2}{3\sqrt{3}\,\varepsilon}$. By applying the assumption A1 and Lemma 2.1, we have $\forall \mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned} F_\lambda(\mathbf{x}) - F_\lambda(\widehat{\mathbf{x}}) = \; & L(\mathbf{x}) + \lambda\sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\, J_\varepsilon(\mathbf{x}_{\mathcal{G}_k}) - L(\widehat{\mathbf{x}}) - \lambda\sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\, J_\varepsilon(\widehat{\mathbf{x}}_{\mathcal{G}_k}) \\ \geq \; & \frac{\sigma}{2}\,\|\mathbf{x}-\widehat{\mathbf{x}}\|_2^2 + \langle \nabla L(\widehat{\mathbf{x}}), \mathbf{x}-\widehat{\mathbf{x}}\rangle + \lambda\sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\,\Big[J_\varepsilon(\mathbf{x}_{\mathcal{G}_k}) - J_\varepsilon(\widehat{\mathbf{x}}_{\mathcal{G}_k})\Big] \qquad (2.9) \\ \geq \; & \frac{\sigma}{2}\,\|\mathbf{x}-\widehat{\mathbf{x}}\|_2^2 + \langle \nabla L(\widehat{\mathbf{x}}), \mathbf{x}-\widehat{\mathbf{x}}\rangle \\ & + \lambda\sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\,\Big[J_\varepsilon'(\widehat{\mathbf{x}}_{\mathcal{G}_k}; \mathbf{x}_{\mathcal{G}_k} - \widehat{\mathbf{x}}_{\mathcal{G}_k}) - \frac{\beta}{2}\|\mathbf{x}_{\mathcal{G}_k} - \widehat{\mathbf{x}}_{\mathcal{G}_k}\|_2^2\Big]. \end{aligned}$$

Due to the stationarity, $\widehat{\mathbf{x}}$ satisfies

$$\langle \nabla L(\widehat{\mathbf{x}}), \mathbf{x}-\widehat{\mathbf{x}}\rangle + \lambda\sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\, J_\varepsilon'(\widehat{\mathbf{x}}_{\mathcal{G}_k}; \mathbf{x}_{\mathcal{G}_k} - \widehat{\mathbf{x}}_{\mathcal{G}_k}) \geq 0.$$

The inequality (2.9) can thus be simplified as

$$\begin{aligned} F_\lambda(\mathbf{x}) - F_\lambda(\widehat{\mathbf{x}}) &\geq \frac{\sigma}{2}\,\|\mathbf{x}-\widehat{\mathbf{x}}\|_2^2 - \frac{\lambda\,\beta}{2}\sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\,\|\mathbf{x}_{\mathcal{G}_k} - \widehat{\mathbf{x}}_{\mathcal{G}_k}\|_2^2 \\ &\geq \frac{\sigma}{2}\,\|\mathbf{x}-\widehat{\mathbf{x}}\|_2^2 - \frac{\lambda\,\beta}{2}\max_{1\leq k\leq m}\sqrt{|\mathcal{G}_k|}\,\|\mathbf{x}-\widehat{\mathbf{x}}\|_2^2 = \frac{\xi}{2}\|\mathbf{x}-\widehat{\mathbf{x}}\|_2^2, \end{aligned}$$

where we define $\xi \triangleq \sigma - \lambda\beta\bar{\mathcal{G}}$ with $\bar{\mathcal{G}} \triangleq \max_{1\leq k\leq m}\sqrt{|\mathcal{G}_k|}$. Observe that if $\sigma - \lambda\beta\bar{\mathcal{G}} \geq 0$, i.e., $\lambda\bar{\mathcal{G}} \leq \frac{\sigma}{\beta} \leq \frac{3\sqrt{3}\varepsilon\sigma}{2}$, then $\xi \geq 0$ and hence $F_\lambda(\mathbf{x}) \geq F_\lambda(\widehat{\mathbf{x}})$ for all $\mathbf{x}$, which implies that any stationary solution $\widehat{\mathbf{x}}$ is a global minimizer. ∎

Lemma 2.1 shows that by adding a strongly convex function, e.g., $\|\cdot\|_2^2$, we can convexify J_ε and obtain G_β. Theorem 2.2 further shows that if the loss function is strongly convex with the modulus σ, the overall objective function is dominated by the convexity. As a result, any stationary point is in fact a global minimizer. Similar results are shown for a class of difference of convex programs in [2]. Besides, the theorem provides some insight about choosing hyperparameters ε and λ. Given the group structure, we can choose the ratio of the two hyperparameters according to the inequality stated in Theorem 2.2. Such choice of hyperparameters guarantees a global optimality for any stationary solution, even before the solution is computed.

3. Algorithm. Inspired by the iterative algorithm for reweighted l_1 minimization [9] and the majorize-minimization (MM) framework [23, 27], we propose to iteratively minimize a surrogate function by assigning a new weight for each group based on the magnitude of the previous iterate. The overall framework is described in Section 3.1, with convergence analysis in Section 3.2. In Section 3.3, we present the algorithms to minimize the surrogate function at each iteration, specifically designed for least squares loss (2.3) and logistic loss (2.4).

3.1. Majorize-minimization framework. Instead of a direct minimization of F_λ in (2.1), we employ the MM framework [23, 27] to minimize a surrogate function iteratively. The surrogate function, denoted by $\widehat{G}_\lambda$, shall satisfy the following conditions,

$$F_\lambda(\mathbf{x}^t) = \widehat{G}_\lambda(\mathbf{x}^t; \mathbf{x}^t) \quad \text{and} \quad F_\lambda(\mathbf{x}) \le \widehat{G}_\lambda(\mathbf{x}; \mathbf{x}^t),\ \forall \mathbf{x}. \tag{3.1}$$

We construct a surrogate function that satisfies (3.1),

$$\begin{aligned} \widehat{G}_\lambda(\mathbf{x}; \mathbf{x}^t) \triangleq\ & L(\mathbf{x}) + \frac{c}{2}\|\mathbf{x} - \mathbf{x}^t\|_2^2 \\ & + \lambda \sum_{k=1}^{m} \sqrt{|\mathcal{G}_k|} \left\{ \log\left(\sqrt{\|\mathbf{x}_{\mathcal{G}_k}^t\|_2^2 + \varepsilon} + \|\mathbf{x}_{\mathcal{G}_k}^t\|_2 \right) + \frac{1}{\sqrt{w_k^t}} \|\mathbf{x}_{\mathcal{G}_k}\|_2 - \frac{1}{\sqrt{w_k^t}} \|\mathbf{x}_{\mathcal{G}_k}^t\|_2 \right\}, \end{aligned} \tag{3.2}$$

where $w_k^t \triangleq \|\mathbf{x}_{\mathcal{G}_k}^t\|_2^2 + \varepsilon$ can be regarded as a weight for the k-th group at the t-th iteration and $c > 0$ yields the strong convexity of $\widehat{G}_\lambda$ necessary for convergence analysis. Note that a straightforward extension of [9] to group sparsity is using the regularization of $\log(\|\mathbf{x}_{\mathcal{G}_k}^t\|_2 + \varepsilon)$. Then, the MM framework yields the reciprocal of $\|\mathbf{x}_{\mathcal{G}_k}^t\|_2 + \varepsilon$ as the weight, which does not align with the square root of the group size $\sqrt{|\mathcal{G}_k|}$ used in this work as well as the existing group sparsity problems (See Table 1). Matching the group norm is one motivation that we choose the function in (2.5). Lemma 3.1 guarantees the mentioned relationship between $F_\lambda(\mathbf{x})$ and $\widehat{G}_\lambda(\mathbf{x}; \mathbf{x}^t)$. Lemma 3.2 shows the relationship between the directional derivatives of the two functions.

LEMMA 3.1. *Given $F_\lambda(\mathbf{x})$, $\widehat{G}_\lambda(\mathbf{x}; \mathbf{x}^t)$, defined in (2.1) and (3.2), the conditions in (3.1) hold.*

Proof. Clearly, the equality in (3.1) holds. To establish the inequality relationship, we consider a univariate function $G(u) \triangleq \log\left(\sqrt{u^2 + \varepsilon} + u\right)$ in the domain of nonnegative real numbers. It is straightforward to verify that $G(u)$ is a concave function on $[0, \infty)$ and hence we have $G(u) \le G(v) + G'(v)(u - v)$ for all $u, v \ge 0$. By

letting $u = \|\mathbf{x}_{\mathcal{G}_k}\|_2$ and $v = \|\mathbf{x}^t_{\mathcal{G}_k}\|_2$, we have

$$\begin{aligned} F_\lambda(\mathbf{x}) &= L(\mathbf{x}) + \lambda \sum_{k=1}^{m} \sqrt{|\mathcal{G}_k|} \log\left(\sqrt{\|\mathbf{x}_{\mathcal{G}_k}\|_2^2 + \varepsilon} + \|\mathbf{x}_{\mathcal{G}_k}\|_2\right) \\ &\le L(\mathbf{x}) + \lambda \sum_{k=1}^{m} \sqrt{|\mathcal{G}_k|} \left[J_\varepsilon(\mathbf{x}^t_{\mathcal{G}_k}) + \frac{1}{\sqrt{w_k^t}} \left(\|\mathbf{x}_{\mathcal{G}_k}\|_2 - \|\mathbf{x}^t_{\mathcal{G}_k}\|_2 \right) \right] \\ &= \widehat{G}_\lambda(\mathbf{x};\, \mathbf{x}^t) - \frac{c}{2}\|\mathbf{x} - \mathbf{x}^t\|_2^2 \le \widehat{G}_\lambda(\mathbf{x};\, \mathbf{x}^t). \end{aligned} \tag{3.3}$$

