Journal of Machine Learning Research 20 (2019) 1-42 Submitted 8/16; Published 8/19

Two-Layer Feature Reduction for Sparse-Group Lasso via Decomposition of Convex Sets

Jie Wang JIEWANGX@USTC.EDU.CN
Department of Electronic Engineering and Information Science
University of Science and Technology of China
Hefei, Anhui, China

Zhanqiu Zhang ZZQ96@MAIL.USTC.EDU.CN
Department of Electronic Engineering and Information Science
University of Science and Technology of China
Hefei, Anhui, China

Jieping Ye JPYE@UMICH.EDU
Department of Computational Medicine and Bioinformatics
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109-2218, USA

Editor: Zhihua Zhang

Abstract

Sparse-Group Lasso (SGL) has been shown to be a powerful regression technique for simultaneously discovering group and within-group sparse patterns by using a combination of the ℓ_1 and ℓ_2 norms. However, in large-scale applications, the complexity of the regularizers entails great computational challenges. In this paper, we propose a novel **t**wo-**l**ayer **f**eature **re**duction method (TLFre) for SGL via a decomposition of its dual feasible set. The two-layer reduction is able to quickly identify the inactive groups and the inactive features, respectively, which are guaranteed to be absent from the sparse representation and can be removed from the optimization. Existing feature reduction methods are only applicable to sparse models with one sparsity-inducing regularizer. To our best knowledge, TLFre is *the first one* that is capable of dealing with *multiple* sparsity-inducing regularizers. Moreover, TLFre has a very low computational cost and can be integrated with any existing solvers. We also develop a screening method—called DPC (**d**ecom**p**osition of **c**onvex set)—for nonnegative Lasso. Experiments on both synthetic and real data sets show that TLFre and DPC improve the efficiency of SGL and nonnegative Lasso by several orders of magnitude.

Keywords: Sparse, Sparse Group Lasso, Screening, Fenchel's Dual, Decomposition, Convex Sets, Composite Function Optimization

1. Introduction

Sparse-Group Lasso (SGL) (Friedman et al.; Simon et al., 2013) is a powerful regression technique in identifying important groups and features simultaneously. To yield sparsity at both group and individual feature levels, SGL combines the Lasso (Tibshirani, 1996) and group Lasso (Yuan and Lin, 2006) penalties. In recent years, SGL has found great success in a wide range of applications, including but not limited to machine learning

(Vidyasagar, 2014; Yogatama and Smith, 2014), signal processing (Sprechmann et al., 2011), bioinformatics (Peng et al., 2010) etc. Many research efforts have been devoted to developing efficient solvers for SGL (Friedman et al.; Simon et al., 2013; Liu and Ye, 2010; Vincent and Hansen, 2014). However, when the feature dimension is extremely high, the complexity of the SGL regularizers imposes great computational challenges. Therefore, there is an increasingly urgent need for nontraditional techniques to address the challenges posed by the massive volume of the data sources.

Recently, El Ghaoui et al. (2012) proposed a promising feature reduction method, called *SAFE screening*, to screen out the so-called *inactive* features, which have zero coefficients in the solution, from the optimization. Thus, the size of the data matrix needed for the training phase can be significantly reduced, which may lead to substantial improvement in the efficiency of solving sparse models. Inspired by SAFE, various exact and heuristic feature screening methods have been proposed for many sparse models such as Lasso (Wang et al., 2013; Liu et al., 2014; Tibshirani et al., 2012; Xiang and Ramadge, 2012), group Lasso (Wang et al., 2013; Wang et al.; Tibshirani et al., 2012), etc. It is worthwhile to mention that the discarded features by exact feature screening methods such as SAFE (El Ghaoui et al., 2012), DOME (Xiang and Ramadge, 2012) and EDPP (Wang et al., 2013) are guaranteed to have zero coefficients in the solution. However, heuristic feature screening methods like strong rule (Tibshirani et al., 2012) may mistakenly discard features that have nonzero coefficients in the solution. Thus, to compute the exact solutions, the authors propose to check the KKT conditions after the screening pass of strong rules. More recently, the idea of exact feature screening has been extended to exact sample screening, which screens out the nonsupport vectors in SVM (Ogawa et al., 2013; Wang et al., 2014) and LAD (Wang et al., 2014). As a promising data reduction tool, exact feature/sample screening would be of great practical importance because they can effectively reduce the data size without sacrificing the optimality (Ogawa et al., 2014).

However, all of the existing feature/sample screening methods are only applicable for the sparse models with one sparsity-inducing regularizer. In this paper, we propose an exact two-layer feature screening method, called TLFre, for the SGL problem. The first and second layer of TLFre aim to quickly identify the inactive groups and the inactive features, respectively, which are guaranteed to have zero coefficients in the solution. To the best of our knowledge, TLFre is the first screening method which is capable of dealing with multiple sparsity-inducing regularizers.

We note that most of the existing exact feature screening methods involve an estimation of the dual optimal solution. The difficulty in developing screening methods for sparse models with multiple sparsity-inducing regularizers like SGL is that the dual feasible set is the sum of simple convex sets. Thus, to determine the feasibility of a given point, we need to know if it is decomposable with respect to the summands, which is itself a nontrivial problem (see Section 2). One of our major contributions is that we derive an elegant decomposition method of any dual feasible solutions of SGL via the framework of Fenchel's duality (see Section 3). Based on the Fenchel's dual problem of SGL, we motivate TLFre by an in-depth exploration of its geometric properties and the optimality conditions in Section 4. We derive the set of the regularization parameter values corresponding to zero solutions. To develop TLFre, we need to estimate the upper bounds involving the dual optimal solution. To this end, we first give an accurate estimation of the dual optimal solution via the normal cones. Then, we formulate the estimation of the upper bounds via nonconvex optimization problems. We show that these nonconvex problems admit closed form solutions.

The rest of this paper is organized as follows. In Section 2, we briefly review some basics of the SGL problem. We then derive the Fenchel's dual of SGL with nice geometric properties under the elegant framework of Fenchel's Duality in Section 3. In Section 4, we develop the TLFre screening rule for SGL. To demonstrate the flexibility of the proposed framework, we extend TLFre to the nonnegative Lasso problem in Section 5. Experiments in Section 6 on both synthetic and real data demonstrate that the speedup gained by the proposed screening rules in solving SGL and nonnegative Lasso can be orders of magnitude. Please see the appendix for detailed proofs that are not presented in the main text.

Notation: Let $\|\cdot\|_1$, $\|\cdot\|$ and $\|\cdot\|_\infty$ be the ℓ_1, ℓ_2 and ℓ_∞ norms, respectively. Denote by $\mathcal{B}_1^n$, $\mathcal{B}^n$, and $\mathcal{B}_\infty^n$ the unit ℓ_1, ℓ_2, and ℓ_∞ norm balls in $\mathbb{R}^n$ (we omit the superscript if it is clear from the context). For a set $\mathcal{C}$, let $\operatorname{int}\mathcal{C}$ be its interior. If $\mathcal{C}$ is closed and convex, we define the projection operator as

$$\mathbf{P}_{\mathcal{C}}(\mathbf{w}) := \operatorname{argmin}_{\mathbf{u}\in\mathcal{C}}\|\mathbf{w}-\mathbf{u}\|.$$

We denote the indicator function of $\mathcal{C}$ by

$$\mathbf{I}_{\mathcal{C}}(\mathbf{w}) = \begin{cases} 0, & \text{if } \mathbf{w}\in\mathcal{C},\\ \infty, & \text{otherwise.}\end{cases}$$

Let $\Gamma_0(\mathbb{R}^n)$ be the class of proper closed convex functions on $\mathbb{R}^n$. For $f\in\Gamma_0(\mathbb{R}^n)$, let ∂f be its subdifferential. The domain of f is the set $\operatorname{dom} f := \{\mathbf{w} : f(\mathbf{w}) < \infty\}$.

For $\mathbf{w}\in\mathbb{R}^n$, let $[\mathbf{w}]_i$ be its i^{th} component. More generally, if $\mathcal{G}\subset\{1,2,\ldots,n\}$ is an index set, we denote the corresponding subvector of $\mathbf{w}$ by $[\mathbf{w}]_{\mathcal{G}}\in\mathbb{R}^{|\mathcal{G}|}$, where $|\mathcal{G}|$ denotes the number of elements in $\mathcal{G}$. For $\gamma\in\mathbb{R}$, let

$$\operatorname{sgn}(\gamma) = \begin{cases}\operatorname{sign}(\gamma), & \text{if } \gamma\neq 0,\\ 0, & \text{otherwise.}\end{cases}$$

We define

$$\operatorname{SGN}(\mathbf{w}) = \left\{\mathbf{s}\in\mathbb{R}^n : [\mathbf{s}]_i \in \begin{cases}\operatorname{sign}([\mathbf{w}]_i), & \text{if } [\mathbf{w}]_i\neq 0;\\ [-1,1], & \text{if } [\mathbf{w}]_i = 0.\end{cases}\right\}$$

We denote by $\gamma_+ = \max(\gamma, 0)$. Then, for $\gamma\geq 0$, the shrinkage operator $\mathcal{S}_\gamma(\mathbf{w}) : \mathbb{R}^n\to\mathbb{R}^n$ can be written as

$$[\mathcal{S}_\gamma(\mathbf{w})]_i = (|[\mathbf{w}]_i| - \gamma)_+ \operatorname{sgn}([\mathbf{w}]_i),\ i = 1,\ldots,n. \tag{1}$$

2. Basics and Motivation

In this section, we briefly review some basics of SGL. Let $\mathbf{y}\in\mathbb{R}^N$ be the response vector and $\mathbf{X}\in\mathbb{R}^{N\times p}$ be the matrix of features. With the group information available, the SGL problem (Friedman et al.) is

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\left\|\mathbf{y} - \sum\nolimits_{g=1}^G \mathbf{X}_g\beta_g\right\|^2 + \lambda_1\sum\nolimits_{g=1}^G \sqrt{n_g}\|\beta_g\| + \lambda_2\|\beta\|_1, \tag{2}$$

where n_g is the number of features in the g^{th} group, $\mathbf{X}_g \in \mathbb{R}^{N \times n_g}$ denotes the predictors in that group with the corresponding coefficient vector β_g, and λ_1, λ_2 are positive regularization parameters. Without loss of generality, let $\lambda_1 = \alpha\lambda$ and $\lambda_2 = \lambda$ with $\alpha > 0$. Then, problem (2) becomes:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \left\| \mathbf{y} - \sum\nolimits_{g=1}^{G} \mathbf{X}_g \beta_g \right\|^2 + \lambda \left(\alpha \sum\nolimits_{g=1}^{G} \sqrt{n_g} \|\beta_g\| + \|\beta\|_1 \right). \tag{3}$$

By the Lagrangian multipliers method (Boyd and Vandenberghe, 2004) (see the appendix), we can derive the dual problem of SGL as follows.

$$\begin{aligned} &\sup_{\theta} \ \frac{1}{2}\|\mathbf{y}\|^2 - \frac{1}{2}\left\| \frac{\mathbf{y}}{\lambda} - \theta \right\|^2 \\ &\text{s.t. } \ \mathbf{X}_g^T \theta \in \mathcal{D}_g^\alpha := \alpha\sqrt{n_g}\mathcal{B} + \mathcal{B}_\infty,\ g = 1, \ldots, G. \end{aligned} \tag{4}$$

It is well-known that the dual feasible set of Lasso is the intersection of closed half spaces (thus a polytope); for group Lasso, the dual feasible set is the intersection of ellipsoids. Surprisingly, the geometric properties of these dual feasible sets play fundamentally important roles in most of the existing screening methods for sparse models with one sparsity-inducing regularizer (Wang et al., 2014; Liu et al., 2014; Wang et al., 2013; El Ghaoui et al., 2012).

When we incorporate multiple sparse-inducing regularizers to the sparse models, problem (4) indicates that the dual feasible set can be much more complicated. Although (4) provides a geometric description of the dual feasible set of SGL, it is not suitable for further analysis. Notice that, *even the feasibility of a given point θ is not easy to determine*, since it is nontrivial to tell if $\mathbf{X}_g^T\theta$ can be decomposed into $\mathbf{b}_1 + \mathbf{b}_2$ with $\mathbf{b}_1 \in \alpha\sqrt{n_g}\mathcal{B}$ and $\mathbf{b}_2 \in \mathcal{B}_\infty$. Therefore, to develop screening methods for SGL, it is desirable to gain deeper understanding of the sum of simple convex sets.

In the next section, we analyze the dual feasible set of SGL in depth via the Fenchel's Duality Theorem. We show that for each $\mathbf{X}_g^T\theta \in \mathcal{D}_g^\alpha$, Fenchel's duality naturally leads to an explicit decomposition $\mathbf{X}_g^T\theta = \mathbf{b}_1 + \mathbf{b}_2$, with one belonging to $\alpha\sqrt{n_g}\mathcal{B}$ and the other one belonging to $\mathcal{B}_\infty$. This lays the foundation of the proposed screening method for SGL.

3. The Fenchel's Dual Problem of SGL

In Section 3.1, we derive the Fenchel's dual of SGL via Fenchel's Duality Theorem. We then motivate TLFre in Section 3.2 and sketch our approach by Algorithm 1. In Section 3.3, we explore the geometric properties of the Fenchel's dual of SGL and derive the effective interval of the parameter λ with a fixed value of α—that is the set of λ given α corresponding to nonzero solutions of SGL.

Algorithm 1 Guidelines for developing TLFre.

1: Given a pair of parameter values (λ, α), we estimate a region Θ that contains the dual optimum $\theta^*(\lambda, \alpha)$ of (4).

2: We solve the following two optimization problems:

$$s_g^* = \sup_{\xi_g} \{\|\mathcal{S}_1(\xi_g)\| : \xi_g \in \Xi_g \supseteq \mathbf{X}_g^T \Theta\}, \text{ where } \mathbf{X}_g^T\Theta = \{\mathbf{X}_g^T \theta : \theta \in \Theta\}, \tag{5}$$

$$t_{g_k}^* = \sup_{\theta} \{|\mathbf{x}_{g_k}^T \theta| : \theta \in \Theta\}, \text{ where } \mathbf{x}_{g_k} \text{ is the } k^{th} \text{ column of } \mathbf{X}_g. \tag{6}$$

3: The TLFre screening rules take the form of

$$s_g^* < \alpha\sqrt{n_g} \Rightarrow \beta_g^*(\lambda, \alpha) = 0, \tag{7}$$

$$t_{g_k}^* \leq 1 \Rightarrow [\beta_g^*(\lambda, \alpha)]_k = 0, \tag{8}$$

where $\beta^*(\lambda, \alpha)$ is the optimal solution of SGL in (3).

3.1. The Fenchel's Dual of SGL via Fenchel's Duality Theorem

To derive the Fenchel's dual problem of SGL, we need the Fenchel's Duality Theorem as stated in Theorem 1. We denote the conjugate of $f \in \Gamma_0(\mathbb{R}^n)$ by $f^* \in \Gamma_0(\mathbb{R}^n)$:

$$f^*(\mathbf{z}) = \sup_{\mathbf{w}} \langle \mathbf{w}, \mathbf{z} \rangle - f(\mathbf{w}). \tag{9}$$

Theorem 1 [Fenchel's Duality Theorem] *Let $f \in \Gamma_0(\mathbb{R}^N)$, $\Omega \in \Gamma_0(\mathbb{R}^p)$, and $\mathcal{T}(\beta) = \mathbf{y} - \mathbf{X}\beta$ be an affine mapping from $\mathbb{R}^p$ to $\mathbb{R}^N$. Let $p^*, d^* \in [-\infty, \infty]$ be primal and dual values defined, respectively, by the Fenchel problems:*

$$p^* = \inf_{\beta \in \mathbb{R}^p} f(\mathbf{y} - \mathbf{X}\beta) + \lambda\Omega(\beta); \;\; d^* = \sup_{\theta \in \mathbb{R}^N} -f^*(\lambda\theta) - \lambda\Omega^*(\mathbf{X}^T\theta) + \lambda\langle \mathbf{y}, \theta \rangle.$$

One has $p^* \geq d^*$. If, furthermore, f and Ω satisfy the condition

$$0 \in \text{int} \left(\text{dom}\, f - \mathbf{y} + \mathbf{X} \text{dom}\, \Omega\right),$$

then $p^* = d^*$, and the supreme is attained in the dual problem if finite.

We omit the proof of Theorem 1 as it is similar to that of Theorem 3.3.5 in (Borwein and Lewis, 2006).

Let $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$ and $\lambda\Omega(\beta)$ be the second term in (3). We can write SGL as

$$\min_\beta f(\mathbf{y} - \mathbf{X}\beta) + \lambda\Omega(\beta). \tag{10}$$

To derive the Fenchel's dual problem of SGL, Theorem 1 implies that we need to find f^* and Ω^*. It is well-known that $f^*(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|^2$. Therefore, we only need to find Ω^*, where the concept *infimal convolution* is needed:

Definition 2 (Bauschke and Combettes, 2011) *Let $h, g \in \Gamma_0(\mathbb{R}^n)$. The infimal convolution of h and g is*

$$(h \Box g)(\xi) = \inf_\eta \, h(\eta) + g(\xi - \eta), \tag{11}$$

and it is exact at a point ξ if there exists a $\eta^*(\xi)$ such that

$$(h \Box g)(\xi) = h(\eta^*(\xi)) + g(\xi - \eta^*(\xi)). \tag{12}$$

$h \Box g$ is exact if it is exact at each point in its domain, and we denote it by $h \boxdot g$.

With the infimal convolution, we derive Ω^* in the following Lemma.

Lemma 3 *Let $\Omega_1^\alpha(\beta) = \alpha \sum_{g=1}^G \sqrt{n_g} \|\beta_g\|$, $\Omega_2(\beta) = \|\beta\|_1$ and $\Omega(\beta) = \Omega_1^\alpha(\beta) + \Omega_2(\beta)$. Moreover, let $\mathcal{C}_g^\alpha = \alpha\sqrt{n_g}\mathcal{B} \subset \mathbb{R}^{n_g}$, $g = 1, \ldots, G$. Then, the following hold:*

(i) $(\Omega_1^\alpha)^*(\xi) = \sum_{g=1}^G \mathbf{I}_{\mathcal{C}_g^\alpha}(\xi_g)$, $\quad (\Omega_2)^*(\xi) = \sum_{g=1}^G \mathbf{I}_{\mathcal{B}_\infty}(\xi_g)$,

(ii) $\Omega^*(\xi) = \left((\Omega_1^\alpha)^* \boxdot (\Omega_2)^*\right)(\xi) = \sum_{g=1}^G \mathbf{I}_{\mathcal{B}}\left(\frac{\xi_g - \mathbf{P}_{\mathcal{B}_\infty}(\xi_g)}{\alpha\sqrt{n_g}}\right)$,

where $\xi_g \in \mathbb{R}^{n_g}$ is the sub-vector of ξ corresponding to the g^{th} group.