∎

LEMMA 3.2. *Let Assumption A1 hold. The directional derivatives of $F_\lambda(\mathbf{x})$ and $\widehat{G}_\lambda(\mathbf{x};\mathbf{x}^t)$ with respect to $\mathbf{x}$ in any given direction $\mathbf{d} \in \mathbb{R}^d$ are equal at the point $\mathbf{x} = \mathbf{x}^t$, i.e.,*

$$F'_\lambda(\mathbf{x}^t;\mathbf{d}) = \widehat{G}_\lambda(\cdot\,;\mathbf{x}^t)'(\mathbf{x}^t;\mathbf{d}). \tag{3.4}$$

Proof. For the case $\mathbf{x}^t_{\mathcal{G}_k} \neq 0$ for all $k = 1, \ldots, m$, it is straightforward to have $\nabla F_\lambda(\mathbf{x}^t) = \nabla_{\mathbf{x}} \widehat{G}_\lambda(\mathbf{x}^t;\mathbf{x}^t)$. If there exists $\mathbf{x}^t_{\mathcal{G}_k} = 0$ for some k, the equivalence in (3.4) can be directly shown by using the definition of directional derivative given in (2.6). ∎

Since $\widehat{G}_\lambda(\mathbf{x};\mathbf{x}^t)$ serves an upper bound of F_λ, we can minimize this upper bound to get a new solution, i.e.,

$$\mathbf{x}^{t+1} \in \operatorname*{argmin}_{\mathbf{x}} \ \widehat{G}_\lambda(\mathbf{x};\, \mathbf{x}^t), \tag{3.5}$$

which can be obtained by solving

$$\mathbf{x}^{t+1} \in \operatorname*{argmin}_{\mathbf{x}} \ L(\mathbf{x}) + \frac{c}{2}\|\mathbf{x} - \mathbf{x}^t\|_2^2 + \lambda \sum_{k=1}^{m} \frac{\sqrt{|\mathcal{G}_k|}}{\sqrt{w_k^t}} \|\mathbf{x}_{\mathcal{G}_k}\|_2 \tag{3.6}$$

where $w_k^t = \|\mathbf{x}^t_{\mathcal{G}_k}\|_2^2 + \varepsilon$. The convergence analysis is based on the MM framework of using (3.5). We rewrite it as (3.6) in order to see the connection to group LASSO, i.e., (3.6) is equivalent to weighted group LASSO for $c = 0$. We can interpret (3.6) that our algorithm assigns a different weight on each group determined by the previous iterate. Since the magnitude of w_k^t is smaller if $\mathbf{x}_{\mathcal{G}_k}$ is closer to the zero vector, the algorithm imposes a larger weight for zero-groups than for nonzero-groups.

3.2. Convergence analysis. We characterize the convergence of the iterative scheme (3.5) in Theorem 3.3. We note that (3.5) is a special case of successive upper-bound minimization (SUM) algorithm [40]. We borrow some techniques from the convergence analysis of the SUM algorithm for a generalized directionally differentiable function given in [40, Theorem 1] to prove Theorem 3.3.

THEOREM 3.3. *Suppose Assumptions A1-A3 hold and $c > 0$. The sequence of $\{\mathbf{x}^t\}_{t=1}^\infty$ produced by (3.5) achieves the following properties:*

(a) We have

$$F_\lambda(\mathbf{x}^t) - F_\lambda(\mathbf{x}^{t+1}) \ge \frac{c}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2, \tag{3.7}$$

and the sequence $\{F_\lambda(\mathbf{x}^t)\}_{t=1}^\infty$ converges;

(b) $\|\mathbf{x}^t - \mathbf{x}^{t+1}\|_2 \to 0$ as $t \to \infty$;

(c) The sequence $\{\mathbf{x}^t\}_{t=1}^{\infty}$ is bounded, and every limit point of $\{\mathbf{x}^t\}_{t=1}^{\infty}$ is a stationary solution of the problem (2.1).

Proof. (*a*) By Lemma 3.1 and the iterative scheme of (3.5), we have

$$\widehat{G}_\lambda(\mathbf{x}^{t+1};\mathbf{x}^t) \le \widehat{G}_\lambda(\mathbf{x}^t;\mathbf{x}^t) = F_\lambda(\mathbf{x}^t). \tag{3.8}$$

The inequalities (3.3) and (3.8) yield (3.7), which implies that $F_\lambda(\mathbf{x}^t)$ is a decreasing sequence with respect to t. By Assumption A3, L is bounded below, so is F_λ. By the Monotone Convergence Theorem, the sequence $\{F_\lambda(\mathbf{x}^t)\}_{t=1}^{\infty}$ converges.

(*b*) By summing (3.7) from $t=0$ to ∞, we have

$$F_\lambda(\mathbf{x}^0) - \lim_{t\to\infty} F_\lambda(\mathbf{x}^t) \ge \frac{c}{2}\sum_{t=0}^{\infty}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_2^2. \tag{3.9}$$

Since $c>0$, (3.9) implies that $\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_2^2 \to 0$.

(*c*) By Assumption A2, $L(\mathbf{x})$ is coercive, so is $F_\lambda(\mathbf{x})$ and hence $\{\mathbf{x}^t\}_{t=1}^{\infty}$ is bounded from (3.9). Therefore, $\{\mathbf{x}^t\}_{t=1}^{\infty}$ has a convergent subsequence, denoted by $\mathbf{x}^{t_j}\to\mathbf{x}^*$ as $t_j\to\infty$. From (b), we have $\mathbf{x}^{t_j+1}\to\mathbf{x}^*$ as well. Using (3.3) and (3.8), we obtain

$$\widehat{G}_\lambda(\mathbf{x}^{t_{j+1}};\mathbf{x}^{t_{j+1}}) = F_\lambda(\mathbf{x}^{t_{j+1}}) \le \widehat{G}_\lambda(\mathbf{x}^{t_{j+1}};\mathbf{x}^{t_j}) \le \widehat{G}_\lambda(\mathbf{x};\mathbf{x}^{t_j}), \quad \forall \mathbf{x}.$$

By taking limit on $t_j\to\infty$, we have $\widehat{G}_\lambda(\mathbf{x}^*;\mathbf{x}^*) \le \widehat{G}_\lambda(\mathbf{x};\mathbf{x}^*), \forall \mathbf{x}$, which implies that $\widehat{G}_\lambda(\cdot;\mathbf{x}^*)'(\mathbf{x}^*;\mathbf{x}-\mathbf{x}^*) \ge 0, \forall \mathbf{x}\in\mathbb{R}^d$. By Lemma 3.2, we obtain $F'_\lambda(\mathbf{x}^*;\mathbf{x}-\mathbf{x}^*)\ge 0$ which shows that every limit point of $\{\mathbf{x}^t\}_{t=1}^{\infty}$ is a stationary solution of (2.1). ∎

3.3. Subproblem solution. We solve for (3.6) at each iteration, which is equivalent to weighted group LASSO for $c=0$. There are many efficient algorithms for group LASSO that can be adapted here, among which we describe a vanilla ADMM framework [4]. In particular, we introduce an auxiliary variable $\mathbf{z}$ and rewrite (3.6) into an equivalent form

$$\min_{\mathbf{x},\mathbf{z}} \; L(\mathbf{z}) + \frac{c}{2}\|\mathbf{z}-\mathbf{x}^t\|_2^2 + \lambda\sum_{k=1}^{m}\frac{\sqrt{|\mathcal{G}_k|}}{\sqrt{w_k^t}}\|\mathbf{x}_{\mathcal{G}_k}\|_2 \qquad \text{s.t.} \qquad \mathbf{x}=\mathbf{z}. \tag{3.10}$$

The corresponding augmented Lagrangian is expressed as

$$\mathcal{L}(\mathbf{x},\mathbf{z};\mathbf{v}) \triangleq L(\mathbf{z}) + \frac{c}{2}\|\mathbf{z}-\mathbf{x}^t\|_2^2 + \lambda\sum_{k=1}^{m}\frac{\sqrt{|\mathcal{G}_k|}}{\sqrt{w_k^t}}\|\mathbf{x}_{\mathcal{G}_k}\|_2 + \langle\rho\mathbf{v},\mathbf{x}-\mathbf{z}\rangle + \frac{\rho}{2}\|\mathbf{x}-\mathbf{z}\|_2^2,$$

where $\mathbf{v}$ is a Lagrangian multiplier (or dual variable) and ρ is a positive parameter. We consider a scaled form by multiplying ρ by $\langle\mathbf{v},\mathbf{x}-\mathbf{z}\rangle$, and hence ADMM iterations proceed as follows:

$$\begin{cases} \mathbf{x}^{\tau+1} \in \operatorname{argmin}_{\mathbf{x}} \mathcal{L}(\mathbf{x},\mathbf{z}^\tau;\mathbf{v}^\tau) \\ \mathbf{z}^{\tau+1} \in \operatorname{argmin}_{\mathbf{z}} \mathcal{L}(\mathbf{x}^{\tau+1},\mathbf{z};\mathbf{v}^\tau) \\ \mathbf{v}^{\tau+1} = \mathbf{v}^\tau + \mathbf{x}^{\tau+1} - \mathbf{z}^{\tau+1}, \end{cases} \tag{3.11}$$

where we use the index of τ to indicate its inner iteration, as opposed to t for the outer iteration of the MM framework (3.6). We then elaborate on how to solve the two minimization problems in (3.11). Specifically for the $\mathbf{x}$-subproblem, it can be decomposed into each group

$$\mathbf{x}_{\mathcal{G}_k}^{\tau+1} \in \operatorname*{argmin}_{\mathbf{x}} \lambda_k\|\mathbf{x}\|_2 + \frac{\rho}{2}\|\mathbf{x}-\mathbf{z}_{\mathcal{G}_k}^\tau + \mathbf{v}_{\mathcal{G}_k}^\tau\|_2^2, \tag{3.12}$$

Algorithm 3.1 Iteratively reweighted algorithm for least squares loss

Set c, ε, ρ, and λ
Set tolerance parameters $\delta_1, \delta_2, \delta_3, \delta_4$ and maximum iteration numbers Δ_1, Δ_2
Initialize $\mathbf{x}^0, \mathbf{z}^0, \mathbf{v}^0, t = 0$
repeat
  for $k = 1, 2, \ldots, m$ **do**
    $w_k^t = \|\mathbf{x}_{\mathcal{G}_k}^t\|_2^2 + \varepsilon$
    $\lambda_k = \lambda\sqrt{|\mathcal{G}_k|/w_k^t}$
  end
  repeat
    for $k = 1, 2, \ldots, m$ **do**
      $\mathbf{x}_{\mathcal{G}_k}^{\tau+1} = S(\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}, \lambda_k/\rho)$
    end
    $\mathbf{z}^{\tau+1} = \left(\frac{1}{n}A^TA + (c+\rho)I_d\right)^{-1}\left(\frac{1}{n}A^T\mathbf{b} + c\mathbf{x}^t + \rho(\mathbf{x}^{\tau+1} + \mathbf{v}^{\tau})\right)$
    $\mathbf{v}^{\tau+1} = \mathbf{v}^{\tau} + \mathbf{x}^{\tau+1} - \mathbf{z}^{\tau+1}$
    $\tau = \tau + 1$
  until $\|A\mathbf{x}^{\tau} - \mathbf{b}\|_2 < \delta_1$ *or* $\dfrac{\|\mathbf{x}^{\tau} - \mathbf{z}^{\tau}\|_2}{\max\{\|\mathbf{x}^{\tau}\|_2, \|\mathbf{z}^{\tau}\|_2, \delta_2\}} < \delta_3$ *or* $\tau = \Delta_1$;
  $\mathbf{x}_{\mathcal{G}_k}^{t+1} = \mathbf{x}_{\mathcal{G}_k}^{\tau}, \ \forall k$
  $t = t + 1$
  $\tau = 0$
until $\|\mathbf{x}^t - \mathbf{x}^{t-1}\|_2 < \delta_4$ *or* $t = \Delta_2$;
Output $\hat{\mathbf{x}} = \mathbf{x}^t$.

where $\lambda_k \triangleq \lambda\frac{\sqrt{|\mathcal{G}_k|}}{\sqrt{w_k^t}}$. The closed-form solution for (3.12) is given by

$$\mathbf{x}_{\mathcal{G}_k}^{\tau+1} = S(z_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}, \lambda_k/\rho),$$

where S is referred to as a *soft-shrinkage* operator defined as

$$S(\mathbf{z}, \lambda) = \begin{cases} \left(1 - \dfrac{\lambda}{\|\mathbf{z}\|_2}\right)\mathbf{z} & \text{if } \|\mathbf{z}\|_2 > \lambda \\ 0 & \text{if } \|\mathbf{z}\|_2 \le \lambda. \end{cases}$$

Note that the shrinkage operator makes solving l_1-related problems very efficient.

For the $\mathbf{z}$-subproblem, we take the derivative of $\mathcal{L}$ with respect to $\mathbf{z}$ and the optimal solution of $\mathbf{z}^{\tau+1}$ should satisfy the following equation

$$\nabla L(\mathbf{z}) + c(\mathbf{z} - \mathbf{x}^t) + \rho(\mathbf{z} - \mathbf{x}^{\tau+1} - \mathbf{v}^{\tau}) = 0. \tag{3.13}$$

The solution of (3.13) can be obtained in different ways depending on the loss function. In this paper, we consider two types of loss functions: least squares loss and logistic loss, which are defined in Section 2. Specifically for least squares loss (2.3), the optimality condition (3.13) gives the closed-form solution for $\mathbf{z}$:

$$z^{\tau+1} = \left(\frac{1}{n}A^TA + (c+\rho)I_d\right)^{-1}\left(\frac{1}{n}A^T\mathbf{b} + c\mathbf{x}^t + \rho(\mathbf{x}^{\tau+1} + \mathbf{v}^{\tau})\right), \tag{3.14}$$

where I_d denotes the identity matrix of size $d \times d$. The overall algorithm for the least squares loss is described in Algorithm 3.1.

Algorithm 3.2 Iteratively reweighted algorithm for logistic loss

Set c, ε, ρ, and λ
Initialize $\mathbf{x}^0, \mathbf{z}^0, \mathbf{v}^0, t = 0, s = 0$
repeat
  for $k = 1, 2, \ldots, m$ **do**
    $w_k^t = \|\mathbf{x}_{\mathcal{G}_k}^t\|_2^2 + \varepsilon$
    $\lambda_k = \lambda\sqrt{|\mathcal{G}_k|/w_k^t}$
  end
  repeat
    for $k = 1, 2, \ldots, m$ **do**
      $\mathbf{x}_{\mathcal{G}_k}^{\tau+1} = S(\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}, \lambda_k/\rho)$
    end
    repeat
      update $\mathbf{z}_{s+1}$ by using (3.16)
      $s = s + 1$
    until $\mathbf{z}_s$ *converges*;
    $\mathbf{z}^{\tau+1} = \mathbf{z}_s$
    $\mathbf{v}^{\tau+1} = \mathbf{v}^{\tau} + \mathbf{x}^{\tau+1} - \mathbf{z}^{\tau+1}$
    $\tau = \tau + 1$
    $s = 0$
  until $\mathbf{x}_{\mathcal{G}_k}^{\tau}$ *converges for all* k;
  $\mathbf{x}_{\mathcal{G}_k}^{t+1} = \mathbf{x}_{\mathcal{G}_k}^{\tau}, \ \forall k$
  $t = t + 1$
  $\tau = 0$
until $\mathbf{x}^t$ *converges*;
Output $\hat{\mathbf{x}} = \mathbf{x}^t$.

There is no closed-form solution of $\mathbf{z}$ for the logistic loss, and we use the Newton's method to find the solution. In particular, we get the first and second derivatives of L^{logit} defined in (2.4) with respect to each component of $\mathbf{z}$ as follows:

$$\begin{aligned} \frac{\partial L^{\text{logit}}}{\partial z_j} &= -\frac{1}{n}\sum_{i=1}^{n}\left[b_i a_{ij} - \frac{b_i a_{ij}}{1 + e^{-b_i \mathbf{a}_i^T \mathbf{z}}}\right], \\ \frac{\partial^2 L^{\text{logit}}}{\partial z_j \partial z_k} &= \frac{1}{n}\sum_{i=1}^{n}\frac{b_i^2 a_{ij} a_{ik}}{2 + e^{b_i \mathbf{a}_i^T \mathbf{z}} + e^{-b_i \mathbf{a}_i^T \mathbf{z}}}, \end{aligned} \tag{3.15}$$

for $j, k = 1, \ldots, d$. Here $A \in \mathbb{R}^{n \times d}$ with a_{ij} as the j-th component of $\mathbf{a}_i$. The Newton's method at the s-th inner iteration is given by

$$\begin{aligned} \mathbf{z}_{s+1} = \mathbf{z}_s - \delta_s &\Big(\nabla_{\mathbf{z}}^2 L^{\text{logit}}(\mathbf{z}_s) + (c + \rho) I_d\Big)^{-1} \\ &\Big(\nabla_{\mathbf{z}} L^{\text{logit}}(\mathbf{z}_s) + c(\mathbf{z}_s - \mathbf{x}^t) + \rho(\mathbf{z}_s - \mathbf{x}^{\tau+1} - \mathbf{v}^{\tau})\Big), \end{aligned} \tag{3.16}$$

where $\delta_s > 0$ is a step size. We summarize the algorithm corresponding to the logistic loss in Algorithm 3.2.

4. Numerical Experiments. We carry out extensive experiments on the problems of linear regression and binary classification. Sections 4.2-4.3 investigate synthetic datasets with group sparse patterns simulated from least squares and logistic

regression models, respectively. In Section 4.4, we examine real data on gene expression profiling for Bardet-Biedl syndrome in mammals with a generalized additive model, where smoothing spline functions of features can be expressed as a matrix expanded in the order of the splines [19]. The learned solution is used to detect which genes are responsible for the syndrome in order to better understand the complex disease. In Section 4.5, we study clinical features of pediatric acute respiratory infections (ARI) with real data collected by World Health Organization ARI Multicentre Study [25]. Making use of the clinically-supported clustering of the features, any advances in group selection can identify significant clinical signs of ARI and guide the development of diagnostic measures.

4.1. Experimental Settings. We implement Algorithm 3.1 in MATLAB to solve group LOG with least squares loss (2.3), while using the R package `grpreg` [5] to solve the group LASSO subproblem with logistic loss (2.4) for convenience. To improve computational efficiency of our method, we use relaxed termination criteria when solving the subproblem for all the linear regression experiments. Every group LASSO iteration (3.6) can be solved quickly at a price of producing a "relaxed" solution. For more details, refer to Table 4 and related discussion on the computational cost of group LOG.

The performance of our method is compared to LASSO, group LASSO, group SCAD, group MCP, OLS (for regression only) (2.3) and the ordinary logistic regression (for classification only) (2.4). We use `CVX` to find the solutions to OLS and LASSO, while group SCAD and group MCP are implemented by `grpreg`. One iteration of Algorithm 3.1, without updating the weight, yields the solution for group LASSO. In the case of binary classification, we use `glmnet` [35] to solve for the ordinary logistic regression as well as LASSO with the logistic loss. All the other logistic regression models promoting group sparsity are solved by `grpreg`.