To prove Lemma 3, we first cite the following technical result.

Theorem 4 (Hiriart-Urruty, 2006) *Let $f_1, \cdots, f_k \in \Gamma_0(\mathbb{R}^n)$. Suppose there is a point in $\cap_{i=1}^k \mathrm{dom}\, f_i$ at which $f_1, \cdots, f_{k-1}$ is continuous. Then, for all $p \in \mathbb{R}^n$:*

$$(f_1 + \cdots + f_k)^*(p) = \min_{p_1 + \cdots + p_k = p} [f_1^*(p_1) + \cdots + f_k^*(p_k)].$$

We now give the proof of Lemma 3.

Proof The first part can be derived directly by the definition as follows:

$$\begin{aligned}(\Omega_1^\alpha)^*(\xi) = & \sup_\beta \langle \beta, \xi \rangle - \Omega_1^\alpha(\beta) = \sum_{g=1}^G \alpha\sqrt{n_g} \left(\sup_{\beta_g} \left\langle \beta_g, \frac{\xi_g}{\alpha\sqrt{n_g}} \right\rangle - \|\beta_g\| \right) \\ = & \sum_{g=1}^G \alpha\sqrt{n_g} \mathbf{I}_{\mathcal{B}} \left(\frac{\xi_g}{\alpha\sqrt{n_g}} \right) = \sum_{g=1}^G \mathbf{I}_{\mathcal{B}} \left(\frac{\xi_g}{\alpha\sqrt{n_g}} \right) = \sum_{g=1}^G \mathbf{I}_{\mathcal{C}_g^\alpha}(\xi_g).\end{aligned}$$

$$(\Omega_2)^*(\xi) = \sup_\beta \langle \beta, \xi \rangle - \Omega_2(\beta) = \mathbf{I}_{\mathcal{B}_\infty}(\xi) = \sum_{g=1}^G \mathbf{I}_{\mathcal{B}_\infty}(\xi_g).$$

To show the second part, Theorem 4 indicates that we only need to show $(\Omega_1^\alpha)^* \Box (\Omega_2)^*(\xi)$ is exact (note that Ω_1^α and Ω_2 are continuous everywhere). Let us now compute $(\Omega_1^\alpha)^* \Box (\Omega_2)^*$.

$$\begin{aligned}\left((\Omega_1^\alpha)^* \Box (\Omega_2)^*\right)(\xi) = & \inf_\eta \, (\Omega_1^\alpha)^*(\xi - \eta) + (\Omega_2)^*(\eta) \qquad (13) \\ = & \sum_{g=1}^G \inf_{\eta_g} \mathbf{I}_{\mathcal{B}} \left(\frac{\xi_g - \eta_g}{\alpha\sqrt{n_g}} \right) + \mathbf{I}_{\mathcal{B}_\infty}(\eta_g) \\ = & \sum_{g=1}^G \inf_{\|\eta_g\|_\infty \le 1} \mathbf{I}_{\mathcal{B}} \left(\frac{\xi_g - \eta_g}{\alpha\sqrt{n_g}} \right)\end{aligned}$$

To solve the optimization problem in (13), i.e.,

$$\mu_g^* = \inf_{\eta_g} \left\{ \mathbf{I}_{\mathcal{B}} \left(\frac{\xi_g - \eta_g}{\alpha\sqrt{n_g}} \right) : \|\eta_g\|_\infty \leq 1 \right\}, \tag{14}$$

we can consider the following problem

$$\nu_g^* = \inf_{\eta_g} \left\{ \frac{1}{\alpha\sqrt{n_g}} \|\xi_g - \eta_g\| : \|\eta_g\|_\infty \leq 1 \right\}. \tag{15}$$

We can see that the optimal solution of problem (15) must also be an optimal solution of problem (14). Let $\eta_g^*(\xi_g)$ be the optimal solution of (15). We can see that $\eta_g^*(\xi_g)$ is indeed the projection of ξ_g on $\mathcal{B}_\infty$, which admits a closed form solution:

$$[\eta_g^*(\xi_g)]_i = [\mathbf{P}_{\mathcal{B}_\infty}(\xi_g)]_i = \begin{cases} 1, & \text{if } [\xi_g]_i > 1, \\ [\xi_g]_i, & \text{if } |[\xi_g]_i| \leq 1, \\ -1, & \text{if } [\xi_g]_i < -1. \end{cases}$$

Thus, problem (14) can be solved as

$$\mu_g^* = \mathbf{I}_{\mathcal{B}} \left(\frac{\xi_g - \mathbf{P}_{\mathcal{B}_\infty}(\xi_g)}{\alpha\sqrt{n_g}} \right).$$

Hence, the infimal convolution in Eq. (13) is exact and Theorem 4 leads to

$$\Omega^*(\xi) = \left((\Omega_1^\alpha)^* \boxdot (\Omega_2)^*\right)(\xi) = \sum_{g=1}^{G} \mathbf{I}_{\mathcal{B}} \left(\frac{\xi_g - \mathbf{P}_{\mathcal{B}_\infty}(\xi_g)}{\alpha\sqrt{n_g}} \right), \tag{16}$$

which completes the proof. ■

Note that $\mathbf{P}_{\mathcal{B}_\infty}(\xi_g)$ admits a closed form solution:

$$[\mathbf{P}_{\mathcal{B}_\infty}(\xi_g)]_i = \operatorname{sgn}([\xi_g]_i) \min(|[\xi_g]_i|, 1).$$

By Theorem 1 and Lemma 3, we derive the Fenchel's dual of SGL in Theorem 5 (see Section B for the proof).

Theorem 5 *For the SGL problem in (3), the following hold:*

(i) *The Fenchel's dual of SGL is given by:*

$$\begin{aligned} &\inf_{\theta} \ \frac{1}{2}\|\frac{\mathbf{y}}{\lambda} - \theta\|^2 - \frac{1}{2}\|\mathbf{y}\|^2, \\ &\text{s.t.} \ \left\|\mathbf{X}_g^T\theta - \mathbf{P}_{\mathcal{B}_\infty}(\mathbf{X}_g^T\theta)\right\| \leq \alpha\sqrt{n_g}, \ g = 1, \ldots, G. \end{aligned} \tag{17}$$

(ii) *Let $\beta^*(\lambda, \alpha)$ and $\theta^*(\lambda, \alpha)$ be the optimal solutions of problems (3) and (17), respectively. Then,*

$$\lambda\theta^*(\lambda, \alpha) = \mathbf{y} - \mathbf{X}\beta^*(\lambda, \alpha), \tag{18}$$

$$\mathbf{X}_g^T\theta^*(\lambda, \alpha) \in \alpha\sqrt{n_g}\partial\|\beta_g^*(\lambda, \alpha)\| + \partial\|\beta_g^*(\lambda, \alpha)\|_1, \ g = 1, \ldots, G. \tag{19}$$

Eq. (18) and Eq. (19) are the so-called KKT conditions (Boyd and Vandenberghe, 2004) and can also be obtained by the Lagrangian multiplier method (see Section A in the appendix).

Remark 6 *We note that the shrinkage operator can also be expressed by*

$$\mathcal{S}_\gamma(\mathbf{w}) = \mathbf{w} - \mathbf{P}_{\gamma\mathcal{B}_\infty}(\mathbf{w}),\ \ \gamma \geq 0. \tag{20}$$

Therefore, problem (17) can be written more compactly as

$$\begin{aligned} &\inf_\theta\ \frac{1}{2}\|\frac{\mathbf{y}}{\lambda} - \theta\|^2 - \frac{1}{2}\|\mathbf{y}\|^2, \\ &\text{s.t.}\ \ \left\|\mathcal{S}_1(\mathbf{X}_g^T\theta)\right\| \leq \alpha\sqrt{n_g},\ g = 1, \ldots, G. \end{aligned} \tag{21}$$

The equivalence between the dual formulations For the SGL problem, its Lagrangian dual in (4) and Fenchel's dual in (17) are indeed equivalent to each other. We bridge them together by the following lemma.

Lemma 7 (Bauschke and Combettes, 2011) *Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be nonempty subsets of $\mathbb{R}^n$. Then $\mathbf{I}_{\mathcal{C}_1} \boxdot \mathbf{I}_{\mathcal{C}_2} = \mathbf{I}_{\mathcal{C}_1+\mathcal{C}_2}$.*

In view of Lemmas 3 and 7, and recall that $\mathcal{D}_g^\alpha = \mathcal{C}_g^\alpha + \mathcal{B}_\infty$, we have

$$\Omega^*(\xi) = \left((\Omega_1^\alpha)^* \boxdot (\Omega_2)^*\right)(\xi) = \sum\nolimits_{g=1}^G \left(\mathbf{I}_{\mathcal{C}_g^\alpha} \boxdot \mathbf{I}_{\mathcal{B}_\infty}\right)(\xi_g) = \sum\nolimits_{g=1}^G \mathbf{I}_{\mathcal{D}_g^\alpha}(\xi_g). \tag{22}$$

Combining Eq. (22) and Theorem 1, we obtain the dual formulation of SGL in (4). Therefore, the dual formulations of SGL in (4) and (17) are the same.

Remark 8 *An appealing advantage of the Fenchel's dual in (17) is that we have a natural decomposition of all points $\xi_g \in \mathcal{D}_g^\alpha$: $\xi_g = \mathbf{P}_{\mathcal{B}_\infty}(\xi_g) + \mathcal{S}_1(\xi_g))$ with $\mathbf{P}_{\mathcal{B}_\infty}(\xi_g) \in \mathcal{B}_\infty$ and $\mathcal{S}_1(\xi_g) \in \mathcal{C}_g^\alpha$. As a result, this leads to a convenient way to determine the feasibility of any dual variable θ by checking if $\mathcal{S}_1(\mathbf{X}_g^T\theta) \in \mathcal{C}_g^\alpha$, $g = 1, \ldots, G$.*

3.2. Motivation of the Two-Layer Screening Rules

We motive the two-layer screening rules via the KKT condition in Eq. (19). As implied by the name, there are two layers in our method. The first layer aims to identify the inactive groups, and the second layer detects the inactive features for the remaining groups.

by Eq. (19), we have the following cases by noting $\partial\|\mathbf{w}\|_1 = \text{SGN}(\mathbf{w})$ and

$$\partial\|\mathbf{w}\| = \begin{cases} \left\{\frac{\mathbf{w}}{\|\mathbf{w}\|}\right\}, & \text{if } \mathbf{w} \neq 0, \\ \{\mathbf{u} : \|\mathbf{u}\| \leq 1\}, & \text{if } \mathbf{w} = 0. \end{cases}$$

Case 1. If $\beta_g^*(\lambda, \alpha) \neq 0$, we have

$$[\mathbf{X}_g^T\theta^*(\lambda, \alpha)]_k \in \begin{cases} \alpha\sqrt{n_g}\frac{[\beta_g^*(\lambda,\alpha)]_k}{\|\beta_g^*(\lambda,\alpha)\|} + \text{sign}([\beta_g^*(\lambda, \alpha)]_k), & \text{if } [\beta_g^*(\lambda, \alpha)]_k \neq 0, \\ [-1, 1], & \text{if } [\beta_g^*(\lambda, \alpha)]_k = 0. \end{cases} \tag{23}$$

In view of Eq. (23), we can see that

$$\text{(a): } \mathcal{S}_1(\mathbf{X}_g^T\theta^*(\lambda,\alpha)) = \alpha\sqrt{n_g}\frac{\beta_g^*(\lambda_1,\lambda_2)}{\|\beta_g^*(\lambda_1,\lambda_2)\|} \text{ and } \|\mathcal{S}_1(\mathbf{X}_g^T\theta^*(\lambda,\alpha))\| = \alpha\sqrt{n_g}, \tag{24}$$

$$\text{(b): If } \left|[\mathbf{X}_g^T\theta^*(\lambda,\alpha]_k\right| \leq 1 \text{ then } [\beta_g^*(\lambda,\alpha)]_k = 0. \tag{25}$$

Case 2. If $\beta_g^*(\lambda,\alpha)=0$, we have

$$[\mathbf{X}_g^T\theta^*(\lambda,\alpha)]_k \in \alpha\sqrt{n_g}[\mathbf{u}_g]_k + [-1,1],\ \|\mathbf{u}_g\| \leq 1. \tag{26}$$

The first layer (group-level) of TLFre From (24) in **Case 1**, we have

$$\left\|\mathcal{S}_1(\mathbf{X}_g^T\theta^*(\lambda,\alpha))\right\| < \alpha\sqrt{n_g} \Rightarrow \beta_g^*(\lambda,\alpha) = 0. \tag{R1}$$

We can see that we can utilize (R1) to identify the inactive groups, and thus it is a group-level screening rule.

The second layer (feature-level) of TLFre Let $\mathbf{x}_{g_k}$ be the k^{th} column of $\mathbf{X}_g$. We have $[\mathbf{X}_g^T\theta^*(\lambda,\alpha)]_k = \mathbf{x}_{g_k}^T\theta^*(\lambda,\alpha)$. In view of (25) and (26), we can see that

$$\left|\mathbf{x}_{g_k}^T\theta^*(\lambda,\alpha)\right| \leq 1 \Rightarrow [\beta_g^*(\lambda,\alpha)]_k = 0. \tag{R2}$$

Different from (R1), (R2) detects the inactive features, and thus it is a feature-level screening rule.

However, we cannot directly apply (R1) and (R2) to identify the inactive groups/features because both need to know $\theta^*(\lambda,\alpha)$. Inspired by the SAFE rules (El Ghaoui et al., 2012), we can first estimate a region Θ containing $\theta^*(\lambda,\alpha)$. Let $\mathbf{X}_g^T\Theta = \{\mathbf{X}_g^T\theta : \theta \in \Theta\}$. Then, (R1) and (R2) can be relaxed as follows:

$$\sup_{\xi_g}\left\{\|\mathcal{S}_1(\xi_g)\| : \xi_g \in \Xi_g \supseteq \mathbf{X}_g^T\Theta\right\} < \alpha\sqrt{n_g} \Rightarrow \beta_g^*(\lambda,\alpha) = 0, \tag{R1*}$$

$$\sup_{\theta}\left\{\left|\mathbf{x}_{g_k}^T\theta\right| : \theta \in \Theta\right\} \leq 1 \Rightarrow [\beta_g^*(\lambda,\alpha)]_k = 0. \tag{R2*}$$

We note that the two optimization problems in (R1*) and (R2*) are the same with (5) and (6), respectively. Therefore, inspired by (R1*) and (R2*), we can develop TLFre via the guidelines as shown in Algorithm 1.

3.3. The Effective Interval of Parameter Values

In this section, we explore the geometric properties of the Fenchel's dual of SGL in depth—based on which we can derive the set of parameter values such that the primal optimum is zero/nonzero. We note that, Simon et al. (Simon et al., 2013) derived similar results for SGL with a different parameterization of the parameter values. However, their approach is based on the primal problem of SGL and the KKT conditions. Our new approach—that is based on the dual perspective—sheds new insights on the geometric properties of SGL. We consider the SGL problem in (3) and (2) in Section 3.3.1 and 3.3.2, respectively.

3.3.1. The Effective Interval of Parameter Values for Problem (3)

Consider the SGL problem in (3). For notational convenience, let

$$\mathcal{F}_g^\alpha = \{\theta : \|\mathcal{S}_1(\mathbf{X}_g^T\theta)\| \leq \alpha\sqrt{n_g}\},\ g = 1, \ldots, G.$$

We denote the feasible set of the Fenchel's dual of SGL by

$$\mathcal{F}^\alpha = \cap_{g=1,\ldots,G}\,\mathcal{F}_g^\alpha.$$

Problem (17) [or (21)] implies that $\theta^*(\lambda, \alpha)$ is the projection of $\mathbf{y}/\lambda$ on $\mathcal{F}^\alpha$, i.e.,

$$\theta^*(\lambda, \alpha) = \mathbf{P}_{\mathcal{F}^\alpha}(\mathbf{y}/\lambda). \tag{27}$$

Thus, if $\mathbf{y}/\lambda \in \mathcal{F}^\alpha$, we have $\theta^*(\lambda, \alpha) = \mathbf{y}/\lambda$. Moreover, (R1) implies that $\beta^*(\lambda, \alpha) = 0$ if $\mathbf{y}/\lambda$ is an *interior* point of $\mathcal{F}^\alpha$. Indeed, we have the following stronger result.

Theorem 9 *For the SGL problem in (3), let*

$$\lambda_{\max}^\alpha = \max_g \{\rho_g : \left\|\mathcal{S}_1(\mathbf{X}_g^T\mathbf{y}/\rho_g)\right\| = \alpha\sqrt{n_g}\}. \tag{28}$$

Then, the following statements are equivalent:

(i) $\dfrac{\mathbf{y}}{\lambda} \in \mathcal{F}^\alpha$, (ii) $\theta^*(\lambda, \alpha) = \dfrac{\mathbf{y}}{\lambda}$, (iii) $\beta^*(\lambda, \alpha) = 0$, (iv) $\lambda \geq \lambda_{\max}^\alpha$.

Remark 10 *Theorem 9 implies that the primal optimum $\beta^*(\lambda, \alpha) \neq 0$ if and only if $\lambda \in (0, \lambda_{\max}^\alpha)$, namely, the effective interval of the parameter λ with a fixed value of α is $(0, \lambda_{\max}^\alpha)$.*

We note that ρ_g in the definition of $\lambda_{\max}^\alpha$ admits a closed form solution. For notational convenience, let $|\mathbf{w}|$ be the vector by taking absolute value of $\mathbf{w}$ component-wisely and $[\mathbf{w}]^{(k)}$ be the vector consisting of the first k components of $\mathbf{w}$.