As for the hyperparameters involved in each method, we mainly tune λ for LASSO, group LASSO, and group LOG, while tuning η for group SCAD and group MCP. For group LOG, we fix $c = 0$[1] and choose specific values of ε for different experiments. For group SCAD and group MCP, we fix $\lambda = 1$ and use the default values of γ suggested by `grpreg`, which are $\gamma_{\text{MCP}} = 3$ and $\gamma_{\text{SCAD}} = 4$. In linear regression (Sections 4.2 and 4.4), we use a hyperparameter set, denoted by $U = \{\lambda_1, \cdots, \lambda_{50}\}$, for LASSO, group LASSO, and group LOG. Following a conventional choice of the set, we have for synthetic data in Section 4.2 that

$$\lambda_i = \begin{cases} 10^{-5}, & i = 1 \\ 5\lambda_{i-1}, & 2 \le i \le 6 \\ 1.1\lambda_{i-1}, & 7 \le i \le 44 \\ 2\lambda_{i-1}, & 45 \le i \le 50. \end{cases}$$

For real data in Section 4.4, we define $\lambda_1 = 10^{-5}$ and $\lambda_i = 1.3\lambda_{i-1}$ for $2 \le i \le 50$. For methods implemented by `grpreg`, we have a hyperparameter set of 50 values, and subsequently, the package automatically returns 50 equally spaced values on the log scale over the relevant range [5]. Group LOG regularized logistic regression uses the hyperparameter set from the first iteration with `grpreg`. The hyperparameter sets for all the methods are confirmed to cover the relevant ranges, i.e., the optimal hyperparameters are not at the edges of the ranges.

[1] We need c to be strictly positive to show convergence in Theorem 3.3, while experimentally we observe that $c = 0$ often gives satisfactory results. We use $c = 0$ throughout the experiments.

We split real data into three sets of data: training set, validation set, and test set. For synthetic data, we only generate a training set and a validation set, while evaluating the performance by comparing to the ground-truth. For linear regression, the optimal hyperparameter is chosen based on the least squared error on the validation set. In binary classification (Sections 4.3 and 4.5), the hyperparameter sets for λ and η are obtained by `glmnet` and `grpreg`, respectively. We select the best λ and η that produces the largest area under the curve (AUC) on the validation set for synthetic data in Section 4.3 and smallest classification error for real data in Section 4.5. The AUC and classification error are defined in the respective sections.

4.2. Synthetic data experiments for linear regression. To test the performance on the linear regression problem, we generate a feature matrix $A \in \mathbb{R}^{100\times 200}$ from the standard Gaussian distribution such that $a_{ij} \overset{\text{iid}}{\sim} N(0,1)$, $j = 1,\ldots,200$ for $i = 1,\ldots,100$. A ground-truth vector is $\mathbf{x}^* = (x_1^*,\ \ldots,\ x_{200}^*)^T \in \mathbb{R}^{200}$. We introduce $\bar{A} \triangleq [\mathbf{1},\ A] \in \mathbb{R}^{100\times 201}$, where $\mathbf{1} \triangleq (1,\ldots,1)^T \in \mathbb{R}^{100}$ for the intercept term x_0^* to capture the center of data. A response vector $\mathbf{b} \in \mathbb{R}^{100}$ is generated such that

$$\mathbf{b} = \bar{A}\bar{\mathbf{x}}^* + \mathbf{e}, \tag{4.1}$$

where $\bar{\mathbf{x}}^* \triangleq (x_0^*,\ \mathbf{x}^{*T})^T \in \mathbb{R}^{201}$ and $\mathbf{e}$ denotes the Gaussian noise such that $e_i \sim N(0, \alpha^2\sigma_A^2)$ with σ_A being the standard deviation of $\bar{A}\bar{\mathbf{x}}^*$ and α being the noise level. We vary the noise level being $\alpha = 0.1, 0.2, 0.3$ for three experimental scenarios from less noisy to more noisy cases.

We let $x_0^* = 0$ and make $\mathbf{x}^*$ have 20 groups each having 10 coefficients. Among the 20 groups of homogeneous size of ten, 13 have nonzero coefficients that are uniformly distributed such that $x_j^* \overset{iid}{\sim} U[-5,5]$ and the other 7 have zero coefficients, i.e., $x_j^* = 0$. We randomly generate 50 ground-truth vectors with corresponding training/validation sets. For each ground-truth $\mathbf{x}^*$, we generate a set of training instances $(A_{tr}, \mathbf{b}_{tr}) \in \mathbb{R}^{100\times 200} \times \mathbb{R}^{100}$ and a validation set $(A_v, \mathbf{b}_v) \in \mathbb{R}^{100\times 200} \times \mathbb{R}^{100}$. We fix $\varepsilon = 0.001$ in this experiment and choose the best value of λ in the hyperparameter set U that gives the smallest squared error $||\mathbf{b}_v - \hat{x}_0\mathbf{1} - A_v\hat{\mathbf{x}}||_2^2$ on the validation set, where $\hat{\mathbf{x}} \in \mathbb{R}^{100}$ denotes the reconstructed solution with the optimal hyperparameter, and $\hat{x}_0$ is the estimated intercept. Note that, for numerical experiments, any learned model does not regularize the intercept and we do not measure the errors between x_0^* and $\hat{x}_0$ either.

We consider five metrics to quantitatively evaluate the performance of a reconstructed solution $\hat{\mathbf{x}}$ of each method. Without loss of generality, we assume that the grouped members of $\mathbf{x}^*$ are ordered by $\mathcal{G}_1,\ldots,\mathcal{G}_{m'}$ $(m' < m)$ that have nonzero coefficients, whereas $\mathcal{G}_{m'+1},\ldots,\mathcal{G}_m$ have zero coefficients. The support of $\mathbf{x}^*$ and the support recovered by $\hat{\mathbf{x}}$ are denoted as $\mathcal{S}$ and $\hat{\mathcal{S}}$, respectively. We present a confusion matrix in Table 2 that is used to compute the following metrics for assessment:

1. Relative error of $\hat{\mathbf{x}} \triangleq \dfrac{\|\hat{\mathbf{x}} - \mathbf{x}^*\|_2}{\|\mathbf{x}^*\|_2}$.
2. Precision of $\hat{\mathbf{x}} \triangleq \dfrac{|\hat{\mathcal{S}} \cap \mathcal{S}|}{|\hat{\mathcal{S}}|}$.
3. Recall of $\hat{\mathbf{x}} \triangleq \dfrac{|\hat{\mathcal{S}} \cap \mathcal{S}|}{|\mathcal{S}|}$.
4. Element accuracy of $\hat{\mathbf{x}} \triangleq \dfrac{|\hat{\mathcal{S}} \cap \mathcal{S}| + |\hat{\mathcal{S}}^c \cap \mathcal{S}^c|}{d}$, with $\hat{\mathbf{x}} \in \mathbb{R}^d$.

	$\mathcal{S} \triangleq \{j : x_j^* \neq 0\}$	$\mathcal{S}^c \triangleq \{j : x_j^* = 0\}$
Positive set $\hat{\mathcal{S}} \triangleq \{j : \hat{x}_j \neq 0\}$	True positives $\hat{\mathcal{S}} \cap \mathcal{S}$	False positives $\hat{\mathcal{S}} \cap \mathcal{S}^c$
Negative set $\hat{\mathcal{S}}^c \triangleq \{j : \hat{x}_j = 0\}$	False negatives $\hat{\mathcal{S}}^c \cap \mathcal{S}$	True negatives $\hat{\mathcal{S}}^c \cap \mathcal{S}^c$

TABLE 2

A confusion matrix for comparing group sparse patterns in a reconstructed solution $\hat{\mathbf{x}}$ to those in the ground-truth $\mathbf{x}^*$.

Method	Relative error	Precision	Recall	Accuracy	Group accuracy
$\alpha = 0.1$					
OLS	0.7045 (0.0517)	0.3500 (0.0000)	1.0000 (0.0000)	0.3500 (0.0000)	0.3500 (0.0000)
LASSO	0.6641 (0.0960)	0.5311 (0.0621)	0.6689 (0.1069)	0.6660 (0.0761)	0.0300 (0.0440)
Group LASSO	0.3517 (0.0677)	0.3862 (0.0398)	1.0000 (0.0000)	0.4360 (0.0783)	0.4360 (0.0783)
Group SCAD	0.4041 (0.1939)	0.9763 (0.0714)	0.9886 (0.0391)	0.9850 (0.0368)	0.9850 (0.0368)
Group MCP	0.4444 (0.2401)	0.9905 (0.0382)	0.9000 (0.1163)	0.9620 (0.0458)	0.9620 (0.0458)
Group LOG	0.1613 (0.0288)	0.9751 (0.0671)	1.0000 (0.0000)	0.9890 (0.0308)	0.9890 (0.0308)
$\alpha = 0.2$					
OLS	0.7508 (0.0369)	0.3500 (0.0000)	1.0000 (0.0000)	0.3500 (0.0000)	0.3500 (0.0000)
LASSO	0.7259 (0.0700)	0.5097 (0.0475)	0.6426 (0.0827)	0.6538 (0.0468)	0.0250 (0.0368)
Group LASSO	0.4862 (0.0814)	0.3871 (0.0316)	0.9971 (0.0202)	0.4410 (0.0698)	0.4410 (0.0698)
Group SCAD	0.5061 (0.1422)	0.9386 (0.1022)	0.9771 (0.0728)	0.9640 (0.0572)	0.9640 (0.0572)
Group MCP	0.5481 (0.1706)	0.9927 (0.0364)	0.8886 (0.1332)	0.9590 (0.0512)	0.9590 (0.0512)
Group LOG	0.3142 (0.0625)	0.9753 (0.0705)	0.9943 (0.0283)	0.9870 (0.0347)	0.9870 (0.0347)
$\alpha = 0.3$					
OLS	0.7404 (0.0471)	0.3500 (0.0000)	1.0000 (0.0000)	0.3500 (0.0000)	0.3500 (0.0000)
LASSO	0.7345 (0.0798)	0.5258 (0.0634)	0.6386 (0.0999)	0.6631 (0.0555)	0.0430 (0.0639)
Group LASSO	0.4761 (0.0729)	0.3903 (0.0310)	1.0000 (0.0000)	0.4480 (0.0670)	0.4480 (0.0670)
Group SCAD	0.5180 (0.1558)	0.9152 (0.1108)	0.9743 (0.0625)	0.9540 (0.0579)	0.9540 (0.0579)
Group MCP	0.5817 (0.1957)	0.9852 (0.0533)	0.8771 (0.1291)	0.9520 (0.0525)	0.9520 (0.0525)
Group LOG	0.3010 (0.0500)	0.9875 (0.0379)	1.0000 (0.0000)	0.9950 (0.0152)	0.9950 (0.0152)

TABLE 3

Results for synthetic regression datasets with homogeneous group sizes. The averages are presented along with their standard deviations in parentheses.

5. Group accuracy of $\hat{\mathbf{x}} \triangleq \dfrac{|\mathcal{M}_{tp}| + |\mathcal{M}_{tn}|}{m}$, where

$$\text{True positive groups of } \hat{\mathbf{x}} \ (\mathcal{M}_{tp}) \triangleq \{k \in \{1, \ldots, m'\} : \mathcal{G}_k \subset \hat{\mathcal{S}}\}$$

$$\text{True negative groups of } \hat{\mathbf{x}} \ (\mathcal{M}_{tn}) \triangleq \{k \in \{m'+1, \ldots, m\} : \mathcal{G}_k \subset \hat{\mathcal{S}}^c\}.$$

The relative error computes the distance between the solution and the ground-truth vector normalized by the l_2 norm of the ground-truth vector. The precision is the fraction of true positives among the nonzero indices in $\hat{\mathbf{x}}$, while the recall is the fraction of true positives among the nonzero indices in $\mathbf{x}^*$. The element accuracy is the fraction of the true positives and true negatives among all the member indices. A group having all true positives or all true negatives is considered as a group that makes success in sparsity recovery. Thus, the group accuracy is the fraction of the groups successfully recovered among all groups.

We report the average values and the standard deviations of the above metrics over 50 training/validation sets in Table 3. The proposed group LOG produces the smallest relative errors, highest accuracy and group accuracy. Specifically, it is worth noting that the group accuracy of group LOG is always larger than 98% under all

Method	Relative error	Accuracy	Group accuracy	Time	Outer iter.	Total iter.
Group LASSO	0.3517 (0.0677)	0.4360 (0.0783)	0.4360 (0.0783)	1.02	NA	948.82
Group LOG (exact)	0.1465 (0.0233)	0.9950 (0.0152)	0.9950 (0.0152)	10.97	10.36	10,088.62
Group LOG (relaxed)	0.1613 (0.0288)	0.9890 (0.0308)	0.9890 (0.0308)	0.50	9.24	462.00

TABLE 4

Comparison of accuracy, wall time and number of iterations for synthetic regression datasets with homogeneous group sizes and $\alpha = 0.1$. The averages are presented along with their standard deviations in parentheses.

the three noise levels, while LASSO suffers from low group accuracy. Each group regularizer such as group LOG and group LASSO produces same accuracy and group accuracy values in such a homogeneous group setting, where the number of correctly recovered coefficients is a product of the group size and the number of correctly recovered groups. This is also observed in Table 6. Our approach and group MCP in Table 3 show great performance in terms of precision. The highest recall values are achieved by OLS, group LASSO, and group LOG. Since OLS does not perform variable selection and produce nonzero coefficients, their recall values are always 1 with poor precision values. Although the trade-off between precision and recall tends to hinder simultaneous achievement of high precision and high recall, our approach shows very satisfactory results of a balanced performance across the five metrics, in comparison to the other methods.

The group LOG results from Table 3 are obtained with relaxed termination criteria in Algorithm 3.1: $\delta_1 = \delta_3 =$ `1e-1`, $\delta_2 =$ `1e-16`, $\delta_4 =$ `1e-6` and $\Delta_1 = 50$, $\Delta_2 = 15$. A more accurate solution can be obtained by setting $\Delta_1 = 1000$, with much higher computation costs. By decreasing the number of iterations solved within the subproblem, the relaxed scheme significantly reduces the computational time. Table 4 compares the performance and computational efficiency of group LASSO, exact group LOG and relaxed group LOG on the simulated datasets with $\alpha = 0.1$. With the maximum iteration number $\Delta_1 = 1000$ for group LASSO, the average iteration number is 948.82. Compared to group LASSO, the exact group LOG using the same strict termination criteria for the subproblem has ten-fold increase in the time spent. The relaxed setting of small inner iterations $\Delta_1 = 50$ maintains performance accuracy, reducing the computational time to nearly half of the time used by group LASSO.

We generate another set of synthetic data with heterogeneous group sizes, as opposed to an equal number of members in each group. We still consider 20 groups, but with group sizes of $[2, 3, 10, 15, 20, 2, 3, 10, 15, 20, 2, 3, 10, 15, 20, 2, 3, 10, 15, 20]$ by setting $|\mathcal{G}_{1+5\cdot k}| = 2$, $|\mathcal{G}_{2+5\cdot k}| = 3$, $|\mathcal{G}_{3+5\cdot k}| = 10$, $|\mathcal{G}_{4+5\cdot k}| = 15$, and $|\mathcal{G}_{5+5\cdot k}| = 20$ for $k = 0,\ 1,\ 2,\ 3$. We then randomly select 7 nonzero groups, whose corresponding coefficients in $\mathbf{x}^*$ are generated by a continuous uniform distribution between -5 and 5. All the coefficients in the rest of 13 groups are set as zero. Table 5 presents the results from the experiments with the heterogeneous group sizes. We can observe the same behavior as in Table 3 that our approach is the best in terms of a balanced performance on all the metrics. Due to the imbalanced sparsity across the groups, it is more difficult to maintain a high group accuracy than for the homogeneous case. As a result, our approach gives slightly worse group accuracy in Table 5 than in Table 3, still with significant improvements over the alternative methods. Group MCP and SCAD do not perform well at a higher noise level of $\alpha = 0.3$.

4.3. Synthetic data experiments for binary classification. In this experiment, we generate a data matrix $\bar{A}$, and a coefficient sequence $\bar{\mathbf{x}}^*$ in the same manner as in Section 4.2 with both homogeneous and heterogeneous group sizes. A response

Method	Relative error	Precision	Recall	Accuracy	Group accuracy
$\alpha = 0.1$					
OLS	0.7107 (0.0590)	0.3359 (0.0644)	1.0000 (0.0000)	0.3359 (0.0644)	0.3500 (0.0000)
LASSO	0.6514 (0.1311)	0.5050 (0.0747)	0.6911 (0.1216)	0.6598 (0.0637)	0.1660 (0.0738)
Group LASSO	0.3493 (0.1165)	0.3667 (0.0656)	0.9987 (0.0064)	0.4184 (0.0861)	0.4970 (0.0860)
Group SCAD	0.3609 (0.1159)	0.3520 (0.0720)	0.9993 (0.0046)	0.3774 (0.0881)	0.4230 (0.0893)
Group MCP	0.3625 (0.1146)	0.3481 (0.0696)	0.9993 (0.0046)	0.3674 (0.0859)	0.3950 (0.0764)
Group LOG	0.2278 (0.1046)	0.9760 (0.0693)	0.9753 (0.0238)	0.9796 (0.0353)	0.9540 (0.0359)
$\alpha = 0.2$					
OLS	0.7343 (0.0508)	0.3478 (0.0754)	1.0000 (0.0000)	0.3478 (0.0754)	0.3500 (0.0000)
LASSO	0.7049 (0.1096)	0.5192 (0.0832)	0.6311 (0.1356)	0.6613 (0.0659)	0.1800 (0.1035)
Group LASSO	0.4495 (0.1012)	0.3890 (0.0880)	0.9971 (0.0089)	0.4442 (0.1159)	0.5030 (0.0883)
Group SCAD	0.7236 (0.3279)	0.8063 (0.1611)	0.5272 (0.3953)	0.7806 (0.1607)	0.7570 (0.1270)
Group MCP	0.7534 (0.3209)	0.7996 (0.1684)	0.4919 (0.4078)	0.7714 (0.1629)	0.7530 (0.1251)
Group LOG	0.3983 (0.0956)	0.9517 (0.0951)	0.9445 (0.0665)	0.9589 (0.0461)	0.9230 (0.0419)
$\alpha = 0.3$					
OLS	0.7814 (0.0458)	0.3380 (0.0792)	1.0000 (0.0000)	0.3380 (0.0792)	0.3500 (0.0000)
LASSO	0.7414 (0.0880)	0.5103 (0.0805)	0.6181 (0.1215)	0.6638 (0.0632)	0.1720 (0.0743)
Group LASSO	0.5210 (0.0853)	0.3897 (0.0835)	0.9955 (0.0118)	0.4667 (0.1078)	0.5260 (0.0933)
Group SCAD	1.0000 (0.0281)	0.5083 (0.1935)	0.0852 (0.1516)	0.6648 (0.0912)	0.6480 (0.0589)
Group MCP	1.0026 (0.0281)	0.5162 (0.2403)	0.0746 (0.1544)	0.6658 (0.0889)	0.6520 (0.0553)
Group LOG	0.5098 (0.0873)	0.9118 (0.1647)	0.9243 (0.0560)	0.9328 (0.0900)	0.8870 (0.0653)

TABLE 5

Results for synthetic regression datasets with heterogeneous group sizes. The averages are presented along with their standard deviations in parentheses.

vector $\mathbf{b}$ is generated from a Bernoulli distribution such that the probability of the i-th response being 1 is

$$P(b_i = 1) = \frac{1}{1 + e^{-\bar{\mathbf{a}}_i \bar{\mathbf{x}}^*}}, \quad i = 1, \ldots, 100. \tag{4.2}$$

where $\bar{\mathbf{a}}_i$ being the i-th row of $\bar{A}$. We generate 50 training/validation sets corresponding to the same sequence $\bar{\mathbf{x}}^*$. The noise level is represented by the empirical Bayes risk r based on 100 observations:

$$r \triangleq \frac{1}{100} \sum_{i=1}^{100} \min \left(\frac{1}{1 + e^{-\bar{\mathbf{a}}_i \bar{\mathbf{x}}^*}}, \frac{1}{1 + e^{\bar{\mathbf{a}}_i \bar{\mathbf{x}}^*}} \right).$$

By scaling $\bar{\mathbf{x}}^*$ to $0.5\bar{\mathbf{x}}^*$ and $0.25\bar{\mathbf{x}}^*$, we can evaluate the performance under scenarios with different risk values of $r = 0.0092, 0.0236, 0.0610$, respectively [32]. A higher risk value of r corresponds to more noisy data that makes a learning task more challenging. We set $\varepsilon = 0.001$ for our method in the first scenario with $\bar{\mathbf{x}}^*$, and $\varepsilon = 0.1$ in the other two scenarios. We rely on the receiver operating characteristic (ROC) curve to tune the hyperparameter. An ROC curve plots true positive rate (TPR) versus false positive rate (FPR) at different thresholds of the estimated probability (4.2) for an individual observation [20]. We find the optimal hyperparameter that produces the largest area under the ROC curve (AUC) on the validation set, which corresponds to the best combination of high TPR and low FPR regardless of the thresholding values.