Lemma 11 *We sort $0 \neq |\mathbf{X}_g^T\mathbf{y}| \in \mathbb{R}^{n_g}$ in descending order and denote it by $\mathbf{z}$.*

(i) *If there exists $[\mathbf{z}]_k$ such that $\|\mathcal{S}_1(\mathbf{X}_g^T\mathbf{y}/[\mathbf{z}]_k)\| = \alpha\sqrt{n_g}$, then $\rho_g = [\mathbf{z}]_k$.*

(ii) *Otherwise, let $\tau_i = \|\mathcal{S}_1(\mathbf{X}_g^T\mathbf{y}/[\mathbf{z}]_i)\|$, $i = 1, \ldots, n_g$, and $\tau_{n_g+1} = \infty$. There exists a k such that $\alpha\sqrt{n_g} \in (\tau_k, \tau_{k+1})$, and $\rho_g \in ([\mathbf{z}]_{k+1}, [\mathbf{z}]_k)$ is the root of*

$$(k - \alpha^2 n_g)\rho^2 - 2\rho\|[\mathbf{z}]^{(k)}\|_1 + \|[\mathbf{z}]^{(k)}\|^2 = 0.$$

We omit the proof of Lemma 11 because it is a direct consequence by noting that

$$\|\mathcal{S}_1(\mathbf{X}_g^T\mathbf{y}/\lambda)\|^2 = \alpha^2 n_g$$

is piecewise quadratic.

3.3.2. The Effective Interval of Parameter Values for Problem (2)

Theorem 9 implies that the optimal solution $\beta^*(\lambda, \alpha)$ is 0 as long as $\mathbf{y}/\lambda \in \mathcal{F}^\alpha$. This geometric property also leads to an explicit characterization of the set of (λ_1, λ_2) such that the corresponding solution of problem (2) is 0. We denote by $\bar{\beta}^*(\lambda_1, \lambda_2)$ the optimal solution of problem (2).

Corollary 12 *For the SGL problem in (2), let*

$$\lambda_1^{\max}(\lambda_2) = \max_g \frac{1}{\sqrt{n_g}} \|\mathcal{S}_{\lambda_2}(\mathbf{X}_g^T \mathbf{y})\|.$$

Then, the following hold.

(i) $\bar{\beta}^*(\lambda_1, \lambda_2) = 0 \Leftrightarrow \lambda_1 \geq \lambda_1^{\max}(\lambda_2)$.

(ii) $\bar{\beta}^*(\lambda_1, \lambda_2) = 0$ *if*

$$\lambda_1 \geq \lambda_1^{\max} := \max_g \frac{1}{\sqrt{n_g}} \|\mathbf{X}_g^T \mathbf{y}\| \text{ or } \lambda_2 \geq \lambda_2^{\max} := \|\mathbf{X}^T \mathbf{y}\|_\infty.$$

By Corollary 12, we can see that the primal optimum $\bar{\beta}^*(\lambda_1, \lambda_2) \neq 0$ if and only if $\lambda_1 \in (0, \lambda_1^{\max}(\lambda_2))$. In other words, the effective interval of the parameter λ_1 with a fixed value of λ_2 is $(0, \lambda_1^{\max}(\lambda_2))$.

4. The Two-Layer Screening Rules for SGL

We follow the guidelines in Algorithm 1 to develop TLFre. In Section 4.1, we give an accurate estimation of $\theta^*(\lambda, \alpha)$ via normal cones (Ruszczyński, 2006). Then, we compute the supreme values in (R1*) and (R2*) by solving nonconvex problems in Section 4.2.

We note that, in many applications, the parameter values that perform the best are usually unknown. To determine appropriate parameter values, commonly used approaches such as cross validation and stability selection involve solving SGL many times over a grip of parameter values. Thus, given $\{\alpha_i\}_{i=1}^{\mathcal{I}}$ and $\lambda_{i,1} > \cdots > \lambda_{i,\mathcal{J}_i}$, we can fix the value of α each time and solve SGL by varying the value of λ. We repeat the process until we solve SGL for all of the parameter values.

We present the TLFre screening rule combined with any solver for solving the SGL problems at a grid of parameters in Algorithm 2 (see Section 4.3 for a detailed explanation). Moreover, Theorem 9 gives the closed form solution of $\beta^*(\lambda, \alpha)$ for any $\lambda \geq \lambda_{\max}^\alpha$. Thus, we assume that the input parameter values in Algorithm 2 satisfy $\lambda_{i,j} < \lambda_{\max}^{\alpha_i}$ for all $i = 1, \ldots, \mathcal{I}$ and $j = 1, \ldots, \mathcal{J}_i$.

Algorithm 2 The TLFre screening rule combined with any solver of SGL.

```
Input: $\{(\lambda_{i,j},\alpha_i) : i = 1,\ldots,\mathcal{I}, j = 1,\ldots,\mathcal{J}_i\}$, where $\lambda_{\max}^{\alpha_i} > \lambda_{i,1} > \cdots > \lambda_{i,\mathcal{J}_i}$ for $i = 1,\ldots,\mathcal{I}$.
Output: $\beta^*(\lambda_{i,j},\alpha_i)$ and the index set $\mathcal{G}_{i,j}$ such that $[\beta^*(\lambda_{i,j},\alpha_i)]_{\mathcal{G}_{i,j}} = 0$ for $i = 1,\ldots,\mathcal{I}$ and $j = 1,\ldots,\mathcal{J}_i$.
Initialize $\mathcal{G}_{i,j} \leftarrow \emptyset$, $i = 1,\ldots,\mathcal{I}$, $j = 1,\ldots,\mathcal{J}_i$.
for $i = 1$ to $\mathcal{I}$ do
    Compute $\lambda_{\max}^{\alpha_i}$ by Eq. (28) and set $\lambda_{i,0} \leftarrow \lambda_{\max}^{\alpha_i}$.
    Set $\theta^*(\lambda_{i,0},\alpha_i) \leftarrow \frac{\mathbf{y}}{\lambda_{i,0}}$ by Theorem 9.
    for $j = 1$ to $\mathcal{J}_i$ do
        /* compute the ball $\Theta$ that contains $\theta^*(\lambda_{i,j},\alpha_i)$ */
        Compute $\mathbf{v}_{\alpha_i}^{\perp}(\lambda_{i,j-1},\lambda_{i,j})$ by Theorem 14.
        Set the center of $\Theta$: $\mathbf{o}_{\alpha_i}(\lambda_{i,j-1},\lambda_{i,j}) \leftarrow \theta^*(\lambda_{i,j-1},\alpha_i) + \frac{1}{2}\mathbf{v}_{\alpha_i}^{\perp}(\lambda_{i,j},\lambda_{i,j-1})$.
        Set the radius of $\Theta$: $r_{\alpha_i}(\lambda_{i,j-1},\lambda_{i,j}) \leftarrow \frac{1}{2}\|\mathbf{v}_{\alpha_i}^{\perp}(\lambda_{i,j},\lambda_{i,j-1})\|$.
        for $g = 1$ to $G$ do
            Compute $s_g^*(\lambda_{i,j},\lambda_{i,j-1};\alpha_i)$ by Theorem 17.
            /* the first layer (group-level screening) of TLFre */
            if $s_g^*(\lambda_{i,j},\lambda_{i,j-1};\alpha_i) < \alpha_i\sqrt{n_g}$ then
                $\beta_g^*(\lambda_{i,j},\alpha_i) = 0$.
                Set $\mathcal{G}_{i,j} \leftarrow \mathcal{G}_{i,j} \cup \{g_k : \mathbf{x}_{g_k}$ is the $k^{th}$ column of $\mathbf{X}_g\}$.
            else
                for $k = 1$ to $n_g$ do
                    Compute $t_{g_k}^*(\lambda_{i,j},\lambda_{i,j-1};\alpha_i)$ by Theorem 18.
                    /* the second layer (feature-level screening) of TLFre */
                    if $t_{g_k}^*(\lambda_{i,j},\lambda_{i,j-1};\alpha_i) \leq 1$ then
                        $[\beta_g^*(\lambda_{i,j},\alpha_i)]_k = 0$.
                        Set $\mathcal{G}_{i,j} \leftarrow \mathcal{G}_{i,j} \cup \{g_k\}$.
                    end if
                end for
            end if
        end for
        Set $\overline{\mathcal{G}}_{i,j} \leftarrow \{k : k = 1,\ldots,p, k \notin \mathcal{G}_{i,j}\}$.
        Compute $[\beta^*(\lambda_{i,j},\alpha_i)]_{\overline{\mathcal{G}}_{i,j}}$ on the reduced data set $\mathbf{X}_{\overline{\mathcal{G}}_{i,j}}$ by any solver.
        Set $\theta^*(\lambda_{i,j},\alpha_i) \leftarrow (\mathbf{y} - \mathbf{X}\beta^*(\lambda_{i,j},\alpha_i))/\lambda_{i,j}$ by Eq. (18).
    end for
end for
```

4.1. Estimation of the Dual Optimal Solution

Due to the geometric property of the dual problem in (17), i.e., $\theta^*(\lambda,\alpha) = \mathbf{P}_{\mathcal{F}^\alpha}(\mathbf{y}/\lambda)$, we have a very useful characterization of the dual optimal solution via the so-called normal cones (Ruszczyński, 2006).

Proposition 13 (Ruszczyński, 2006; Bauschke and Combettes, 2011) *For a closed convex set $\mathcal{C} \in \mathbb{R}^n$ and a point $\mathbf{w} \in \mathcal{C}$, the normal cone to $\mathcal{C}$ at $\mathbf{w}$ is defined by*

$$N_{\mathcal{C}}(\mathbf{w}) = \{\mathbf{v} : \langle \mathbf{v}, \mathbf{w}' - \mathbf{w} \rangle \leq 0, \forall \mathbf{w}' \in \mathcal{C}\}. \tag{29}$$

Then, the following hold:

(i) $N_{\mathcal{C}}(\mathbf{w}) = \{\mathbf{v} : \mathbf{P}_{\mathcal{C}}(\mathbf{w} + \mathbf{v}) = \mathbf{w}\}$.

(ii) $\mathbf{P}_{\mathcal{C}}(\mathbf{w} + \mathbf{v}) = \mathbf{w}, \forall\, \mathbf{v} \in N_{\mathcal{C}}(\mathbf{w})$.

(iii) *Let* $\overline{\mathbf{w}} \notin \mathcal{C}$. *Then,* $\mathbf{w} = \mathbf{P}_{\mathcal{C}}(\overline{\mathbf{w}}) \Leftrightarrow \overline{\mathbf{w}} - \mathbf{w} \in N_{\mathcal{C}}(\mathbf{w})$.

(iv) *Let* $\overline{\mathbf{w}} \notin \mathcal{C}$ *and* $\mathbf{w} = \mathbf{P}_{\mathcal{C}}(\overline{\mathbf{w}})$. *Then,* $\mathbf{P}_{\mathcal{C}}(\mathbf{w} + t(\overline{\mathbf{w}} - \mathbf{w})) = \mathbf{w}$ *for all* $t \geq 0$.

By Theorem 9, $\theta^*(\bar{\lambda}, \alpha)$ is known if $\bar{\lambda} = \lambda^{\alpha}_{\max}$. Thus, we can estimate $\theta^*(\lambda, \alpha)$ in terms of $\theta^*(\bar{\lambda}, \alpha)$. Due to the same reason, *we only consider the cases with $\lambda < \lambda^{\alpha}_{\max}$ for $\theta^*(\lambda, \alpha)$* to be estimated.

Theorem 14 *For the SGL problem in (3), suppose that $\theta^*(\bar{\lambda}, \alpha)$ is known with $\bar{\lambda} \leq \lambda^{\alpha}_{\max}$. Let ρ_g, $g = 1, \ldots, G$, be defined by Theorem 9. For $\lambda \in (0, \bar{\lambda})$, let*

$$\mathbf{n}_{\alpha}(\bar{\lambda}) = \begin{cases} \frac{\mathbf{y}}{\bar{\lambda}} - \theta^*(\bar{\lambda}, \alpha), & \text{if } \bar{\lambda} < \lambda^{\alpha}_{\max}, \\ \mathbf{X}_* \mathcal{S}_1\left(\mathbf{X}_*^T \frac{\mathbf{y}}{\lambda^{\alpha}_{\max}}\right), & \text{if } \bar{\lambda} = \lambda^{\alpha}_{\max}, \end{cases} \text{ where } \mathbf{X}_* = \text{argmax}_{\mathbf{X}_g} \rho_g,$$

$$\mathbf{v}_{\alpha}(\lambda, \bar{\lambda}) = \frac{\mathbf{y}}{\lambda} - \theta^*(\bar{\lambda}, \alpha),$$

$$\mathbf{v}_{\alpha}^{\perp}(\lambda, \bar{\lambda}) = \mathbf{v}_{\alpha}(\lambda, \bar{\lambda}) - \frac{\langle \mathbf{v}_{\alpha}(\lambda, \bar{\lambda}), \mathbf{n}_{\alpha}(\bar{\lambda}) \rangle}{\|\mathbf{n}_{\alpha}(\bar{\lambda})\|^2} \mathbf{n}_{\alpha}(\bar{\lambda}).$$

Then, the following hold:

(i) $\mathbf{n}_{\alpha}(\bar{\lambda}) \in N_{\mathcal{F}^{\alpha}}(\theta^*(\bar{\lambda}, \alpha))$,

(ii) $\|\theta^*(\lambda, \alpha) - (\theta^*(\bar{\lambda}, \alpha) + \frac{1}{2}\mathbf{v}_{\alpha}^{\perp}(\lambda, \bar{\lambda}))\| \leq \frac{1}{2}\|\mathbf{v}_{\alpha}^{\perp}(\lambda, \bar{\lambda})\|$.

For notational convenience, we denote

$$\mathbf{o}_{\alpha}(\lambda, \bar{\lambda}) = \theta^*(\bar{\lambda}, \alpha) + \frac{1}{2}\mathbf{v}_{\alpha}^{\perp}(\lambda, \bar{\lambda}). \tag{30}$$

Theorem 14 shows that $\theta^*(\lambda, \alpha)$ lies inside the ball of radius $\frac{1}{2}\|\mathbf{v}_{\alpha}^{\perp}(\lambda, \bar{\lambda})\|$ centered at $\mathbf{o}_{\alpha}(\lambda, \bar{\lambda})$.

4.2. Solving for the Supreme Values via Nonconvex Optimization

We solve the optimization problems in (R1*) and (R2*). To simplify notations, let

$$\Theta = \{\theta : \|\theta - \mathbf{o}_{\alpha}(\lambda, \bar{\lambda})\| \leq \frac{1}{2}\|\mathbf{v}_{\alpha}^{\perp}(\lambda, \bar{\lambda})\|\}, \tag{31}$$

$$\Xi_g = \left\{\xi_g : \|\xi_g - \mathbf{X}_g^T \mathbf{o}_{\alpha}(\lambda, \bar{\lambda})\| \leq \frac{1}{2}\|\mathbf{v}_{\alpha}^{\perp}(\lambda, \bar{\lambda})\|\|\mathbf{X}_g\|_2\right\}, \; g = 1, \ldots, G. \tag{32}$$

Theorem 14 indicates that $\theta^*(\lambda,\alpha) \in \Theta$. Moreover, we can see that $\mathbf{X}_g^T\Theta \subseteq \Xi_g$, $g = 1,\ldots,G$. To develop the TLFre rule by (R1*) and (R2*), we need to solve the following optimization problems:

$$s_g^*(\lambda,\bar{\lambda};\alpha) = \sup_{\xi_g} \{\|\mathcal{S}_1(\xi_g)\| : \xi_g \in \Xi_g\},\ g = 1,\ldots,G, \tag{33}$$

$$t_{g_k}^*(\lambda,\bar{\lambda};\alpha) = \sup_{\theta} \{|\mathbf{x}_{g_k}^T\theta| : \theta \in \Theta\},\ k = 1,\ldots,n_g,\ g = 1,\ldots,G. \tag{34}$$

4.2.1. The Solution of Problem (33)

We consider the following equivalent problem of (33):

$$\frac{1}{2}\left(s_g^*(\lambda,\bar{\lambda};\alpha)\right)^2 = \sup_{\xi_g} \left\{\frac{1}{2}\|\mathcal{S}_1(\xi_g)\|^2 : \xi_g \in \Xi_g\right\}. \tag{35}$$

We can see that the objective function of problem (35) is *continuously differentiable* and the feasible set is a ball. Thus, problem (35) is *nonconvex* because we need to *maximize* a convex function subject to a convex set. We first derive the necessary optimality conditions in Lemma 15 and then deduce the closed form solutions of problems (33) and (35) in Theorem 17.

Lemma 15 *Let Ξ_g^* be the set of optimal solutions of (35) and $\xi_g^* \in \Xi_g^*$. Then, the following hold:*
(i) *Suppose that ξ_g^* is an interior point of Ξ_g. Then, Ξ_g is a subset of $\mathcal{B}_\infty$.*
(ii) *Suppose that ξ_g^* is a boundary point of Ξ_g. Then, there exists $\mu^* \geq 0$ such that*

$$\mathcal{S}_1(\xi_g^*) = \mu^*\left(\xi_g^* - \mathbf{X}_g^T\mathbf{o}_\alpha(\lambda,\bar{\lambda})\right). \tag{36}$$

(iii) *Suppose that there exists $\xi_g^0 \in \Xi_g$ and $\xi_g^0 \notin \mathcal{B}_\infty$. Then, we have*
(iiia) *$\xi_g^* \notin \mathcal{B}_\infty$ and ξ_g^* is a boundary point of Ξ_g, i.e.,*

$$\|\xi_g^* - \mathbf{X}_g^T\mathbf{o}_\alpha(\lambda,\bar{\lambda})\| = \frac{1}{2}\|\mathbf{v}_\alpha^\perp(\lambda,\bar{\lambda})\|\|\mathbf{X}_g\|_2.$$

(iiib) *The optimality condition in Eq. (36) holds with $\mu^* > 0$.*

To show Lemma 15, we need the following proposition.

Proposition 16 (Hiriart-Urruty, 1988) *Suppose that $h \in \Gamma_0$ and $\mathcal{C}$ is a nonempty closed convex set. If $\mathbf{w}^* \in \mathcal{C}$ is a local maximum of h on $\mathcal{C}$, then $\partial h(\mathbf{w}^*) \subseteq N_{\mathcal{C}}(\mathbf{w}^*)$.*

We now present the proof of Lemma 15.

Proof To simplify notations, let

$$\mathbf{c} = \mathbf{X}_g^T\mathbf{o}_\alpha(\lambda,\bar{\lambda}) \text{ and } r = \frac{1}{2}\|\mathbf{v}_\alpha^\perp(\lambda,\bar{\lambda})\|\|\mathbf{X}_g\|_2. \tag{37}$$

By Eq. (1), we have

$$h(\mathbf{w}) := \frac{1}{2}\|\mathcal{S}_1(\mathbf{w})\|^2 = \frac{1}{2}\sum_i \left(|[\mathbf{w}]_i| - 1\right)_+^2. \tag{38}$$

It is easy to see that $h(\cdot)$ is continuously differentiable. Indeed, we have

$$\nabla h(\mathbf{w}) = \mathcal{S}_1(\mathbf{w}). \tag{39}$$

Then, problem (35) can be written as

$$\frac{1}{2}(s_g^*(\lambda, \bar{\lambda}; \alpha))^2 = \sup_{\xi_g} \left\{ h(\xi_g) = \frac{1}{2}\sum_i ([\xi_g]_i - 1)_+^2 : \xi_g \in \Xi_g \right\}, \tag{40}$$

where $\Xi_g = \{\xi_g : \|\xi_g - \mathbf{c}\| \leq r\}$. Then, Proposition 16 results in

$$\mathcal{S}_1(\xi_g^*) = \nabla h(\xi_g^*) \subseteq \partial h(\xi_g^*) \subseteq N_{\Xi_g}(\xi_g^*). \tag{41}$$

(i) Suppose that ξ_g^* is an interior point of Ξ_g. Then, we have $N_{\Xi_g}(\xi_g^*) = 0$. By Eq. (41), we can see that

$$0 = \mathcal{S}_1(\xi_g^*) \Rightarrow 0 = \frac{1}{2}\|\mathcal{S}_1(\xi_g^*)\|^2 = \frac{1}{2}(s_g^*(\lambda, \bar{\lambda}; \alpha))^2 = \sup_{\xi_g} \left\{ \frac{1}{2}\|\mathcal{S}_1(\xi_g)\|^2 : \xi_g \in \Xi_g \right\}.$$

Therefore, we have

$$\|\mathcal{S}_1(\xi_g)\| = 0, \, \forall \, \xi_g \in \Xi_g. \tag{42}$$

Because $\mathcal{S}_1(\xi_g) = \xi_g - \mathbf{P}_{\mathcal{B}_\infty}(\xi_g)$ (see Remark 6), Eq. (42) implies that

$$\xi_g = \mathbf{P}_{\mathcal{B}_\infty}(\xi_g), \, \forall \, \xi_g \in \Xi_g \Rightarrow \xi_g \in \mathcal{B}_\infty, \, \forall \, \xi_g \in \Xi_g.$$

This completes the proof.

(ii) Suppose that ξ_g^* is a boundary point of Ξ_g. We can see that

$$N_{\Xi_g}(\xi_g^*) = \{\mu(\xi_g^* - \mathbf{c}), \, \mu \geq 0\}. \tag{43}$$

Then, Eq. (36) follows by combining Eq. (43) and the optimality condition in (41).

(iii) Suppose that there exists $\xi_g^0 \in \Xi_g$ and $\xi_g^0 \notin \mathcal{B}_\infty$.

(iiia) The definition of ξ_g^0 leads to

$$0 < \|\mathcal{S}_1(\xi_g^0)\| \leq \|\mathcal{S}_1(\xi_g^*)\| \Rightarrow \xi_g^* \notin \mathcal{B}_\infty.$$

Moreover, we can see that ξ_g^* is a boundary point of Ξ_g. Because if ξ_g^* is an interior point of Ξ_g, the first part implies that $\Xi_g \subset \mathcal{B}_\infty$. This contradicts with the existence of ξ_g^0. Thus, ξ_g^* must be a boundary point of Ξ_g, i.e. $\|\xi_g^* - \mathbf{c}\| = r$.

(iiib) Because ξ_g^* is a boundary point of Ξ_g, the second part implies that Eq. (36) holds. Moreover, from (iiia), we know that $\xi_g^* \notin \mathcal{B}_\infty$. Therefore, both sides of Eq. (36) are nonzero and thus $\mu^* > 0$. This completes the proof.

■

Based on the necessary optimality conditions in Lemma 15, we derive the closed form solutions of (33) and (35) in the following Theorem. The notations are the same as the ones in the proof of Lemma 15 [see Eq. (37) and Eq. (38)].

Theorem 17 *For problems (33) and (35), let* $\mathbf{c} = \mathbf{X}_g^T \mathbf{o}_\alpha(\lambda, \bar{\lambda})$*,* $r = \frac{1}{2}\|\mathbf{v}_\alpha^\perp(\lambda, \bar{\lambda})\| \|\mathbf{X}_g\|_2$ *and* Ξ_g^* *be the set of the optimal solutions.*

(i) *Suppose that* $\mathbf{c} \notin \mathcal{B}_\infty$*, i.e.,* $\|\mathbf{c}\|_\infty > 1$*. Let* $\mathbf{u} = r\mathcal{S}_1(\mathbf{c})/\|\mathcal{S}_1(\mathbf{c})\|$*. Then,*

$$s_g^*(\lambda, \bar{\lambda}; \alpha) = \|\mathcal{S}_1(\mathbf{c})\| + r \ \text{ and } \ \Xi_g^* = \{\mathbf{c} + \mathbf{u}\}. \tag{44}$$

(ii) *Suppose that* $\mathbf{c}$ *is a boundary point of* $\mathcal{B}_\infty$*, i.e.,* $\|\mathbf{c}\|_\infty = 1$*. Then,*

$$s_g^*(\lambda, \bar{\lambda}; \alpha) = r \ \text{ and } \ \Xi_g^* = \{\mathbf{c} + \mathbf{u} : \mathbf{u} \in N_{\mathcal{B}_\infty}(\mathbf{c}), \|\mathbf{u}\| = r\}. \tag{45}$$

(iii) *Suppose that* $\mathbf{c} \in \operatorname{int}\mathcal{B}_\infty$*, i.e.,* $\|\mathbf{c}\|_\infty < 1$*. Let* $i^* \in \mathcal{I}^* = \{i : |[\mathbf{c}]_i| = \|\mathbf{c}\|_\infty\}$*. Then,*

$$s_g^*(\lambda, \bar{\lambda}; \alpha) = (\|\mathbf{c}\|_\infty + r - 1)_+, \tag{46}$$

$$\Xi_g^* = \begin{cases} \Xi_g, & \text{if } \Xi_g \subset \mathcal{B}_\infty, \\ \{\mathbf{c} + r \cdot \operatorname{sgn}([\mathbf{c}]_{i^*})\mathbf{e}_{i^*} : i^* \in \mathcal{I}^*\}, & \text{if } \Xi_g \not\subset \mathcal{B}_\infty \text{ and } \mathbf{c} \neq 0, \\ \{r \cdot \mathbf{e}_{i^*}, -r \cdot \mathbf{e}_{i^*} : i^* \in \mathcal{I}^*\}, & \text{if } \Xi_g \not\subset \mathcal{B}_\infty \text{ and } \mathbf{c} = 0, \end{cases}$$

where $\mathbf{e}_i$ *is the* i^{th} *standard basis vector.*

Proof

(i) Suppose that $\mathbf{c} \notin \mathcal{B}_\infty$. By the third part of Lemma 15, we have

$$\xi_g^* \notin \mathcal{B}_\infty, \ \ \|\xi_g^* - \mathbf{c}\| = r, \tag{47}$$

$$\xi_g^* - \mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*) = \mathcal{S}_1(\xi_g^*) = \mu^*(\xi_g^* - \mathbf{c}), \ \mu^* > 0. \tag{48}$$

By Eq. (48), we can see that $\mu^* \neq 1$ because otherwise we would have $\mathbf{c} = \mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*) \in \mathcal{B}_\infty$. Moreover, we can only consider the cases with $\mu^* > 1$ because $\|\mathcal{S}_1(\xi_g^*)\| = \mu^* r$ and we aim to maximize $\|\mathcal{S}_1(\xi_g^*)\|$. Therefore, if we can find a solution with $\mu^* > 1$, there is no need to consider the cases with $\mu^* \in (0, 1)$.

Suppose that $\mu^* > 1$. Then, Eq. (48) leads to

$$\mathbf{c} = \mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*) + \left(1 - \frac{1}{\mu^*}\right)\left(\xi_g^* - \mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*)\right), \tag{49}$$

$$\xi_g^* = \mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*) + \frac{\mu^*}{\mu^* - 1}\left(\mathbf{c} - \mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*)\right). \tag{50}$$

In view of part (iv) of Proposition 13 and Eq. (49), we have

$$\mathbf{P}_{\mathcal{B}_\infty}(\mathbf{c}) = \mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*). \tag{51}$$

Therefore, Eq. (50) can be rewritten as

$$\mathcal{S}_1(\xi_g^*) = \xi_g^* - \mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*) = \frac{\mu^*}{\mu^*-1}\left(\mathbf{c} - \mathbf{P}_{\mathcal{B}_\infty}(\mathbf{c})\right) = \frac{\mu^*}{\mu^*-1}\mathcal{S}_1(\mathbf{c}). \tag{52}$$

Combining Eq. (48) and Eq. (52), we have

$$\frac{\mu^*}{\mu^*-1}\|\mathcal{S}_1(\mathbf{c})\| = \mu^*\|\xi_g^* - \mathbf{c}\| = \mu^* r \Rightarrow \mu^* = 1 + \frac{\|\mathcal{S}_1(\mathbf{c})\|}{r} > 1. \tag{53}$$

The statement holds by plugging Eq. (53) and Eq. (51) into Eq. (50) and Eq. (52). Moreover, the above discussion implies that Ξ_g^* only contains one element as shown in Eq. (44).

(ii) Suppose that $\mathbf{c}$ is a boundary point of $\mathcal{B}_\infty$. Then, we can find a point $\xi_g^0 \in \Xi_g$ and $\xi_g^0 \notin \mathcal{B}_\infty$. By the third part of Lemma 15, we also have Eq. (47) and Eq. (48) hold. We claim that $\mu^* \in (0,1]$. The argument is as follows.

Suppose that $\mu^* > 1$. By the same argument as in the proof of the first part, we can see that Eq. (52) holds. Because $\mathcal{S}_1(\xi_g^*) \neq 0$ by Eq. (47), we have $\mathcal{S}_1(\mathbf{c}) \neq 0$. This implies that $\mathbf{c} \notin \mathcal{B}_\infty$. Thus, we have a contradiction, which implies that $\mu^* \in (0,1]$.

Let us consider the cases with $\mu^* = 1$. Because $\|\mathcal{S}_1(\xi_g^*)\| = \mu^* r$ [see Eq. (48)] and we want to maximize $\|\mathcal{S}_1(\xi_g^*)\|$, there is no need to consider the cases with $\mu^* \in (0,1)$ if we can find solutions of problem (33) with $\mu^* = 1$. Therefore, Eq. (48) leads to

$$\mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*) = \mathbf{c}.$$

By part (iii) of Proposition 13, we can see that

$$\mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*) = \mathbf{c} \Leftrightarrow \xi_g^* - \mathbf{c} \in N_{\mathcal{B}_\infty}(\mathbf{c}). \tag{54}$$

Combining Eq. (54) and Eq. (47), the statement holds immediately, which confirms that $\mu^* = 1$.

(iii) Suppose that $\mathbf{c}$ is an interior point of $\mathcal{B}_\infty$.

(a) We first consider the cases with $\Xi_g \subset \mathcal{B}_\infty$. Then, we can see that

$$\mathcal{S}_1(\xi) = 0,\ \forall \xi \in \Xi_g \Rightarrow \Xi_g^* = \Xi_g.$$

In other words, an arbitrary point of Ξ_g is an optimal solution of problem (33). Thus, we have

$$\begin{aligned}&\mathbf{c} + r \cdot \mathrm{sgn}(\mathbf{e}_{i^*})\mathbf{e}_{i^*} \in \Xi_g^*,\\ &s_g^*(\lambda, \bar{\lambda}; \alpha) = 0.\end{aligned}$$

On the other hand, we can see that

$$\mathbf{c} - r\mathbf{e}_i \in \Xi_g \subset \mathcal{B}_\infty,\ \mathbf{c} + r\mathbf{e}_i \in \Xi_g \subset \mathcal{B}_\infty,\ i = 1, \ldots, n_g \Rightarrow \|\mathbf{c}\|_\infty + r \leq 1.$$

Therefore, we have

$$(\|\mathbf{c}\|_\infty + r - 1)_+ = 0,$$

and thus

$$s_g^*(\lambda, \bar{\lambda}; \alpha) = (\|\mathbf{c}\|_\infty + r - 1)_+.$$

(b) Suppose that $\Xi_g \not\subset \mathcal{B}_\infty$, i.e., there exists $\xi^0 \in \Xi_g$ such that $\xi^0 \notin \mathcal{B}_\infty$. By the third part of Lemma 15, we have Eq. (47) and Eq. (48) hold. Moreover, in view of the proof of the first and second part, we can see that $\mu^* \in (0,1)$. Therefore, Eq. (48) leads to

$$(1-\mu^*)\xi_g^* + \mu^*\mathbf{c} = \mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*). \tag{55}$$

By rearranging the terms of Eq. (55), we have

$$\mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*) - \mathbf{c} = (1-\mu^*)(\xi_g^* - \mathbf{c}). \tag{56}$$

Because $\mu^* \in (0,1)$, Eq. (55) implies that $\mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*)$ lies on the line segment connecting ξ_g^* and $\mathbf{c}$. Thus, we have

$$\|\xi_g^* - \mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*)\| + \|\mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*) - \mathbf{c}\| = \|\xi_g^* - \mathbf{c}\| = r. \tag{57}$$

Therefore, to maximize $\|\mathcal{S}_1(\xi_g^*)\| = \|\xi_g^* - \mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*)\|$, we need to minimize $\|\mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*) - \mathbf{c}\|$. Because $\xi_g^* \notin \mathcal{B}_\infty$, we can see that $\mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*)$ is a boundary point of $\mathcal{B}_\infty$. Therefore, we need to solve the following minimization problem:

$$\min_{\phi_g} \{\|\phi_g - \mathbf{c}\| : \|\phi_g\|_\infty = 1\}. \tag{58}$$

Suppose that $\mathbf{c} = 0$. We can see that the set of optimal solutions of problem (58) is

$$\Phi_g^* = \{\mathbf{e}_i\}_{i=1}^{n_g} \cup \{-\mathbf{e}_i\}_{i=1}^{n_g}.$$

For each $\phi_g^* \in \Phi_g^*$, we set it as $\mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*)$. In view of Eq. (56) and Eq. (47), the statement follows immediately.

Suppose that $\mathbf{c} \neq 0$. Recall that $\mathcal{I}^* = \{i^* : |[\mathbf{c}]_{i^*}| = \|\mathbf{c}\|_\infty\}$. It is easy to see that

$$\Phi_g^* = \left\{ \phi_{i^*} : [\phi_{i^*}]_k = \begin{cases} \text{sgn}([\mathbf{c}]_{i^*}), & \text{if } k = i^*, \\ [\mathbf{c}]_k, & \text{otherwise,} \end{cases} \quad i^* \in \mathcal{I}^* \right\}.$$

We can see that

$$\phi_{i^*} - \mathbf{c} = (1 - |[\mathbf{c}]_{i^*}|)\text{sgn}([\mathbf{c}]_{i^*})\mathbf{e}_{i^*},\ i^* \in \mathcal{I}^*.$$

For each ϕ_{i^*}, we set it to $\mathbf{P}_{\mathcal{B}_\infty}(\xi_g^*)$. Then, we can see that the statement holds by Eq. (56) and Eq. (47). This completes the proof.

■

4.2.2. The Solution of Problem (34)

Problem (34) can be solved directly via the Cauchy-Schwarz inequality.

Theorem 18 *For problem (34), we have*

$$t^*_{g_k}(\lambda, \bar{\lambda}; \alpha) = |\mathbf{x}_{g_k}^T \mathbf{o}_\alpha(\lambda, \bar{\lambda})| + \frac{1}{2}\|\mathbf{v}_\alpha^\perp(\lambda, \bar{\lambda})\|\|\mathbf{x}_{g_k}\|.$$

We are now ready to present the proposed screening rule TLFre.

4.3. The Proposed Two-Layer Screening Rules

To develop the two-layer screening rules for SGL, we only need to plug the supreme values $s_g^*(\lambda, \bar{\lambda}; \alpha)$ and $t^*_{g_k}(\lambda, \bar{\lambda}; \alpha)$ in (R1*) and (R2*). We present the TLFre rule as follows.

Theorem 19 *For the SGL problem in (3), suppose that we are given a grid of parameter values $\{\alpha_i\}_{i=1}^{\mathcal{I}}$ and $\lambda_{\max}^{\alpha_i} = \lambda_{i,0} > \lambda_{i,1} > \ldots > \lambda_{i,\mathcal{J}_i}$ for each α_i. Moreover, assume that $\beta^*(\lambda_{i,j-1}, \alpha_i)$ is known for an integer $0 < j < \mathcal{J}_i$. Let $\theta^*(\lambda_{i,j-1}, \alpha_i)$, $\mathbf{v}_{\alpha_i}^\perp(\lambda_{i,j}, \lambda_{i,j-1})$ and $s_g^*(\lambda_{i,j}, \lambda_{i,j-1}; \alpha_i)$ be given by Eq. (18), Theorems 14 and 17, respectively. Then, for $g = 1, \ldots, G$, the following holds*

$$s_g^*(\lambda_{i,j}, \lambda_{i,j-1}; \alpha_i) < \alpha_i \sqrt{n_g} \Rightarrow \beta_g^*(\lambda_{i,j}, \alpha_i) = 0. \tag{$\mathcal{L}_1$}$$

For the g^{th} group that does not pass the rule in ($\mathcal{L}_1$), we have $[\beta_g^*(\lambda_{i,j}, \alpha_i)]_k = 0$ if

$$t^*_{g_k}(\lambda_{i,j}, \lambda_{i,j-1}; \alpha_i) \leq 1, \tag{$\mathcal{L}_2$}$$

where

$$\begin{aligned} &t^*_{g_k}(\lambda_{i,j}, \lambda_{i,j-1}; \alpha_i) \\ &= \left| \mathbf{x}_{g_k}^T \left(\frac{\mathbf{y} - \mathbf{X}\beta^*(\lambda_{i,j-1}, \alpha_i)}{\lambda_{i,j-1}} + \frac{1}{2}\mathbf{v}_{\alpha_i}^\perp(\lambda_{i,j}, \lambda_{i,j-1}) \right) \right| + \frac{1}{2}\|\mathbf{v}_{\alpha_i}^\perp(\lambda_{i,j}, \lambda_{i,j-1})\|\|\mathbf{x}_{g_k}\|. \end{aligned}$$

($\mathcal{L}_1$) and ($\mathcal{L}_2$) are the first and second layer screening rules of TLFre, respectively.