Tables 6-7 present the results of binary classification for the homogeneous and heterogeneous group sizes, respectively, using the measures introduced in Section 4.2 to evaluate the performance. The proposed approach is significantly better than the alternative methods in terms of relative errors, accuracy, and group accuracy. Our solutions yield precision and recall values above 0.7 in all cases, whereas the other approaches perform poorly in terms of either precision or recall. Furthermore, our ap-

Method	Relative error	Precision	Recall	Accuracy	Group accuracy
$r = 0.0092$					
Logistic regression	0.9756 (0.0555)	0.3500 (0.0000)	1.0000 (0.0000)	0.3500 (0.0000)	0.3500 (0.0000)
LASSO	0.9515 (0.0264)	0.5217 (0.1232)	0.3600 (0.1219)	0.6432 (0.0312)	0.1110 (0.1756)
Group LASSO	0.9588 (0.0204)	0.4932 (0.0923)	0.9286 (0.1392)	0.6141 (0.1224)	0.6140 (0.1225)
Group SCAD	0.9436 (0.0548)	0.7105 (0.1543)	0.7571 (0.1876)	0.7840 (0.0798)	0.7840 (0.0798)
Group MCP	0.9250 (0.0619)	0.8518 (0.1728)	0.5943 (0.1740)	0.8190 (0.0856)	0.8190 (0.0856)
Group LOG	0.9036 (0.0293)	0.8407 (0.1518)	0.7314 (0.2172)	0.8450 (0.0865)	0.8450 (0.0865)
$r = 0.0236$					
Logistic regression	1.3618 (0.1053)	0.3500 (0.0000)	1.0000 (0.0000)	0.3500 (0.0000)	0.3500 (0.0000)
LASSO	0.9274 (0.0381)	0.5285 (0.1154)	0.3534 (0.1150)	0.6489 (0.0339)	0.1280 (0.1582)
Group LASSO	0.9224 (0.0306)	0.5058 (0.1064)	0.9429 (0.0913)	0.6300 (0.1169)	0.6300 (0.1169)
Group SCAD	0.9298 (0.0575)	0.6839 (0.1526)	0.7943 (0.1504)	0.7830 (0.1008)	0.7830 (0.1008)
Group MCP	0.9107 (0.0811)	0.8218 (0.1647)	0.5914 (0.1384)	0.8080 (0.0791)	0.8080 (0.0791)
Group LOG	0.8542 (0.0589)	0.8391 (0.1799)	0.7400 (0.1993)	0.8360 (0.1005)	0.8360 (0.1005)
$r = 0.0610$					
Logistic regression	2.5882 (0.1892)	0.3500 (0.0000)	1.0000 (0.0000)	0.3500 (0.0000)	0.3500 (0.0000)
LASSO	0.9431 (0.0421)	0.5281 (0.1233)	0.3291 (0.1287)	0.6457 (0.0339)	0.1450 (0.1745)
Group LASSO	0.8991 (0.0514)	0.5594 (0.1480)	0.8771 (0.1756)	0.6670 (0.1304)	0.6670 (0.1304)
Group SCAD	0.9632 (0.1023)	0.6791 (0.1730)	0.7771 (0.1532)	0.7670 (0.1008)	0.7670 (0.1008)
Group MCP	1.0273 (0.1936)	0.7991 (0.1955)	0.5543 (0.1545)	0.7880 (0.0901)	0.7880 (0.0901)
Group LOG	0.8401 (0.0768)	0.8231 (0.1955)	0.7029 (0.1495)	0.8220 (0.1041)	0.8220 (0.1041)

TABLE 6

Results for synthetic logistic datasets with homogeneous group sizes. The averages are presented along with their standard deviations in parentheses. The average empirical Bayes risk value, r, is annotated for each scenario.

Method	Relative error	Precision	Recall	Accuracy	Group accuracy
$r = 0.0168$					
Logistic regression	0.9903 (0.0786)	0.3439 (0.0829)	1.0000 (0.0000)	0.3439 (0.0829)	0.3500 (0.0000)
LASSO	0.9478 (0.0292)	0.5213 (0.1356)	0.3728 (0.1205)	0.6526 (0.0528)	0.2020 (0.1340)
Group LASSO	0.9545 (0.0207)	0.5328 (0.1563)	0.9403 (0.1085)	0.6567 (0.1458)	0.6520 (0.1286)
Group SCAD	0.9279 (0.0848)	0.7541 (0.1774)	0.7958 (0.1692)	0.8135 (0.1083)	0.7740 (0.1065)
Group MCP	0.9157 (0.0844)	0.8977 (0.1293)	0.6284 (0.1973)	0.8376 (0.0909)	0.7990 (0.0836)
Group LOG	0.8806 (0.0554)	0.8823 (0.1280)	0.8333 (0.1719)	0.8910 (0.0899)	0.8320 (0.0885)
$r = 0.0371$					
Logistic regression	1.4067 (0.2025)	0.3439 (0.0829)	1.0000 (0.0000)	0.3439 (0.0829)	0.3500 (0.0000)
LASSO	0.9250 (0.0416)	0.5139 (0.1016)	0.3693 (0.1004)	0.6567 (0.0559)	0.2150 (0.1041)
Group LASSO	0.9257 (0.0293)	0.5287 (0.1439)	0.9583 (0.0631)	0.6644 (0.1389)	0.6630 (0.1224)
Group SCAD	0.8977 (0.1177)	0.7504 (0.1652)	0.8987 (0.1046)	0.8360 (0.1021)	0.7830 (0.0896)
Group MCP	0.8683 (0.1301)	0.8590 (0.1755)	0.6528 (0.2016)	0.8359 (0.1137)	0.7900 (0.0904)
Group LOG	0.7981 (0.0907)	0.8732 (0.1369)	0.8405 (0.1810)	0.8853 (0.0919)	0.8310 (0.0820)
$r = 0.0765$					
Logistic regression	2.6696 (0.5039)	0.3439 (0.0829)	1.0000 (0.0000)	0.3439 (0.0829)	0.3500 (0.0000)
LASSO	0.9186 (0.0455)	0.5181 (0.1181)	0.3597 (0.1155)	0.6556 (0.0575)	0.2030 (0.1263)
Group LASSO	0.8800 (0.0580)	0.5167 (0.1479)	0.9068 (0.1370)	0.6460 (0.1162)	0.6430 (0.0979)
Group SCAD	0.9409 (0.1138)	0.7020 (0.1772)	0.8241 (0.1481)	0.7914 (0.1062)	0.7390 (0.0791)
Group MCP	1.0063 (0.2089)	0.8436 (0.1586)	0.6501 (0.1998)	0.8275 (0.1157)	0.7770 (0.0828)
Group LOG	0.8013 (0.0960)	0.8541 (0.2007)	0.7697 (0.1868)	0.8535 (0.1298)	0.7990 (0.0889)

TABLE 7

Results for synthetic logistic datasets with heterogeneous group sizes. The averages are presented along with their standard deviations in parentheses. The average empirical Bayes risk value, r, is annotated for each scenario.

proach works in a reasonable manner with a higher empirical Bayes risk corresponding to a more difficult learning problem.

4.4. Genomic data experiments for Bardet-Biedl syndrome. We consider a generalized additive model that associates the expression level of TRIM32, a gene responsible for Bardet-Biedl syndrome [11], with univariate smooth functions of 20 gene expression levels [19]. Each of the 20 smooth functions is represented by five B-Spline basis functions, forming a group. The feature matrix for the additive model has 120 observations with the 100 features, each group of 5 consecutive features belonging

Method	MSEP	Coefficient selection rate	Group selection rate
OLS	2.6024 (7.4440)	1.0000 (0.0000)	1.0000 (0.0000)
LASSO	0.0242 (0.0203)	0.1974 (0.1932)	0.5650 (0.3280)
Group LASSO	0.0231 (0.0197)	0.5208 (0.3288)	0.5210 (0.3289)
Group SCAD	0.0625 (0.1506)	0.2520 (0.1705)	0.2520 (0.1705)
Group MCP	0.1523 (0.4291)	0.1280 (0.0809)	0.1280 (0.0809)
Group LOG	0.0235 (0.0166)	0.2030 (0.1899)	0.2030 (0.1899)

TABLE 8

Results from gene expression data experiments for Bardet-Biedl syndrome. MSEP stands for mean squared error of prediction. The averages based on 50 repetitions are presented along with their standard deviations in parentheses.

to one gene. The gene expression dataset [42] is from the microarray experiments on eye tissues harvested from 120 male rats, which is publicly available in the R package `gglasso` [51]. An important task associated with the dataset is to examine the regulation of the genes related with the Bardet-Biedl syndrome, a genetic disease that may cause progressive visual impairment, kidney abnormalities, learning difficulties, etc [17].