We also write Theorem 19 in an algorithmic manner in Algorithm 2. For each pair of parameter values $(\lambda_{i,j}, \alpha_i)$, we first apply TLFre to identify the inactive groups and inactive features, namely, the zero components of $\beta^*(\lambda_{i,j}, \alpha_i)$. Then, we remove the inactive groups and inactive features from the data matrix and apply an arbitrary solver to solve the SGL problem on the remaining features.

Specifically, lines 7 to 9 in Algorithm 2 compute the ball that contains $\theta^*(\lambda_{i,j}, \alpha_i)$ in terms of $\theta^*(\lambda_{i,j-1}, \alpha_i)$ (see remark 20). Lines 13 till 15 apply the first layer of TLFre, i.e., the group-level screening, to identify the inactive groups. Take the g^{th} group for an example. If the first layer identifies the g^{th} group as an inactive group, we can set $\beta_g^*(\lambda_{i,j}, \alpha_i) = 0$. Otherwise, lines 20 till 23 apply the second layer of TLFre, i.e., the feature-level screening, to identify the inactive features in the g^{th} group. The index set $\mathcal{G}_{i,j}$ stores the indices of inactive features, i.e., if $k \in \mathcal{G}_{i,j}$, then $[\beta^*(\lambda_{i,j}, \alpha_i)]_k = 0$. After we scan the entire matrix, the index set $\mathcal{G}_{i,j}$ contains all the indices of inactive features identified by TLFre, i.e., $[\beta^*(\lambda_{i,j}, \alpha_i)]_{\mathcal{G}_{i,j}} = 0$. Thus, the remaining unknowns are indeed $[\beta^*(\lambda_{i,j}, \alpha_i)]_{\overline{\mathcal{G}}_{i,j}}$ (see line

27). Then, we apply an arbitrary solver to solve for $[\beta^*(\lambda_{i,j},\alpha_i)]_{\overline{\mathcal{G}}_{i,j}}$ on the reduced data matrix $\mathbf{X}_{\overline{\mathcal{G}}_{i,j}}$, where $\mathbf{X}_{\overline{\mathcal{G}}_{i,j}} = (\mathbf{x}_{k_1},\ldots,\mathbf{x}_{k_q})$, $q = |\overline{\mathcal{G}}_{i,j}|$, and $k_\ell \in \overline{\mathcal{G}}_{i,j}$ for $\ell = 1,\ldots,q$. Once we have solved for $[\beta^*(\lambda_{i,j},\alpha_i)]_{\overline{\mathcal{G}}_{i,j}}$, we indeed have solved for $\beta^*(\lambda_{i,j},\alpha_i)$ as $[\beta^*(\lambda_{i,j},\alpha_i)]_{\mathcal{G}_{i,j}} = 0$. Then, line 29 computes $\theta^*(\lambda_{i,j},\alpha_i)$ by Eq. (18), based on which we can estimate $\theta^*(\lambda_{i,j+1},\alpha_i)$ and apply TLFre to identify the zero components of $\beta^*(\lambda_{i,j+1},\alpha_i)$.

Remark 20 *As shown by Theorem 19 and Algorithm 2, TLFre estimates the dual optimum at $(\lambda_{i,j},\alpha_i)$, i.e., $\theta^*(\lambda_{i,j},\alpha_i)$, in terms of a known dual optimum at a different pair of parameter values $(\lambda_{i,j-1},\alpha_i)$, i.e., $\theta^*(\lambda_{i,j-1},\alpha_i)$. Then, we can apply TLFre to identify the inactive groups and inactive features and solve the TLFre problems on a reduced data matrix. Thus, to initialize TLFre, we may need to solve the SGL problem on the entire data matrix once to compute $\theta^*(\lambda_{i,0},\alpha_i)$, which can be time consuming. However, we note that, Theorem 9 not only gives the effective interval of λ for a fixed value of α, but also a closed form solution of the dual optimum for any $\lambda \geq \lambda_{\max}^{\alpha}$. Thus, as we have done in Algorithm 2 (see line 3), we can always set $\lambda_{i,0} = \lambda_{\max}^{\alpha_i}$ and $\theta^*(\lambda_{i,0},\alpha_i) = \mathbf{y}/\lambda_{i,0}$ to initialize the computation of TLFre. This implies that, if we combine TLFre and an arbitrary solver, we do not need to solve the SGL problem on the entire data matrix even once.*

5. Extension to Nonnegative Lasso

The framework of TLFre is applicable to a large class of sparse models with multiple regularizers. As an example, we extend TLFre to nonnegative Lasso:

$$\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\beta\right\|^2 + \lambda\|\beta\|_1 : \beta \in \mathbb{R}_+^p \right\}, \tag{59}$$

where $\lambda > 0$ is the regularization parameter and $\mathbb{R}_+^p$ is the nonnegative orthant of $\mathbb{R}^p$. In Section 5.1, we transform the constraint $\beta \in \mathbb{R}_+^p$ to a regularizer and derive the Fenchel's dual of the nonnegative Lasso problem. We then motivate the screening method—called DPC since the key step is to **d**ecom**p**ose a **c**onvex set via Fenchel's Duality Theorem—via the KKT conditions in Section 5.2. In Section 5.3, we analyze the geometric properties of the dual problem and derive the set of parameter values leading to zero solutions. We then develop the screening method for nonnegative Lasso in Section 5.4.

5.1. The Fenchel's Dual of Nonnegative Lasso

Let $\mathbf{I}_{\mathbb{R}_+^p}$ be the indicator function of $\mathbb{R}_+^p$. By noting that $\mathbf{I}_{\mathbb{R}_+^p} = \lambda\mathbf{I}_{\mathbb{R}_+^p}$ for any $\lambda > 0$, we can rewrite the nonnegative Lasso problem in (59) as

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\beta\right\|^2 + \lambda\|\beta\|_1 + \lambda\mathbf{I}_{\mathbb{R}_+^p}(\beta). \tag{60}$$

In other words, we incorporate the constraint $\beta \in \mathbb{R}_+^p$ to the objective function as an additional regularizer. As a result, the nonnegative lasso problem in (60) has two regularizers. Thus, similar to SGL, we can derive the Fenchel's dual of nonnegative Lasso via Theorem 1.

We now proceed by following a similar procedure as the one in Section 3.1. We note that the nonnegative Lasso problem in (60) can also be formulated as the one in (10) with $f(\cdot) = \frac{1}{2}\|\cdot\|^2$ and $\Omega(\beta) = \|\beta\|_1 + \mathbf{I}_{\mathbb{R}^p_+}(\beta)$. To derive the Fenchel's dual of nonnegative Lasso, we need to find f^* and Ω^* by Theorem 1. Since we have already seen that $f^*(\cdot) = \frac{1}{2}\|\cdot\|^2$ in Section 3.1, we only need to find $\Omega^*(\cdot)$. The following result is indeed a counterpart of Lemma 3.

Lemma 21 *Let $\Omega_2(\beta) = \|\beta\|_1$, $\Omega_3 = \mathbf{I}_{\mathbb{R}^p_+}(\beta)$, and $\Omega(\beta) = \Omega_2(\beta) + \Omega_3(\beta)$. Then,*

(i) *$(\Omega_2)^*(\xi) = \mathbf{I}_{\mathcal{B}_\infty}(\xi)$ and $(\Omega_3)^*(\xi) = \mathbf{I}_{\mathbb{R}^p_-}(\xi)$, where $\mathbb{R}^p_-$ is the nonpositive orthant of $\mathbb{R}^p$.*

(ii) *$\Omega^*(\xi) = ((\Omega_2)^* \boxdot (\Omega_3)^*)(\xi) = \mathbf{I}_{\mathbb{R}^p_-}(\xi - \mathbf{1})$, where $\mathbb{R}^p \ni \mathbf{1} = (1, 1, \dots, 1)^T$.*

We omit the proof of Lemma 21 since it is very similar to that of Lemma 3.

Remark 22 *Consider the second part of Lemma 21. Let $\mathcal{C}_{\mathbf{1}} = \{\xi : \xi \leq \mathbf{1}\}$, where "$\leq$" is defined component-wisely. We can see that*

$$\mathbf{I}_{\mathbb{R}^p_-}(\xi - \mathbf{1}) = \mathbf{I}_{\mathcal{C}_{\mathbf{1}}}(\xi).$$

On the other hand, Lemma 7 implies that

$$\Omega^*(\xi) = ((\Omega_2)^* \boxdot (\Omega_3)^*)(\xi) = \mathbf{I}_{\mathcal{B}_\infty + \mathbb{R}^p_-}(\xi).$$

Thus, we have $\mathcal{B}_\infty + \mathbb{R}^p_- = \mathcal{C}_{\mathbf{1}}$. The second part of Lemma 21 decomposes each $\xi \in \mathcal{B}_\infty + \mathbb{R}^p_-$ into two components: $\mathbf{1}$ and $\xi - \mathbf{1}$ that belong to $\mathcal{B}_\infty$ and $\mathbb{R}^p_-$, respectively.

By Theorem 1 and Lemma 21, we can derive the Fenchel's dual of nonnegative Lasso in the following theorem (which is indeed the counterpart of Theorem 5).

Theorem 23 *For the nonnegative Lasso problem, the following hold:*

(i) *The Fenchel's dual of nonnegative Lasso is given by:*

$$\inf_{\theta} \left\{ \frac{1}{2} \left\| \frac{\mathbf{y}}{\lambda} - \theta \right\|^2 - \frac{1}{2}\|\mathbf{y}\|^2 : \langle \mathbf{x}_i, \theta \rangle \leq 1, \, i = 1, \dots, p \right\}. \tag{61}$$

(ii) *Let $\beta^*(\lambda)$ and $\theta^*(\lambda)$ be the optimal solutions of problems (60) and (61), respectively.Then,*

$$\lambda\theta^*(\lambda) = \mathbf{y} - \mathbf{X}\beta^*(\lambda), \tag{62}$$

$$\mathbf{X}^T\theta^*(\lambda) \in \partial\|\beta^*(\lambda)\|_1 + \partial\mathbf{I}_{\mathbb{R}^p_+}(\beta^*(\lambda)). \tag{63}$$

We omit the proof of Theorem 23 since it is very similar to that of Theorem 5.

5.2. Motivation of the Screening Method via KKT Conditions

The key to develop the DPC rule for nonnegative lasso is the KKT condition in (63). We can see that $\partial\|\mathbf{w}\|_1 = \mathrm{SGN}(\mathbf{w})$ and

$$\partial \mathbf{I}_{\mathbb{R}^p_+}(\mathbf{w}) = \left\{ \xi \in \mathbb{R}^p : [\xi]_i = \begin{cases} 0, & \text{if } [\mathbf{w}]_i > 0, \\ \rho,\ \rho \le 0, & \text{if } [\mathbf{w}]_i = 0, \end{cases} \right\}.$$

Therefore, the KKT condition in (63) implies that

$$\langle \mathbf{x}_i, \theta^*(\lambda)\rangle \in \begin{cases} 1, & \text{if } [\beta^*(\lambda)]_i > 0, \\ \varrho,\ \varrho \le 1, & \text{if } [\beta^*(\lambda)]_i = 0. \end{cases} \tag{64}$$

By Eq. (64), we have the following rule:

$$\langle \mathbf{x}_i, \theta^*(\lambda)\rangle < 1 \Rightarrow [\beta^*(\lambda)]_i = 0. \tag{R3}$$

Because $\theta^*(\lambda)$ is unknown, we can apply (R3) to identify the inactive features—which have 0 coefficients in $\beta^*(\lambda)$. Similar to TLFre, we can first find a region Θ that contains $\theta^*(\lambda)$. Then, we can relax (R3) as follows:

$$\sup_{\theta\in\Theta} \langle \mathbf{x}_i, \theta\rangle < 1 \Rightarrow [\beta^*(\lambda)]_i = 0. \tag{R3*}$$

Inspired by (R3*), we develop DPC via the following three steps:

Step 1. Given λ, we estimate a region Θ that contains $\theta^*(\lambda)$.

Step 2. We solve the optimization problem $\omega_i = \sup_{\theta\in\Theta} \langle \mathbf{x}_i, \theta\rangle$.

Step 3. By plugging in ω_i computed from **Step 2**, (R3*) leads to the desired screening method DPC for nonnegative Lasso.

5.3. Geometric Properties of the Fenchel's Dual of Nonnegative Lasso

In view of the Fenchel's dual of nonnegative Lasso in (61), we can see that the optimal solution is indeed the projection of $\mathbf{y}/\lambda$ onto the feasible set $\mathcal{F} = \{\theta : \langle \mathbf{x}_i, \theta\rangle \le 1, i = 1, \ldots, p\}$, i.e.,

$$\theta^*(\lambda) = \mathbf{P}_{\mathcal{F}}\left(\frac{\mathbf{y}}{\lambda}\right). \tag{65}$$

Therefore, if $\mathbf{y}/\lambda \in \mathcal{F}$, Eq. (65) implies that $\theta^*(\lambda) = \mathbf{y}/\lambda$. If further $\mathbf{y}/\lambda$ is an interior point of $\mathcal{F}$, R3* implies that $\beta^*(\lambda) = 0$. The next theorem gives the set of parameter values leading to 0 solutions of nonnegative Lasso.

Theorem 24 *For the nonnegative Lasso problem (60), Let $\lambda_{\max} = \max_i \langle \mathbf{x}_i, \mathbf{y}\rangle$. Then, the following statements are equivalent:*

(i) $\dfrac{\mathbf{y}}{\lambda} \in \mathcal{F}$, (ii) $\theta^*(\lambda) = \dfrac{\mathbf{y}}{\lambda}$, (iii) $\beta^*(\lambda) = 0$, (iv) $\lambda \ge \lambda_{\max}$.

We omit the proof of Theorem 24 since it is very similar to that of Theorem 9.

5.4. The Proposed Screening Rule for Nonnegative Lasso

We follow the three steps in Section 5.2 to develop the screening rule for nonnegative Lasso. We first estimate a region that contains $\theta^*(\lambda)$. Because $\theta^*(\lambda)$ admits a closed form solution with $\lambda \geq \lambda_{\max}$ by Theorem 24, we focus on the cases with $\lambda < \lambda_{\max}$.

Theorem 25 *For the nonnegative Lasso problem, suppose that $\theta^*(\bar\lambda)$ is known with $\bar\lambda \leq \lambda_{\max}$. For any $\lambda \in (0, \bar\lambda)$, we define*

$$\mathbf{n}(\bar\lambda) = \begin{cases} \frac{\mathbf{y}}{\bar\lambda} - \theta^*(\bar\lambda), & \text{if } \bar\lambda < \lambda_{\max}^{\alpha}, \\ \mathbf{x}_*, & \text{if } \bar\lambda = \lambda_{\max}, \end{cases} \quad \text{where } \mathbf{x}_* = \operatorname{argmax}_{\mathbf{x}_i} \langle \mathbf{x}_i, \mathbf{y} \rangle,$$

$$\mathbf{v}(\lambda, \bar\lambda) = \frac{\mathbf{y}}{\lambda} - \theta^*(\bar\lambda),$$

$$\mathbf{v}(\lambda, \bar\lambda)^{\perp} = \mathbf{v}(\lambda, \bar\lambda) - \frac{\langle \mathbf{v}(\lambda, \bar\lambda), \mathbf{n}(\bar\lambda)\rangle}{\|\mathbf{n}(\bar\lambda)\|^2}\mathbf{n}(\bar\lambda).$$

Then, the following hold:

(i) $\mathbf{n}(\bar\lambda) \in N_{\mathcal{F}}(\theta^*(\bar\lambda))$,

(ii) $\left\|\theta^*(\lambda) - \left(\theta^*(\bar\lambda) + \frac{1}{2}\mathbf{v}^{\perp}(\lambda, \bar\lambda)\right)\right\| \leq \frac{1}{2}\|\mathbf{v}^{\perp}(\lambda, \bar\lambda)\|$.

Proof We only show that $\mathbf{n}(\lambda_{\max}) \in N_{\mathcal{F}}(\theta^*(\lambda_{\max}))$ since the proof of the other statement is very similar to that of Theorem 14.

By Proposition 13 and Theorem 24, it suffices to show that

$$\langle \mathbf{x}_*, \theta - \mathbf{y}/\lambda_{\max}\rangle \leq 0, \forall\, \theta \in \mathcal{F}. \tag{66}$$

Because $\theta \in \mathcal{F}$, we have $\langle \mathbf{x}_*, \theta\rangle \leq 1$. The definition of $\mathbf{x}_*$ implies that $\langle \mathbf{x}_*, \mathbf{y}/\lambda_{\max}\rangle = 1$. Thus, the inequality in (66) holds, which completes the proof. ■

Theorem 25 implies that $\theta^*(\lambda)$ is in a ball—denoted by $\mathcal{B}(\lambda, \bar\lambda)$—of radius $\frac{1}{2}\|\mathbf{v}^{\perp}(\lambda, \bar\lambda)\|$ centered at $\theta^*(\bar\lambda) + \frac{1}{2}\mathbf{v}^{\perp}(\lambda, \bar\lambda)$. Simple calculations lead to

$$\omega_i = \sup_{\theta \in \mathcal{B}(\lambda, \bar\lambda)} \langle \mathbf{x}_i, \theta\rangle = \left\langle \mathbf{x}_i, \theta^*(\bar\lambda) + \frac{1}{2}\mathbf{v}^{\perp}(\lambda, \bar\lambda)\right\rangle + \frac{1}{2}\|\mathbf{v}^{\perp}(\lambda, \bar\lambda)\|\|\mathbf{x}_i\|. \tag{67}$$

By plugging ω_i into (R3*), we have the DPC screening rule for nonnegative Lasso as follows.