We randomly select 80 observations for training $(A_{tr}, \mathbf{b}_{tr}) \in \mathbb{R}^{80\times 100} \times \mathbb{R}^{80}$, 10 observations for validation $(A_v, \mathbf{b}_v) \in \mathbb{R}^{10\times 100} \times \mathbb{R}^{10}$, and the remaining 30 for testing $(A_{test}, \mathbf{b}_{test}) \in \mathbb{R}^{30\times 100} \times \mathbb{R}^{30}$. Fixing $\varepsilon = 0.001$, we train the model, tune the optimal hyperparameter, and report the averages together with standard deviations of the following three measures based on 50 random separations of the data:

1. Mean squared error of prediction (MSEP) $\triangleq \|\mathbf{b}_{test} - \hat{x}_0\mathbf{1} - A_{test}\hat{\mathbf{x}}\|_2^2$.
2. Coefficient selection rate $\triangleq \dfrac{1}{d}\displaystyle\sum_{j=1}^{d} I(\hat{x}_j \neq 0)$, where I is the indicator function.
3. Group selection rate $\triangleq \dfrac{1}{m}\displaystyle\sum_{k=1}^{m} I(\hat{\mathbf{x}}_{\mathcal{G}_k} \neq 0)$, where $\hat{\mathbf{x}}_{\mathcal{G}_k}$ is a coefficient subsequence corresponding to the k-th group.

MSEP is computed based on the test set. The coefficient selection rate is the average proportion of nonzero coefficients in $\hat{\mathbf{x}}$, and the group selection rate is the average ratio of groups having at least one nonzero coefficient to the total groups. The results are summarized in Table 8, showing that our approach has the highest MSEP. Based the coefficient and group selection rates, our method selects 20 ($\approx 100 \cdot 0.2032$) nonzero coefficients and 4 ($\approx 20 \cdot 0.2050$) groups on average with at least one nonzero element. On the other hand, group LASSO, and group SCAD select a larger number of nonzero coefficients and groups, whereas group MCP tends to produce a smaller number. LASSO does not enhance a group of zero coefficients, as shown by its low coefficient selection rate but high group selection rate. Although the correct number of nonzero groups is unknown for the real data, the proposed method generally gives a good compromise in comparison to the other methods considered, avoiding extreme sparse patterns.

4.5. Clinical data experiments for pediatric pneumonia. This experiment examines clustered clinical signs to predict pediatric pneumonia, an acute respiratory infection (ARI) that affects lungs of young infants. According to [41], ARI may be the main cause of mortality of infants under 3 months of age in developing countries, thus a successful classifier may be used to diagnose and treat ARI of young infants at a low cost.

We adopt a clinical data that is collected in Ethiopia by World Health Organi-

Method	Test Error	Coefficient selection rate	Group selection rate
Logistic regression	0.2886 (0.0950)	0.9688 (0.0120)	1.0000 (0.0000)
LASSO	0.1800 (0.0973)	0.1790 (0.1288)	0.4013 (0.2835)
Group LASSO	0.1829 (0.0709)	0.3431 (0.2977)	0.3175 (0.2829)
Group SCAD	0.1733 (0.0744)	0.1119 (0.1145)	0.0975 (0.1019)
Group MCP	0.1733 (0.0780)	0.0888 (0.0984)	0.0800 (0.0911)
Group LOG	0.1714 (0.0860)	0.1634 (0.2323)	0.1475 (0.2153)

TABLE 9

Results from clinical data experiments for ARI in young infants. The averages and the standard deviations are computed based on 50 repetitions.

zation ARI (WHO/ARI) Multicentre Study and provided in the R package `hdrm` [6]. The data studies 816 infants, among which we restrict our interest to 116 with positive nutrition scores, meaning those who received adequate nutrition. Based on the agreement made by clinicians who participated in the WHO/ARI study [18], the dataset is processed to have 59 clinical signs, exclusively belonging to each of 16 groups with different sizes; we present the clinical signs in Appendix A. We then use the data to predict whether an infant with certain clinical signs has pneumonia or not. We randomly split 116 infants into three sets: 75 for training $(A_{tr}, \mathbf{b}_{tr}) \in \mathbb{R}^{75\times 59}\times\mathbb{R}^{75}$, 20 for validation $(A_v, \mathbf{b}_v) \in \mathbb{R}^{20\times 59} \times \mathbb{R}^{20}$, and 21 for testing $(A_{test}, \mathbf{b}_{test}) \in \mathbb{R}^{21\times 59} \times \mathbb{R}^{21}$ For a dataset $(A, \mathbf{b}) \in \mathbb{R}^{q\times d} \times \mathbb{R}^q$ with $\mathbf{a}_i$ being the i-th row of A, we define the classification error on $(A, \mathbf{b})$ as

$$\frac{1}{2\cdot q}\sum_{i=1}^{q}\left|2\cdot I\left(\frac{1}{1+e^{-\hat{x}_0-\mathbf{a}_i\hat{\mathbf{x}}}}\geq 0.5\right)-1-b_i\right|,$$

where $\hat{\mathbf{x}}$ is a reconstructed solution and $\hat{x}_0$ is the estimated intercept, based on which we can define a classifier. With $\varepsilon = 0.0001$, we choose the optimal hyperparameter that produces a classifier to achieve the smallest classification error on the validation set $(A_v, \mathbf{b}_v)$.

In addition to the coefficient and group selection rates used in Section 4.4, we include the classification error on the test set $(A_{test}, \mathbf{b}_{test})$ to evaluate the prediction performance. The results in Table 9 are based on 50 random separations of the data, indicating that our approach achieves the smallest test errors on average. As for coefficient and group selection rates to infer sparse patterns in the recovered signals, our approach gives 9 ($\approx 0.1634 \cdot 59$) nonzero clinical signs and 2 ($\approx 0.1475 \cdot 16$) clusters with at least one nonzero clinical signs on average. The clustering of the clinical signs is ignored in both logistic regression method and LASSO. Similar to Section 4.4, the recovered coefficients by group LOG have more zeros than those convex methods (LASSO and group LASSO), while less than those nonconvex ones (group MCP and group SCAD). The proposed regularization can be considered as a good balance between the existing convex and nonconvex approaches.

5. Conclusion. We introduced a novel log-composite regularizer to promote group structured sparsity in supervised learning problems. Though the proposed regularizer is nonconvex and nondifferentiable, we adopted an efficient algorithm that iteratively solves a convex program with larger weights to penalize zero-groups, computing the weights by the previous iterate. We theoretically demonstrated that the iterates of the algorithm converge to a stationary point. We further showed that the stationary point is a global minimizer of the objective function that is composed of the regularization and a loss function under some assumptions. We conducted comprehensive experiments on synthetic and real data for two specific applications of linear

regression and binary classification. We evaluated the performance based on five metrics for the synthetic data and three metrics for the real data. We demonstrated that the proposed approach generally outperforms the state-of-the-art methods; specifically worth noting is that it often exhibits a well-balanced performance in comparison to the existing convex and nonconvex methods.

REFERENCES

[1] M. Ahn, *Consistency bounds and support recovery of d-stationary solutions of sparse sample average approximations*, Journal of Global Optimization, (2019), pp. 1–26.

[2] M. Ahn, J. S. Pang, and J. Xin, *Difference-of-convex learning: directional stationarity, optimality, and sparsity*, SIAM Journal on Optimization, 27 (2017), pp. 1637–1665.

[3] S. Bakin, *Adaptive regression and model selection in data mining problems*, PhD thesis, The Australian National University, 1999.

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

[5] P. Breheny and J. Huang, *Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors*, Statistics and Computing, 25 (2015), pp. 173–187.

[6] P. Breheny and J. Huang, *High-Dimensional Regression Modeling: Methodology, Applications, and Software*, Texts in Statistical Science, Chapman and Hall/CRC, 2019.

[7] E. J. Candès, J. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics, 59 (2006), pp. 1207–1223.

[8] E. J. Candès and T. Tao, *Near-optimal signal recovery from random projections: Universal encoding strategies?*, IEEE Transactions on Information Theory, 52 (2006), pp. 5406–5425.

[9] E. J. Candés, M. B. Wakin, and S. Boyd, *Enhancing sparsity by reweighted l1 minimization*, Journal of Fourier Analysis and Applications, 14 (2008), pp. 877–905.

[10] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Journal on Scientific Computing, 20 (1998), pp. 33–61.

[11] A. P. Chiang, J. S. Beck, H.-J. Yen, M. K. Tayeh, T. E. Scheetz, R. E. Swiderski, D. Y. Nishimura, T. A. Braun, K.-Y. A. Kim, J. Huang, K. Elbedour, R. Carmi, D. C. Slusarski, T. L. Casavant, E. M. Stone, and V. C. Sheffield, *Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet–Biedl syndrome gene (BBS11)*, Proceedings of the National Academy of Sciences, 103 (2006), pp. 6287–6292.

[12] F. E. Curtis, Y. Dai, and D. P. Robinson, *A subspace acceleration method for minimization involving a group sparsity-inducing regularizer*, SIAM Journal on Optimization, (2020).

[13] W. Deng, W. Yin, and Y. Zhang, *Group sparse optimization by alternating direction method*, in Wavelets and Sparsity XV, vol. 8858, SPIE, 2013, pp. 242 – 256.

[14] D. L. Donoho, *Compressed sensing*, IEEE Transactions on Information Theory, 52 (2006), pp. 1289–1306.

[15] D. L. Donoho and X. Huo, *Uncertainty principles and ideal atomic decomposition*, IEEE Transactions on Information Theory, 47 (2001), pp. 2845–2862.