Theorem 26 *For the nonnegative Lasso problem, suppose that we are given a sequence of parameter values $\lambda_{\max} = \lambda^{(0)} > \lambda^{(1)} > \ldots > \lambda^{(\mathcal{J})}$. Then, $[\beta^*(\lambda^{(j+1)})]_i = 0$ if $\beta^*(\lambda^{(j)})$ is known and the following holds:*

$$\left\langle \mathbf{x}_i, \frac{\mathbf{y} - \mathbf{X}\beta^*(\lambda^{(j)})}{\lambda^{(j)}} + \frac{1}{2}\mathbf{v}^{\perp}(\lambda^{(j+1)}, \lambda^{(j)})\right\rangle + \frac{1}{2}\|\mathbf{v}^{\perp}(\lambda^{(j+1)}, \lambda^{(j)})\|\|\mathbf{x}_i\| < 1. \tag{68}$$

6. Experiments

We evaluate TLFre for SGL and DPC for nonnegative Lasso in Sections 6.1 and 6.2, respectively, on both synthetic and real data sets. To the best of knowledge, the TLFre and DPC are the first screening methods for SGL and nonnegative Lasso, respectively. The code is available at `http://dpc-screening.github.io/`.

6.1. TLFre for SGL

We perform experiments to evaluate TLFre on synthetic and real data sets in Sections 6.1.1 and 6.1.2, respectively. To measure the performance of TLFre, we compute the *rejection ratios* of $(\mathcal{L}_1)$ and $(\mathcal{L}_2)$, respectively. Specifically, let m be the number of features that have 0 coefficients in the solution, $\overline{\mathcal{G}}$ be the index set of groups that are discarded by $(\mathcal{L}_1)$ and $\overline{p}$ be the number of inactive features that are detected by $(\mathcal{L}_2)$. The rejection ratios of $(\mathcal{L}_1)$ and $(\mathcal{L}_2)$ are defined by $r_1 = \frac{\sum_{g \in \overline{\mathcal{G}}} n_g}{m}$ and $r_2 = \frac{|\overline{p}|}{m}$, respectively. Moreover, we report the *speedup* gained by TLFre, i.e., the ratio of the running time of solver without screening to the running time of solver with TLFre.

To determine appropriate values of α and λ by cross validation or stability selection, we can run TLFre with as many parameter values as we need. Given a data set, for illustrative purposes only, we select seven values of α from $\{\tan(\psi) : \psi = 5^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 85^\circ\}$. Then, for each value of α, we run TLFre along a sequence of 100 values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}^{\alpha}$ from 1 to 0.01. Thus, 700 pairs of parameter values of (λ, α) are sampled in total.

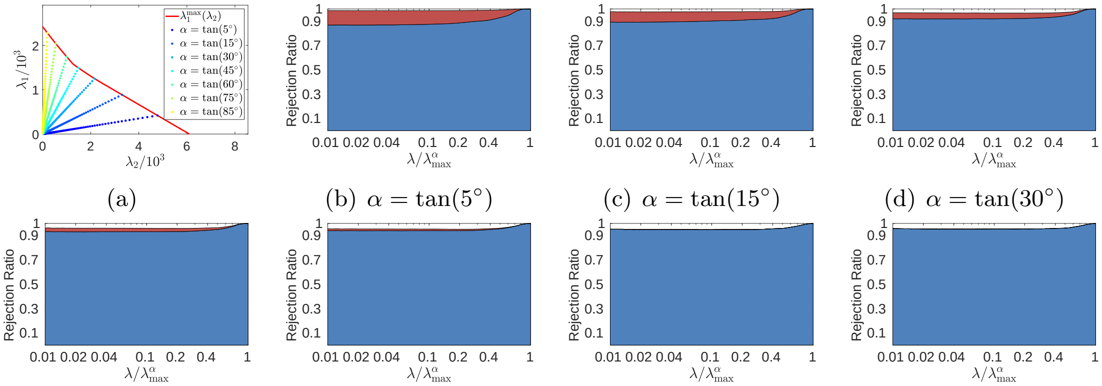
We use *sgLeastR* from the SLEP package (Liu et al., 2009) as the solver for SGL, which is one of the state-of-the-arts (Zhang et al., 2018b) [see Section G for a comparison between sgLeastR and another popular solver (Lin et al., 2014)].

For the non-screening case, we apply sgLeastR directly to solve SGL with different parameter values. We use zero as the initial point.

6.1.1. Simulation Studies

We perform experiments on two synthetic data sets that are commonly used in the literature (Tibshirani et al., 2012; Zou and Hastie, 2005). The true model is $\mathbf{y} = \mathbf{X}\beta^* + 0.01\epsilon$, $\epsilon \sim N(0,1)$. We generate two data sets with 1000×160000 entries: Synthetic 1 and Synthetic 2. We randomly divide the 160000 features into 16000 groups. For Synthetic 1, the entries of the data matrix $\mathbf{X}$ are i.i.d. standard Gaussian with pairwise correlation zero, i.e., $\mathrm{corr}(\mathbf{x}_i, \mathbf{x}_i) = 0$. For Synthetic 2, the entries of the data matrix $\mathbf{X}$ are drawn from i.i.d. standard Gaussian with pairwise correlation $0.5^{|i-j|}$, i.e., $\mathrm{corr}(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}$. To construct β^*, we first randomly select γ_1 percent of groups. Then, for each selected group, we randomly select γ_2 percent of features. The selected components of β^* are populated from a standard Gaussian and the remaining ones are set to 0. We set $\gamma_1 = \gamma_2 = 10$ for Synthetic 1 and $\gamma_1 = \gamma_2 = 20$ for Synthetic 2.

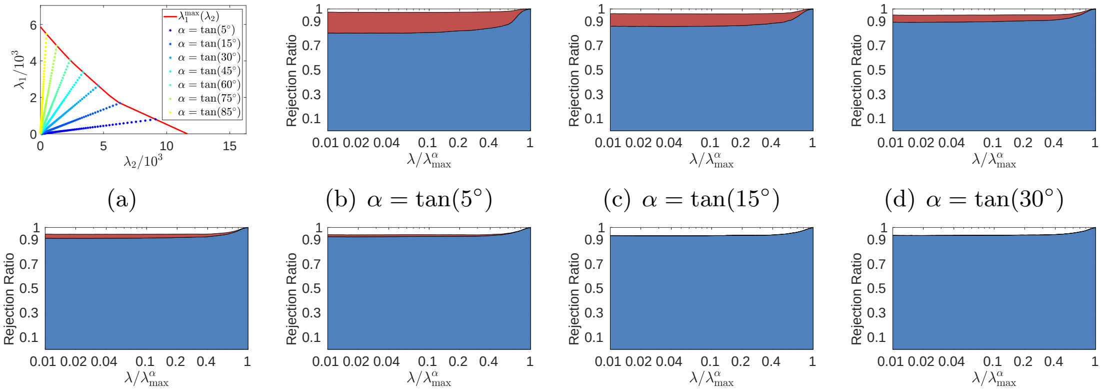
Fig. 1(a) and Fig. 2(a) show the plots of $\lambda_1^{\max}(\lambda_2)$ (see Corollary 12) and the sampled parameter values of λ and α (recall that $\lambda_1 = \alpha\lambda$ and $\lambda_2 = \lambda$). For the other figures, the blue and red regions represent the rejection ratios of $(\mathcal{L}_1)$ and $(\mathcal{L}_2)$, respectively. We can see that TLFre is very effective in discarding inactive groups/features; that is, more than



(a) (b) $\alpha = \tan(5°)$ (c) $\alpha = \tan(15°)$ (d) $\alpha = \tan(30°)$

(e) $\alpha = \tan(45°)$ (f) $\alpha = \tan(60°)$ (g) $\alpha = \tan(75°)$ (h) $\alpha = \tan(85°)$

Figure 1: Rejection ratios of TLFre on the Synthetic 1 data set.



(a) (b) $\alpha = \tan(5°)$ (c) $\alpha = \tan(15°)$ (d) $\alpha = \tan(30°)$

(e) $\alpha = \tan(45°)$ (f) $\alpha = \tan(60°)$ (g) $\alpha = \tan(75°)$ (h) $\alpha = \tan(85°)$

Figure 2: Rejection ratios of TLFre on the Synthetic 2 data set.

90% of inactive features can be detected. Moreover, we can observe that the first layer screening ($\mathcal{L}_1$) becomes more effective with a larger α. Intuitively, this is because the group Lasso penalty plays a more important role in enforcing the sparsity with a larger value of α (recall that $\lambda_1 = \alpha\lambda$). The top and middle parts of Table 1 indicate that the speedup gained by TLFre is very significant (up to 80 times) and TLFre is very efficient. Compared to the running time of the solver without screening, the running time of TLFre is negligible. The running time of TLFre includes that of computing $\|\mathbf{X}_g\|_2$, $g = 1, \ldots, G$, which can be efficiently computed by the power method (Halko et al., 2011). Indeed, this can be shared for TLFre with different parameter values.

6.1.2. Experiments on Real Data Sets

We perform experiments on two commonly used real data sets – the Alzheimer's Disease Neuroimaging Initiative (ADNI) data set (`http://adni.loni.usc.edu/`) and the news20.binary (Chang and Lin, 2011) data set. Details of these data sets are as follows.

Table 1: Running time (in seconds) for solving SGL along a sequence of 100 tuning parameter values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}^{\alpha}$ from 1.0 to 0.01 by (a): the solver (Liu et al., 2009) without screening; (b): the solver combined with TLFre. The data sets are Synthetic 1 and Synthetic 2.

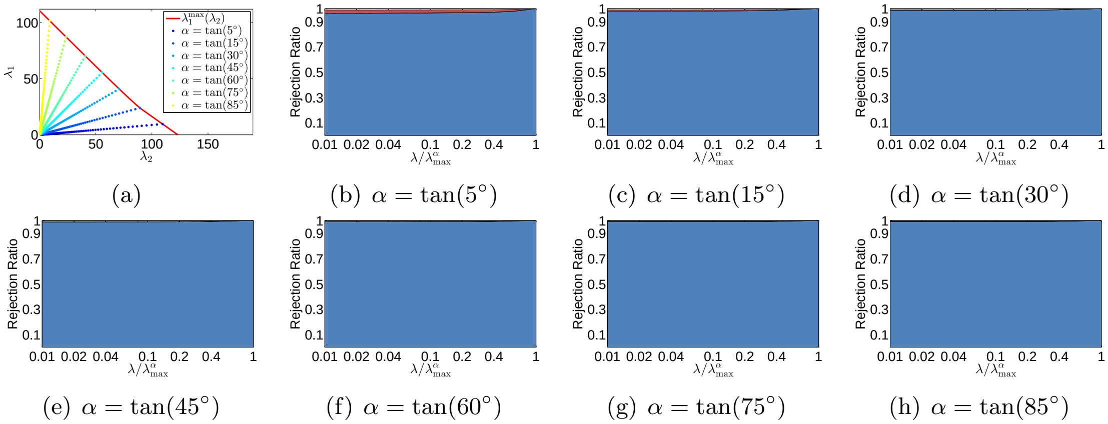
	α	tan(5°)	tan(15°)	tan(30°)	tan(45°)	tan(60°)	tan(75°)	tan(85°)
Synthetic 1	solver	15555.28	16124.08	16106.24	16293.04	16426.44	16836.16	16862.36
	TLFre	37.84	43.08	46.92	51.16	54.24	53.4	53.24
	TLFre+solver	184.16	275.36	680.08	1196.04	1465.16	1629.96	1657.00
	speedup	**84.46**	**58.55**	**23.68**	**13.62**	**11.21**	**10.33**	**10.18**
Synthetic 2	solver	15709.72	16615.16	16286.04	16826.48	16919.41	17178.08	17350.36
	TLFre	41.72	47.28	54.08	54.72	59.08	58.56	60.12
	TLFre+solver	328.52	906.72	1452.28	1702.61	1912.76	2181.23	2180.24
	speedup	**47.82**	**18.32**	**11.21**	**9.88**	**8.85**	**7.88**	**7.96**

Table 2: Running time (in seconds) for solving SGL along a sequence of 100 tuning parameter values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}^{\alpha}$ from 1.0 to 0.01 by (a): the solver (Liu et al., 2009) without screening; (b): the solver combined with TLFre. We perform experiments on the ADNI data sets. The response vectors are GMV and WMV, respectively.

	α	tan(5°)	tan(15°)	tan(30°)	tan(45°)	tan(60°)	tan(75°)	tan(85°)
ADNI+GMV	solver	30652.56	30755.63	30838.29	31096.10	30850.78	30728.27	30572.35
	TLFre	64.08	64.56	64.96	65.00	64.89	65.17	65.05
	TLFre+solver	372.04	383.17	386.80	402.72	391.63	385.98	382.62
	speedup	**82.39**	**80.27**	**79.73**	**77.22**	**78.78**	**79.61**	**79.90**
ADNI+WMV	solver	29751.27	29823.15	29927.52	30078.62	30115.89	29927.58	29896.77
	TLFre	62.91	63.33	63.39	63.99	64.13	64.31	64.36
	TLFre+solver	363.43	364.78	386.15	393.03	395.87	400.11	399.48
	speedup	**81.86**	**81.76**	**77.50**	**76.53**	**76.08**	**74.80**	**74.84**

Table 3: Running time (in seconds) for solving SGL along a sequence of 100 tuning parameter values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}^{\alpha}$ from 1.0 to 0.01 by (a): the solver (Liu et al., 2009) without screening; (b): the solver combined with TLFre. We perform experiments on the news20.binary data sets.

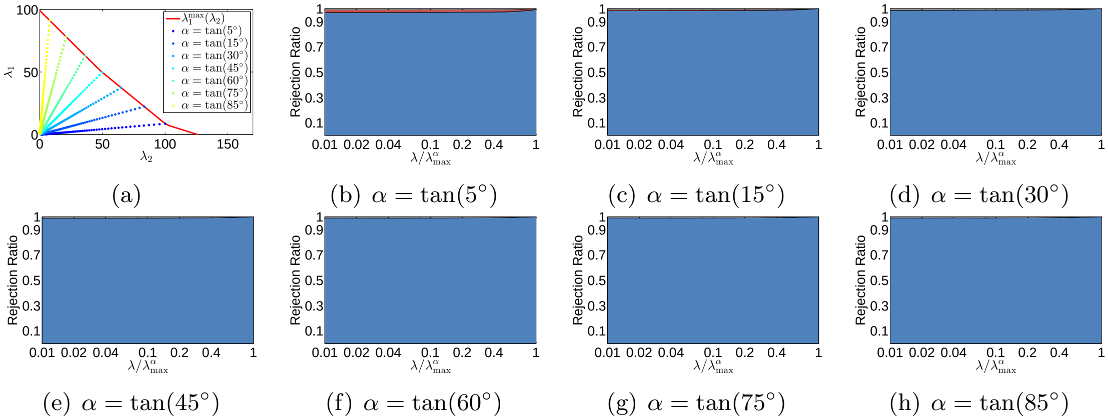
	α	tan(5°)	tan(15°)	tan(30°)	tan(45°)	tan(60°)	tan(75°)	tan(85°)
news20.binary	solver	1233401.05	1231570.22	1277630.92	1299353.68	1292879.86	1216554.09	1347890.85
	TLFre	350.51	337.75	332.01	346.98	352.78	353.52	362.43
	TLFre+solver	1434.49	1465.37	1539.37	1598.87	1608.90	1659.78	1709.35
	speedup	**859.82**	**840.45**	**829.97**	**812.67**	**803.58**	**793.21**	**788.54**



(a) (b) $\alpha = \tan(5°)$ (c) $\alpha = \tan(15°)$ (d) $\alpha = \tan(30°)$

(e) $\alpha = \tan(45°)$ (f) $\alpha = \tan(60°)$ (g) $\alpha = \tan(75°)$ (h) $\alpha = \tan(85°)$

Figure 3: Rejection ratios of TLFre on the ADNI data set with grey matter volume as response.



(a) (b) $\alpha = \tan(5°)$ (c) $\alpha = \tan(15°)$ (d) $\alpha = \tan(30°)$

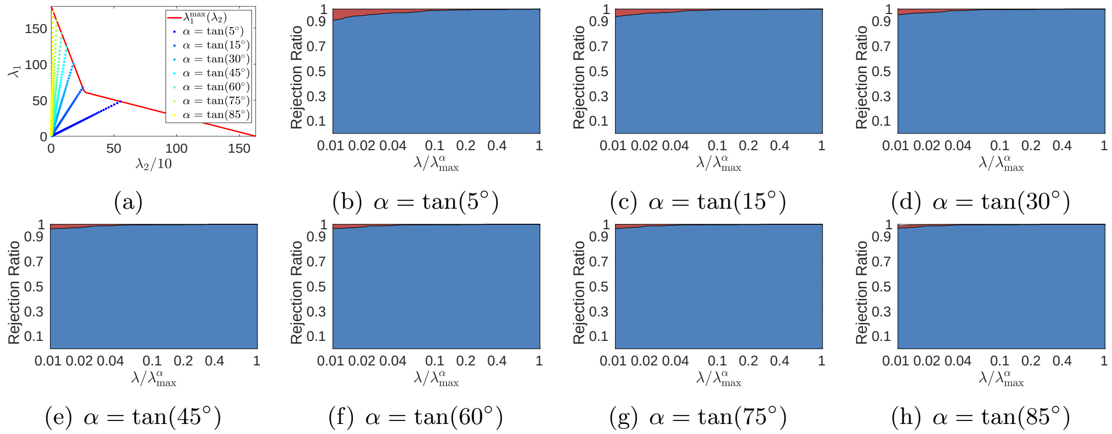
(e) $\alpha = \tan(45°)$ (f) $\alpha = \tan(60°)$ (g) $\alpha = \tan(75°)$ (h) $\alpha = \tan(85°)$

Figure 4: Rejection ratios of TLFre on the ADNI data set with white matter volume as response.

ADNI The data matrix of the ADNI data set consists of 747 samples with 426040 single nucleotide polymorphisms (SNPs), which are divided into 94765 groups. The response vectors are the grey matter volume (GMV) and white matter volume (WMV), respectively.

news20.binary The news20.binary data set consists of 19996 samples with 1355191 features, which are divided into 67760 groups. The entries of the response vectors are the labels of the corresponding samples, which are 1 or -1.

Fig. 3(a), Fig. 4(a), and Fig. 5(a) show the plots of $\lambda_1^{\max}(\lambda_2)$ (see Corollary 12) and the sampled parameter values of α and λ. The other figures present the rejection ratios of $(\mathcal{L}_1)$ and $(\mathcal{L}_2)$ by blue and red regions, respectively. We can see that almost all of the inactive groups/features are discarded by TLFre. The rejection ratios of $r_1 + r_2$ are very close to 1 in all cases. Tables 2 and 3 show that TLFre leads to a very significant speedup (about 80 times on the ADNI data set and 800 times on the news20.binary data set). In other words, the solver without screening needs about 8.5 and 360 hours to solve the 100 SGL problems for



(a) (b) $\alpha = \tan(5°)$ (c) $\alpha = \tan(15°)$ (d) $\alpha = \tan(30°)$

(e) $\alpha = \tan(45°)$ (f) $\alpha = \tan(60°)$ (g) $\alpha = \tan(75°)$ (h) $\alpha = \tan(85°)$

Figure 5: Rejection ratios of TLFre on the news20.binary data set.

each value of α on the ADNI and news20.binary data set, respectively. However, combined with TLFre, the solver needs only 6∼8 minutes and 24∼29 minutes, respectively. Moreover, we can observe that the computational cost of TLFre is negligible compared to that of the solver without screening. This demonstrates the efficiency of TLFre.

6.2. DPC for Nonnegative Lasso

In this experiment, we evaluate the performance of DPC on two synthetic data sets and six real data sets. We integrate DPC with the solver, nnLeastR, (Liu et al., 2009) to solve the nonnegative Lasso problem along a sequence of 100 parameter values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}$ from 1.0 to 0.01. The two synthetic data sets are the same as the ones we used in Section 6.1.1. To construct β^*, we first randomly select 10 percent of features. The corresponding components of β^* are populated from a standard Gaussian and the remaining ones are set to 0.

We use *nnLeastR* from the SLEP package (Liu et al., 2009) as the solver for nonnegative Lasso, which is one of the state-of-the-arts [see Section G for a comparison between nnLeastR and another popular solver (Lin et al., 2014)].

For the non-screening case, we apply nnLeastR directly to solve SGL with different parameter values. We use zero as the initial point.

We list the six real data sets and the corresponding experimental settings as follows.

Breast Cancer data set (West et al., 2001; Shevade and Keerthi, 2003): this data set contains 7129 gene expression values of 44 tumor samples (thus the data matrix $\mathbf{X}$ is of 44×7129). The response vector $\mathbf{y} \in \{1, -1\}^{44}$ contains the binary label of each sample.

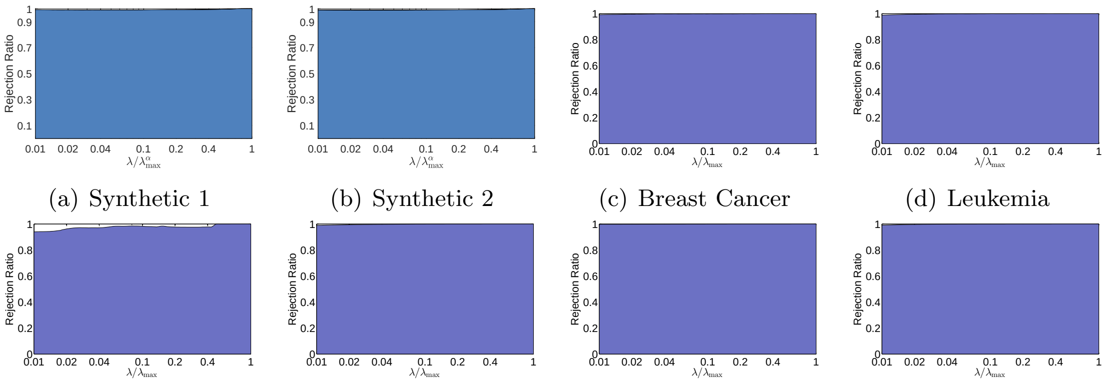
Leukemia data set (Armstrong et al., 2002): this data set contains 11225 gene expression values of 52 samples ($\mathbf{X} \in \mathbb{R}^{52\times 11225}$). The response vector $\mathbf{y}$ contains the binary label of each sample.

Prostate Cancer data set (Petricoin et al., 2002): this data set contains 15154 measurements of 132 patients ($\mathbf{X} \in \mathbb{R}^{132\times 15154}$). By protein mass spectrometry, the features are indexed by time-of-flight values, which are related to the mass over charge ratios of the constituent proteins in the blood. The response vector $\mathbf{y}$ contains the binary label of each sample.

PIE face image data set (Sim et al., 2003; Cai et al., 2007): this data set contains 11554 gray face images (each has 32×32 pixels) of 68 people, taken under different poses, illumination conditions and expressions. In each trial, we first randomly pick an image as the response $\mathbf{y} \in \mathbb{R}^{1024}$, and then use the remaining images to form the data matrix $\mathbf{X} \in \mathbb{R}^{1024\times 11553}$. We run 100 trials and report the average performance of DPC.

MNIST handwritten digit data set (Lecun et al., 1998): this data set contains grey images of scanned handwritten digits (each has 28×28 pixels). The training and test sets contain $60,000$ and $10,000$ images, respectively. We first randomly select 5000 images for each digit from the training set and get a data matrix $\mathbf{X} \in \mathbb{R}^{784\times 50000}$. Then, in each trial, we randomly select an image from the testing set as the response $\mathbf{y} \in \mathbb{R}^{784}$. We run 100 trials and report the average performance of the screening rules.

Street View House Number (SVHN) data set (Netzer et al., 2001): this data set contains color images of street view house numbers (each has 32×32 pixels), including 73257 images for training and 26032 for testing. In each trial, we first randomly select an image as the response $\mathbf{y} \in \mathbb{R}^{3072}$, and then use the remaining ones to form the data matrix $\mathbf{X} \in \mathbb{R}^{3072\times 99288}$. We run 20 trials and report the average performance.



(a) Synthetic 1 (b) Synthetic 2 (c) Breast Cancer (d) Leukemia

(e) Prostate Cancer (f) PIE (g) MNIST (h) SVHN

Figure 6: Rejection ratios of DPC on eight data sets.

Table 4: Running time (in seconds) for solving nonnegative Lasso along a sequence of 100 tuning parameter values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}$ from 1.0 to 0.01 by (a): the solver (Liu et al., 2009) without screening; (b): the solver combined with DPC.

	Synthetic 1	Synthetic 2	Breast Cancer	Leukemia	Prostate Cancer	PIE	MNIST	SVHN
solver	13140.84	13853.84	23.40	34.04	187.82	674.04	3000.69	24761.07
DPC	3.08	3.59	0.03	0.06	0.23	1.16	3.53	30.59
DPC+solver	61.56	69.52	2.18	3.37	6.37	5.01	9.31	104.93
speedup	**213.52**	**199.30**	**10.73**	**10.10**	**29.49**	**134.54**	**322.31**	**235.98**

We present the *rejection ratios*—the ratio of the number of inactive features identified by DPC to the actual number of inactive features—in Fig. 6. We also report the running time of the solver with and without DPC, the time for running DPC, and the corresponding *speedup* in Table 4.

Fig. 6 shows that DPC is very effective in identifying the inactive features even for small parameter values: the rejection ratios are very close to 100% for the entire sequence of parameter values on the eight data sets. Table 4 shows that DPC leads to a very significant speedup on all the data sets. Take MNIST as an example. The solver without DPC takes 50 minutes to solve the 100 nonnegative Lasso problems. However, combined with DPC, the solver only needs 10 seconds. The speedup gained by DPC on the MNIST data set is thus more than 300 times. Similarly, on the SVHN data set, the running time for solving the 100 nonnegative Lasso problems by the solver without DPC is close to seven hours. However, combined with DPC, the solver takes less than two minutes to solve all the 100 nonnegative Lasso problems, leading to a speedup about 230 times. Moreover, we observe that the computational cost of DPC is very low—which is negligible compared to that of the solver without DPC.

7. Conclusion

In this paper, we propose a novel feature reduction method for SGL via decomposition of convex sets. We also derive the set of parameter values that lead to zero solutions of SGL. To the best of our knowledge, TLFre is the first method which is applicable to sparse models with multiple sparsity-inducing regularizers. More importantly, the proposed approach provides novel framework for developing screening methods for complex sparse models with multiple sparsity-inducing regularizers, e.g., ℓ_1 SVM that performs both sample and feature selection, fused Lasso and tree Lasso with more than two regularizers. To demonstrate the flexibility of the proposed framework, we develop the DPC screening rule for the nonnegative Lasso problem. Experiments on both synthetic and real data sets demonstrate the effectiveness and efficiency of TLFre and DPC. We plan to generalize the idea of TLFre to ℓ_1 SVM, fused Lasso and tree Lasso, which are expected to consist of multiple layers of screening.

Acknowledgments

This work was supported in part by NSFC (61822604, 61836006).

Appendix A. The Lagrangian Dual Problem of SGL

We derive the dual problem of SGL in (4) via the Lagrangian multiplier method.

By introducing an auxiliary variable

$$\mathbf{z} = \mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g \beta_g, \tag{69}$$

the SGL problem in (3) becomes:

$$\min_{\beta} \left\{ \frac{1}{2}\|\mathbf{z}\|^2 + \alpha\lambda \sum_{g=1}^{G} \sqrt{n_g}\|\beta_g\| + \lambda\|\beta\|_1 : \mathbf{z} = \mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g\beta_g \right\}.$$

Let $\lambda\theta$ be the Lagrangian multiplier, the Lagrangian function is

$$\begin{aligned} L(\beta, \mathbf{z}; \theta) =& \frac{1}{2}\|\mathbf{z}\|^2 + \alpha\lambda \sum_{g=1}^{G} \sqrt{n_g}\|\beta_g\| + \lambda\|\beta\|_1 + \langle \lambda\theta, \mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g\beta_g - \mathbf{z}\rangle & (70) \\ =& \alpha\lambda \sum_{g=1}^{G} \sqrt{n_g}\|\beta_g\| + \lambda\|\beta\|_1 - \lambda\langle \theta, \sum_{g=1}^{G} \mathbf{X}_g\beta_g\rangle + \frac{1}{2}\|\mathbf{z}\|^2 - \lambda\langle\theta, \mathbf{z}\rangle + \lambda\langle\theta, \mathbf{y}\rangle. & (71) \end{aligned}$$

Let

$$\begin{aligned} f_1(\beta) =& \sum_{g=1}^{G} f_1^g(\beta_g) = \sum_{g=1}^{G} \left(\alpha\lambda\sqrt{n_g}\|\beta_g\| + \lambda\|\beta_g\|_1 - \lambda\langle\theta, \mathbf{X}_g\beta_g\rangle\right), \\ f_2(\mathbf{z}) =& \frac{1}{2}\|\mathbf{z}\|^2 - \lambda\langle\theta, \mathbf{z}\rangle. \end{aligned}$$

To derive the dual problem, we need to minimize the Lagrangian function with respect to β and $\mathbf{z}$. In other words, we need to minimize f_1 and f_2, respectively. We first consider

$$\min_{\beta_g} f_1^g(\beta_g) = \alpha\lambda\sqrt{n_g}\|\beta_g\| + \lambda\|\beta_g\|_1 - \lambda\langle\theta, \mathbf{X}_g\beta_g\rangle.$$

By the Fermat's rule, we have

$$0 \in \partial f_1^g(\beta_g) = \alpha\lambda\sqrt{n_g}\partial\|\beta_g\| + \lambda\partial\|\beta_g\|_1 - \lambda\mathbf{X}_g^T\theta, \tag{72}$$

which leads to

$$\mathbf{X}_g^T\theta = \alpha\sqrt{n_g}\zeta_1 + \zeta_2,\ \ \zeta_1 \in \partial\|\beta_g\|,\ \zeta_2 \in \partial\|\beta_g\|_1. \tag{73}$$

By noting that

$$\langle\zeta_1, \beta_g\rangle = \|\beta_g\|,\ \ \langle\zeta_2, \beta_g\rangle = \|\beta_g\|_1,$$

we have

$$\langle\mathbf{X}_g^T\theta, \beta_g\rangle = \alpha\sqrt{n_g}\|\beta_g\| + \|\beta_g\|_1.$$

Thus, we can see that

$$0 = \min_{\beta_g} f_1^g(\beta_g). \tag{74}$$

Moreover, because $\zeta_1 \in \partial\|\beta_g\|$, $\zeta_2 \in \partial\|\beta_g\|_1$, Eq. (73) implies that

$$\mathbf{X}_g^T\theta \in \alpha\sqrt{n_g}\mathcal{B} + \mathcal{B}_\infty. \tag{75}$$

To minimize f_2, the Fermat's rule results in

$$\mathbf{z} = \lambda\theta, \tag{76}$$

and thus

$$-\frac{\lambda^2}{2}\|\theta\|^2 = \min_{\mathbf{z}} f_2(z). \tag{77}$$

In view of Eq. (70), Eq. (74), Eq. (77) and Eq. (75), the dual problem of SGL can be written as

$$\sup_{\theta} \left\{ \frac{1}{2}\|\mathbf{y}\|^2 - \frac{1}{2}\left\|\theta - \frac{\mathbf{y}}{\lambda}\right\|^2 : \mathbf{X}_g^T\theta \in \alpha\sqrt{n_g}\mathcal{B} + \mathcal{B}_\infty,\, g = 1, \ldots, G \right\},$$

which is equivalent to (4).

Recall that $\beta^*(\lambda, \alpha)$ and $\theta^*(\lambda, \alpha)$ are the primal and dual optimal solutions of SGL, respectively. By Eq. (69), Eq. (72) and Eq. (76), we can see that the KKT conditions are

$$\begin{aligned} \lambda\theta^*(\lambda, \alpha) =& \mathbf{y} - \mathbf{X}\beta^*(\lambda, \alpha), \\ \mathbf{X}_g^T\theta^*(\lambda, \alpha) \in& \alpha\sqrt{n_g}\partial\|\beta_g^*(\lambda, \alpha)\| + \partial\|\beta_g^*(\lambda, \alpha)\|_1,\ g = 1, \ldots, G. \end{aligned}$$

Appendix B. Proof of Theorem 5

To show Theorem 5, we need the Fenchel-Young inequality as follows:

Lemma 27 [Fenchel-Young inequality] (Borwein and Lewis, 2006) *Any point* $\mathbf{z} \in \mathbb{R}^n$ *and* $\mathbf{w}$ *in the domain of a function* $h : \mathbb{R}^n \to (-\infty, \infty]$ *satisfy the inequality*

$$h(\mathbf{w}) + h^*(\mathbf{z}) \geq \langle \mathbf{w}, \mathbf{z} \rangle.$$

Equality holds if and only if $\mathbf{z} \in \partial h(\mathbf{w})$.

We now give the proof of Theorem 5.

Proof We first show the first part. Combining Theorem 1 and Lemma 3, the Fenchel's dual of SGL can be written as:

$$\sup_{\theta} -\frac{\lambda^2}{2}\|\theta\|^2 - \sum\nolimits_{g=1}^{G} \lambda \mathbf{I}_{\mathcal{B}}\left(\frac{\mathbf{X}_g^T\theta - \mathbf{P}_{\mathcal{B}_\infty}(\mathbf{X}_g^T\theta)}{\alpha\sqrt{n_g}}\right) + \lambda\langle \mathbf{y}, \theta \rangle,$$

which is equivalent to problem (17).

To show the second half, we have the following inequalities by Fenchel-Young inequality:

$$f(\mathbf{y} - \mathbf{X}\beta) + f^*(\lambda\theta) \geq \langle \mathbf{y} - \mathbf{X}\beta, \lambda\theta \rangle, \tag{78}$$

$$\lambda\Omega(\beta) + \lambda\Omega^*(\mathbf{X}^T\theta) \geq \lambda\langle \beta, \mathbf{X}^T\theta \rangle. \tag{79}$$

We sum the inequalities in (78) and (79) together and get

$$f(\mathbf{y} - \mathbf{X}\beta) + \lambda\Omega(\beta) \geq -f^*(\lambda\theta) - \lambda\Omega^*(\mathbf{X}^T\theta) + \lambda\langle \mathbf{y}, \theta \rangle. \tag{80}$$

Clearly, the left and right hand sides of inequality (80) are the objective functions of the pair of Fenchel's problems. Because $\mathrm{dom}\, f = \mathbb{R}^N$ and $\mathrm{dom}\, \Omega = \mathbb{R}^p$, we have

$$0 \in \mathrm{int}\,(\mathrm{dom}\, f - \mathbf{y} + \mathbf{X}\mathrm{dom}\, \Omega).$$

Thus, the equality in (80) holds at $\beta^*(\lambda, \alpha)$ and $\theta^*(\lambda, \alpha)$, i.e.,

$$f(\mathbf{y} - \mathbf{X}\beta^*(\lambda, \alpha)) + \lambda\Omega(\beta^*(\lambda, \alpha)) = -f^*(\lambda\theta^*(\lambda, \alpha)) - \lambda\Omega^*(\mathbf{X}^T\theta^*(\lambda, \alpha)) + \lambda\langle \mathbf{y}, \theta^*(\lambda, \alpha)\rangle.$$

Therefore, the equality holds in both (78) and (79) at $\beta^*(\lambda, \alpha)$ and $\theta^*(\lambda, \alpha)$. By applying Lemma 27 again, we have

$$\begin{aligned} \lambda\theta^*(\lambda, \alpha) &\in \partial f(\mathbf{y} - \mathbf{X}\beta^*(\lambda, \alpha)) = \mathbf{y} - \mathbf{X}\beta^*(\lambda, \alpha), \\ \mathbf{X}^T\theta^*(\lambda, \alpha) &\in \partial\Omega(\beta^*(\lambda, \alpha)) = \partial\Omega_1^\alpha(\beta^*(\lambda, \alpha)) + \partial\Omega_2(\beta^*(\lambda, \alpha)), \end{aligned}$$

which completes the proof. ■

Appendix C. Proof of Theorem 9

Proof The equivalence between (i) and (ii) can be see from the fact that

$$\theta^*(\lambda, \alpha) = \mathbf{P}_{\mathcal{F}^\alpha}(\mathbf{y}/\lambda).$$

Next, we show (ii)⇔(iii). Let us first show (ii)⇒(iii). We assume that $\theta^*(\lambda, \alpha) = \mathbf{y}/\lambda$. By the KKT condition in (18), we have $\mathbf{X}\beta^*(\lambda, \alpha) = 0$. We claim that $\beta^*(\lambda, \alpha) = 0$. To see this, let $\beta' \neq 0$ with $\mathbf{X}\beta' = 0$ be another optimal solution of SGL. We denote by h the objective function of SGL in (3). Then, we have

$$h(0) = \frac{1}{2}\|\mathbf{y}\|^2 < h(\beta') = \frac{1}{2}\|\mathbf{y}\|^2 + \lambda_1 \sum_g \sqrt{n_g}\|\beta'_g\| + \lambda_2\|\beta'\|_1,$$

which contradicts with the assumption $\beta' \neq 0$ is also an optimal solution. This contradiction indicates that $\beta^*(\lambda, \alpha)$ must be 0. The converse direction, i.e., (ii)⇐(iii), can be derived directly from the KKT condition in Eq. (18).