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

[17] E. Forsythe and P. L. Beales, *Bardet–Biedl syndrome*, European Journal of Human Genetics, 21 (2013), pp. 8–13.

[18] F. E. Harrell Jr, P. A. Margolis, S. Gove, K. E. Mason, E. K. Mulholland, D. Lehmann, L. Muhe, S. Gatchalian, and H. F. Eichenwald, *Development of a clinical prediction model for an ordinal outcome: the world health organization multicentre study of clinical signs and etiological agents of pneumonia, sepsis and meningitis in young infants*, Statistics in Medicine, 17 (1998), pp. 909–944.

[19] T. Hastie and R. Tibshirani, *Generalized additive models*, vol. 43 of Monographs on Statistics and Applied Probability, Chapman and Hall/CRC, 1990.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer Series in Statistics, Springer Science & Business Media, 2009.

[21] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The*

Lasso and Generalizations, vol. 143 of Monographs on Statistics and Applied Probability, Chapman and Hall/CRC, 2015.

[22] J. Huang, P. Breheny, and S. Ma, *A selective review of group selection in high-dimensional models*, Statistical Science, 27 (2012), pp. 481–499.

[23] D. R. Hunter and K. Lange, *A tutorial on MM algorithms*, The American Statistician, 58 (2004), pp. 30–37.

[24] E. C. J. and T. Tao, *The Dantzig selector: Statistical estimation when p is much larger than n*, The Annals of Statistics, 35 (2007), pp. 2313–2351.

[25] D. T. Jamison, J. G. Breman, A. R. Measham, G. Alleyne, M. Claeson, D. B. Evans, P. Jha, A. Mills, and P. Musgrove, *Disease control priorities in developing countries*, The World Bank, 2006.

[26] T. Kronvall and A. Jakobsson, *Hyperparameter selection for group-sparse regression: A probabilistic approach*, Signal Processing, 151 (2018), pp. 107–118.

[27] K. Lange, *MM optimization algorithms*, vol. 147 of Other Titles in Applied Mathematics, SIAM, 2016.

[28] F. Lauer and H. Ohlsson, *Finding sparse solutions of systems of polynomial equations via group-sparsity optimization*, Journal of Global Optimization, 62 (2015), pp. 319–349.

[29] X. Liao, H. Li, and L. Carin, *Generalized alternating projection for weighted-$L_{2,1}$ minimization with applications to model-based compressive sensing*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 797–823.

[30] Y. Lou and M. Yan, *Fast l1-l2 minimization via a proximal operator*, Journal of Scientific Computing, 74 (2018), pp. 767–785.

[31] Y. Lou, P. Yin, Q. He, and J. Xin, *Computing sparse representation in a highly coherent dictionary based on difference of L_1 and L_2*, Journal of Scientific Computing, 64 (2015), pp. 178–196.

[32] L. Meier, S. Van De Geer, and P. Bühlmann, *The group lasso for logistic regression*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70 (2008), pp. 53–71.

[33] M. Nikolova, *Local strong homogeneity of a regularized estimator*, SIAM Journal on Applied Mathematics, 61 (2000), pp. 633–658.

[34] J. S. Pang, M. Razaviyayn, and A. Alvarado, *Computing b-stationary points of nonsmooth DC programs*, Mathematics of Operations Research, 42 (2017), pp. 95–118.

[35] M. Y. Park and T. Hastie, *L1-regularization path algorithm for generalized linear models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69 (2007), pp. 659–677.

[36] D. N. Phan and H. A. Le-Thi, *Group variable selection via $\ell_{p,0}$ regularization and application to optimal scoring*, Neural Networks, 118 (2019), pp. 220–234.

[37] Z. Qin and D. Goldfarb, *Structured sparsity via alternating direction methods*, Journal of Machine Learning Research, 13 (2012), pp. 1435–1468.

[38] Y. Rahimi, C. Wang, H. Dong, and Y. Lou, *A scale invariant approach for sparse signal recovery*, SIAM Journal on Scientific Computing, 41 (2019), pp. A3649–A3672.

[39] A. Rakotomamonjy, *Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms*, Signal Processing, 91 (2011), pp. 1505–1526.

[40] M. Razaviyayn, M. Hong, and Z.-Q. Luo, *A unified convergence analysis of block successive minimization methods for nonsmooth optimization*, SIAM Journal on Optimization, 23 (2013), pp. 1126–1153.

[41] I. Rudan, C. Boschi-Pinto, Z. Biloglav, K. Mulholland, and H. Campbell, *Epidemiology and etiology of childhood pneumonia*, Bulletin of the World Health Organization, 86 (2008), pp. 408–416.

[42] T. E. Scheetz, K.-Y. A. Kim, R. E. Swiderski, A. R. Philp, T. A. Braun, K. L. Knudtson, A. M. Dorrance, G. F. DiBona, J. Huang, T. L. Casavant, V. C. Sheffield, and E. M. Stone, *Regulation of gene expression in the mammalian eye and its relevance to eye disease*, Proceedings of the National Academy of Sciences, 103 (2006), pp. 14429–14434.

[43] S. Shin, J. Fine, and Y. Liu, *Adaptive estimation with partially overlapping models*, Statistica Sinica, 26 (2016), pp. 235–253.

[44] S. Shin, Y. Liu, S. R. Cole, and J. P. Fine, *Ensemble estimation and variable selection with semiparametric regression models*, Biometrika, 107 (2020), pp. 433–448.

[45] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58 (1996), pp. 267–288.

[46] C. Wang, M. Yan, Y. Rahimi, and Y. Lou, *Accelerated schemes for the L_1/L_2 minimization*, IEEE Transactions on Signal Processing, 68 (2020), pp. 2660–2669.

[47] L. Wang, G. Chen, and H. Li, *Group SCAD regression analysis for microarray time course gene expression data*, Bioinformatics, 23 (2007), pp. 1486–1494.

[48] Y. Wang, W. Yin, and J. Zeng, *Global convergence of ADMM in nonconvex nonsmooth optimization*, Journal of Scientific Computing, 78 (2019), pp. 29–63.
[49] D. Wipf and S. Nagarajan, *Iterative reweighted l1 and l2 methods for finding sparse solutions*, IEEE Journal of Selected Topics in Signal Processing, 4 (2010), pp. 317–329.
[50] Y. Wu and Y. Liu, *Variable selection in quantile regression*, Statistica Sinica, (2009), pp. 801–817.
[51] Y. Yang and H. Zou, *A fast unified algorithm for solving group-lasso penalize learning problems*, Statistics and Computing, 25 (2015), pp. 1129–1141.
[52] P. Yin, E. Esser, and J. Xin, *Ratio and difference of l_1 and l_2 norms and sparse representation with coherent dictionaries*, Communications in Information and Systems, 14 (2014), pp. 87–109.
[53] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68 (2007), pp. 49–67.
[54] C. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, The Annals of Statistics, (2010), pp. 894–942.
[55] S. Zhang and J. Xin, *Minimization of transformed L_1 penalty: theory, difference of convex function algorithm, and robust application in compressed sensing*, Mathematical Programming, 169 (2018), pp. 307–336.
[56] Y. Zhang, N. Zhang, D. Sun, and K.-C. Toh, *An efficient hessian based algorithm for solving large-scale sparse group lasso problems*, Mathematical Programming, 179 (2018), pp. 223–263.
[57] Z. Zhao, S. Wang, C. Sun, R. Yan, and X. Chen, *Sparse multiperiod group lasso for bearing multifault diagnosis*, IEEE Transactions on Instrumentation and Measurement, 69 (2019), pp. 419–431.
[58] Z. Zhao, S. Wu, B. Qiao, S. Wang, and X. Chen, *Enhanced sparse period-group lasso for bearing fault diagnosis*, IEEE Transactions on Industrial Electronics, 66 (2018), pp. 2143–2153.
[59] H. Zou, *The adaptive lasso and its oracle properties*, Journal of the American Statistical Association, 101 (2006), pp. 1418–1429.

Appendix A. Clinical signs of ARI. We list clustered clinical signs of ARI in Table 10.

Group name	Sign abbreviation	Clinical sign
bul.conv	abb	bulging fontanelle
	convul	hx convulsion
hydration	abk	sunken fontanelle
	hdi	hx diarrhoea
	deh	dehydrated
	stu	skin turgor
	dcp	digital capillary refill
drowsy	hcl	less activity
	qcr	quality of crying
	csd	drowsy state
	slpm	sleeping more
	wake	wakes less easy
	aro	arousal
	mvm	amount of movement
	att	attentive
agitated	hcm	crying more
	hcs	crying less
	slpl	sleeping less
	con	consolability
	csa	agitated state
reffort	nfl	nasal flaring
	lcw	lower chest in-drawing
	gru	grunting
	ccy	central cyanosis
breath	hap	hx stop breathing
	apn	apnoea
	hrat	heart rate
ausc	whz	wheezing
	coh	cough heard
	crs	crepitation
	str	stridor
hxprob	hfb	fast breathing
	hdb	diõculty breathing
	rr	adjusted respiratory rate
	inc	respiratory distress
	sr1	respiratory state 1
	sr2	respiratory state 2
feeding	hfa	hx abnormal feeding
	absu	sucking ability
	afe	drinking ability
	hvo	vomit more
labor	chi	previous child died
	fde	fever at delivery
	ldy	days in labour
	twb	water broke
abdominal	abd	abdominal distension
	jau	jaundice
	omph	omphalitis
fever.ill	temp	temperature (in Celsius)
	hfe	hx fever
pustular	conj	conjuctivitis
	oto	otoscopy impression
	puskin	pustular skin rash
birth	biwt	birth weight
	bat	birth preterm
growth	hcir	head circumference
	wght	weight (in grams)
	lgth	length (in centimeters)
age	age	age (in days)

TABLE 10

Clustering of clinical signs of ARI in young infants. The group regularization methods including our approach utilize the clinically-guided clustering for prediction of ARI.