Finally, we show the equivalence (i)⇔(iv). Indeed, in view of the dual problem in (21), we can see that $\mathbf{y}/\lambda \in \mathcal{F}^\alpha$ if and only if

$$\|\mathcal{S}_1(\mathbf{X}_g^T\mathbf{y}/\lambda)\| \leq \alpha\sqrt{n_g},\ g = 1, \ldots, G. \tag{81}$$

We note that $\|\mathcal{S}_1(\mathbf{X}_g^T\mathbf{y}/\lambda)\|$ is monotonically decreasing with respect to λ. Thus, the inequality in (81) is equivalent to (iv), which completes the proof. ■

Appendix D. Proof of Corollary 12

Before we prove Corollary 12, we first derive the Fenchel's dual of (2). By letting $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$ and $\Omega(\beta) = \lambda_1 \sum_{g=1}^{G} \sqrt{n_g}\|\beta_g\| + \lambda_2\|\beta\|_1$, the SGL problem in (2) can be written as:

$$\min_{\beta}\; f(\mathbf{y} - \mathbf{X}\beta) + \Omega(\beta).$$

Then, by Fenchel's Duality Theorem, the Fenchel's dual problem of (2) is

$$\inf_{\theta} \left\{ \frac{1}{2}\|\mathbf{y} - \theta\|^2 - \frac{1}{2}\|\mathbf{y}\|^2 : \left\|\mathcal{S}_{\lambda_2}(\mathbf{X}_g^T\theta)\right\| \leq \lambda_1\sqrt{n_g},\, g = 1, \ldots, G \right\}. \tag{82}$$

Let $\bar{\beta}^*(\lambda_1, \lambda_2)$ and $\bar{\theta}^*(\lambda_1, \lambda_2)$ be the optimal solutions of problem (2) and (82). The optimality conditions can be written as

$$\bar{\theta}^*(\lambda_1, \lambda_2) = \mathbf{y} - \mathbf{X}\bar{\beta}^*(\lambda_1, \lambda_2), \tag{83}$$

$$\mathbf{X}_g^T\bar{\theta}^*(\lambda_1, \lambda_2) \in \lambda_1\sqrt{n_g}\partial\|\bar{\beta}_g^*(\lambda_1, \lambda_2)\| + \lambda_2\partial\|\bar{\beta}_g^*(\lambda_1, \lambda_2)\|_1,\, g = 1, \ldots, G. \tag{84}$$

We denote by $\mathcal{F}(\lambda_1, \lambda_2)$ the feasible set of problem (82). It is easy to see that

$$\bar{\theta}^*(\lambda_1, \lambda_2) = \mathbf{P}_{\mathcal{F}(\lambda_1, \lambda_2)}(\mathbf{y}).$$

We now present the proof of Corollary 12.

Proof For notational convenience, let

(i). $\mathbf{y} \in \mathcal{F}(\lambda_1, \lambda_2)$,

(ii). $\bar{\theta}^*(\lambda_1, \lambda_2) = \mathbf{y}$,

(iii). $\bar{\beta}^*(\lambda_1, \lambda_2) = 0$,

(iv). $\lambda_1 \geq \lambda_1^{\max}(\lambda_2) = \max_g \frac{1}{\sqrt{n_g}}\|\mathcal{S}_{\lambda_2}(\mathbf{X}_g^T\mathbf{y})\|$.

The first half of the statement is (iii)⇔(iv). Indeed, by a similar argument as in the proof of Theorem 9, we can see that the above statements are all equivalent to each other.

We now show the second half. We first show that

$$\lambda_1 \geq \lambda_1^{\max} \Rightarrow \bar{\beta}^*(\lambda_1, \lambda_2) = 0. \tag{85}$$

By the first half, we only need to show

$$\lambda_1 \geq \lambda_1^{\max} \Rightarrow \mathbf{y} \in \mathcal{F}(\lambda_1, \lambda_2).$$

Indeed, the definition of λ_1 implies that

$$\|\mathbf{X}_g^T\mathbf{y}\| \leq \lambda_1\sqrt{n_g},\, g = 1, \ldots, G.$$

We note that for any $\lambda_2 \geq 0$, we have

$$\|\mathcal{S}_{\lambda_2}(\mathbf{X}_g^T\mathbf{y})\| \leq \|\mathbf{X}_g^T\mathbf{y}\|.$$

Therefore, we can see that

$$\|\mathcal{S}_{\lambda_2}(\mathbf{X}_g^T\mathbf{y})\| \leq \|\mathbf{X}_g^T\mathbf{y}\| \leq \lambda_1\sqrt{n_g},\, g = 1, \ldots, G \Rightarrow \mathbf{y} \in \mathcal{F}(\lambda_1, \lambda_2).$$

The proof of (85) is complete.

Similarly, to show that $\lambda_2 \geq \lambda_2^{\max} \Rightarrow \bar{\beta}^*(\lambda_1, \lambda_2)$, we only need to show

$$\lambda_2 \geq \lambda_2^{\max} \Rightarrow \mathbf{y} \in \mathcal{F}(\lambda_1, \lambda_2).$$

By the definition of λ_2, we can see that

$$\|\mathbf{X}_g^T\mathbf{y}\|_\infty \leq \lambda_2,\, g = 1, \ldots, G \Rightarrow \|\mathcal{S}_{\lambda_2}(\mathbf{X}_g^T\mathbf{y})\| = 0 \leq \lambda_1\sqrt{n_g},\, g = 1, \ldots, G.$$

Thus, we have $\mathbf{y} \in \mathcal{F}(\lambda_1, \lambda_2)$, which completes the proof. ■

Appendix E. Proof of Theorem 14

Proof

(i) Suppose that $\bar{\lambda} < \lambda_{\max}^{\alpha}$. Theorem 9 implies that $\mathbf{y}/\bar{\lambda} \notin \mathcal{F}^{\alpha}$ and thus

$$\mathbf{y}/\bar{\lambda} - \mathbf{P}_{\mathcal{F}^\alpha}\left(\mathbf{y}/\bar{\lambda}\right) = \mathbf{y}/\bar{\lambda} - \theta^*(\bar{\lambda}, \alpha) \neq 0.$$

By the third part of Proposition 13, we can see that

$$\mathbf{y}/\bar{\lambda} - \theta^*(\bar{\lambda}, \alpha) \in N_{\mathcal{F}^\alpha}(\theta^*(\bar{\lambda}, \alpha)). \tag{86}$$

Thus, the statement holds for all $\bar{\lambda} < \lambda_{\max}^{\alpha}$.

Suppose that $\bar{\lambda} = \lambda_{\max}^{\alpha}$. By Theorem 9, we have

$$\theta^*(\bar{\lambda}, \alpha) = \mathbf{y}/\bar{\lambda} \in \mathcal{F}^{\alpha}.$$

In view of the definition of $\mathbf{X}_*$, we have

$$\left\|\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^{\alpha}}\right)\right\| = \alpha\sqrt{n_*},$$

where n_* is the number of feature contained in $\mathbf{X}_*$. Moreover, it is easy to see that

$$\|\mathcal{S}_1(\mathbf{X}_*^T\theta)\| \leq \alpha\sqrt{n_*},\, \forall\theta \in \mathcal{F}^{\alpha}.$$

Therefore, to prove the statement, we need to show that

$$\left\langle \mathbf{X}_*\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^{\alpha}}\right), \theta - \frac{\mathbf{y}}{\lambda_{\max}^{\alpha}}\right\rangle \leq 0,\, \forall\theta \in \mathcal{F}^{\alpha}. \tag{87}$$

Recall Remark 6, we have the following identity [see Eq. (20)]

$$\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right)=\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}-\mathbf{P}_{\mathcal{B}_\infty}\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right).$$

Thus, we have

$$\begin{aligned}
&\left\langle\mathbf{X}_*\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right),\theta-\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right\rangle \qquad (88)\\
=&\left\langle\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right),\mathbf{X}_*^T\left(\theta-\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right)+\mathbf{P}_{\mathcal{B}_\infty}\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right)-\mathbf{P}_{\mathcal{B}_\infty}\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right)\right\rangle\\
=&\left\langle\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right),\mathbf{X}_*^T\theta-\mathbf{P}_{\mathcal{B}_\infty}\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right)\right\rangle-\left\|\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right)\right\|^2\\
=&\left\langle\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right),\mathbf{X}_*^T\theta-\mathbf{P}_{\mathcal{B}_\infty}\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right)\right\rangle-\alpha^2 n_*.
\end{aligned}$$

Consider the first term on the right hand side of Eq. (88), we have

$$\begin{aligned}
&\left\langle\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right),\mathbf{X}_*^T\theta-\mathbf{P}_{\mathcal{B}_\infty}\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right)\right\rangle \qquad (89)\\
=&\left\langle\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right),\mathbf{X}_*^T\theta-\mathbf{P}_{\mathcal{B}_\infty}(\mathbf{X}_*^T\theta)+\mathbf{P}_{\mathcal{B}_\infty}(\mathbf{X}_*^T\theta)-\mathbf{P}_{\mathcal{B}_\infty}\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right)\right\rangle\\
=&\left\langle\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right),\mathcal{S}_1(\mathbf{X}_*^T\theta)\right\rangle+\left\langle\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right),\mathbf{P}_{\mathcal{B}_\infty}(\mathbf{X}_*^T\theta)-\mathbf{P}_{\mathcal{B}_\infty}\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right)\right\rangle.
\end{aligned}$$

Let $\mathcal{P}=\{i:[\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}]_i>1\}$ and $\mathcal{N}=\{i:[\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}]_i<-1\}$. We note that the second term on the right hand side of Eq. (89) can be written as

$$\begin{aligned}
&\left\langle\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right),\mathbf{P}_{\mathcal{B}_\infty}(\mathbf{X}_*^T\theta)-\mathbf{P}_{\mathcal{B}_\infty}\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right)\right\rangle \qquad (90)\\
=&\sum_{i\in\mathcal{P}}\left([\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}]_i-1\right)\left([\mathbf{P}_{\mathcal{B}_\infty}(\mathbf{X}_*^T\theta)]_i-1\right)+\sum_{j\in\mathcal{N}}\left([\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}]_j+1\right)\left([\mathbf{P}_{\mathcal{B}_\infty}(\mathbf{X}_*^T\theta)]_j+1\right).
\end{aligned}$$

Because $\|\mathbf{P}_{\mathcal{B}_\infty}(\mathbf{X}_*^T\theta)\|_\infty\leq 1$, we can see that Eq. (90) is non-positive. Therefore, by Eq. (89), we have

$$\begin{aligned}
\left\langle\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right),\mathbf{X}_*^T\theta-\mathbf{P}_{\mathcal{B}_\infty}\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right)\right\rangle\leq&\left\langle\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right),\mathcal{S}_1(\mathbf{X}_*^T\theta)\right\rangle \qquad (91)\\
\leq&\left\|\mathcal{S}_1\left(\mathbf{X}_*^T\frac{\mathbf{y}}{\lambda_{\max}^\alpha}\right)\right\|\|\mathcal{S}_1(\mathbf{X}_*^T\theta)\|\\
\leq&\alpha^2 n_*.
\end{aligned}$$

Combining Eq. (88) and the inequality in (91), we can see that the inequality in (87) holds. Thus, the statement holds for $\bar{\lambda}=\lambda_{\max}^\alpha$. This completes the proof.

(ii) We now show the second half. It is easy to see that the statement is equivalent to

$$\|\theta^*(\lambda,\alpha)-\theta^*(\bar{\lambda},\alpha)\|^2 \leq \langle \theta^*(\lambda,\alpha)-\theta^*(\bar{\lambda},\alpha),\, \mathbf{v}_\alpha^\perp(\lambda,\bar{\lambda})\rangle. \tag{92}$$

Thus, we will show that the inequality in (92) holds.

Because of the first half, we have

$$\langle \mathbf{n}_\alpha(\bar{\lambda}),\, \theta-\theta^*(\bar{\lambda},\alpha)\rangle \leq 0,\ \ \forall\, \theta\in\mathcal{F}^\alpha. \tag{93}$$

By letting $\theta=\theta^*(\lambda,\alpha)$, the inequality in (93) leads to

$$\langle \mathbf{n}_\alpha(\bar{\lambda}),\, \theta^*(\lambda,\alpha)-\theta^*(\bar{\lambda},\alpha)\rangle \leq 0. \tag{94}$$

In view of the first half and by letting $\theta=0$, the inequality in (93) leads to

$$\langle \mathbf{n}_\alpha(\bar{\lambda}),\, 0-\theta^*(\bar{\lambda},\alpha)\rangle \leq 0 \Rightarrow \begin{cases} \langle \mathbf{n}_\alpha(\bar{\lambda}),\, \mathbf{y}\rangle \geq 0, & \text{if } \bar{\lambda}=\lambda_{\max}^\alpha,\\ \|\mathbf{y}\|/\bar{\lambda}\geq \|\theta^*(\bar{\lambda},\alpha)\|, & \text{if } \bar{\lambda}<\lambda_{\max}^\alpha. \end{cases} \tag{95}$$

Moreover, the first half also leads to $\frac{\mathbf{y}}{\lambda}-\theta^*(\lambda,\alpha)\in N_{\mathcal{F}^\alpha}(\theta^*(\lambda,\alpha))$. Thus, we have

$$\langle \frac{\mathbf{y}}{\lambda}-\theta^*(\lambda,\alpha), \theta-\theta^*(\lambda,\alpha)\rangle \leq 0,\ \ \forall\, \theta\in\mathcal{F}^\alpha. \tag{96}$$

By letting $\theta=\theta^*(\bar{\lambda},\alpha)$, the inequality in (96) results in

$$\langle \frac{\mathbf{y}}{\lambda}-\theta^*(\lambda,\alpha), \theta^*(\bar{\lambda},\alpha)-\theta^*(\lambda,\alpha)\rangle \leq 0,\ \ \forall\, \theta\in\mathcal{F}^\alpha. \tag{97}$$

We can see that the inequality in (97) is equivalent to

$$\|\theta^*(\lambda,\alpha)-\theta^*(\bar{\lambda},\alpha)\|^2 \leq \langle \theta^*(\lambda,\alpha)-\theta^*(\bar{\lambda},\alpha),\, \mathbf{v}_\alpha(\lambda,\bar{\lambda})\rangle. \tag{98}$$

On the other hand, the right hand side of (92) can be rewritten as

$$\begin{aligned}
&\langle \theta^*(\lambda,\alpha)-\theta^*(\bar{\lambda},\alpha),\, \mathbf{v}_\alpha^\perp(\lambda,\bar{\lambda})\rangle \qquad (99)\\
=&\langle \theta^*(\lambda,\alpha)-\theta^*(\bar{\lambda},\alpha),\, \mathbf{v}_\alpha(\lambda,\bar{\lambda})\rangle - \langle \theta^*(\lambda,\alpha)-\theta^*(\bar{\lambda},\alpha),\, \mathbf{v}_\alpha(\lambda,\bar{\lambda})-\mathbf{v}_\alpha^\perp(\lambda,\bar{\lambda})\rangle\\
=&\langle \theta^*(\lambda,\alpha)-\theta^*(\bar{\lambda},\alpha),\, \mathbf{v}_\alpha(\lambda,\bar{\lambda})\rangle - \left\langle \theta^*(\lambda,\alpha)-\theta^*(\bar{\lambda},\alpha),\, \frac{\langle \mathbf{v}_\alpha(\lambda,\bar{\lambda}), \mathbf{n}_\alpha(\bar{\lambda})\rangle}{\|\mathbf{n}_\alpha(\bar{\lambda})\|^2}\mathbf{n}_\alpha(\bar{\lambda})\right\rangle.
\end{aligned}$$

In view of (94), (98) and (99), we can see that (92) holds if $\langle \mathbf{v}_\alpha(\lambda,\bar{\lambda}), \mathbf{n}_\alpha(\bar{\lambda})\rangle\geq 0$. Indeed,

$$\begin{aligned}
\langle \mathbf{v}_\alpha(\lambda,\bar{\lambda}), \mathbf{n}_\alpha(\bar{\lambda})\rangle &= \big\langle \mathbf{y}/\lambda-\theta^*(\bar{\lambda},\alpha), \mathbf{n}_\alpha(\bar{\lambda})\big\rangle \qquad (100)\\
&= \big(1/\lambda-1/\bar{\lambda}\big)\,\langle \mathbf{y}, \mathbf{n}_\alpha(\bar{\lambda})\rangle + \langle \mathbf{y}/\bar{\lambda}-\theta^*(\bar{\lambda},\alpha), \mathbf{n}_\alpha(\bar{\lambda})\rangle
\end{aligned}$$

Consider the first term on the right hand side of Eq. (100). By the first half of (95), we have

$$\langle \mathbf{y}, \mathbf{n}_\alpha(\bar{\lambda})\rangle \geq 0,\ \ \text{if } \bar{\lambda}=\lambda_{\max}^\alpha. \tag{101}$$

Suppose that $\bar{\lambda} < \lambda_{\max}^{\alpha}$. By the second half of (95), we can see that

$$\langle \mathbf{y}, \mathbf{n}_{\alpha}(\bar{\lambda}) \rangle = \langle \mathbf{y}, \mathbf{y}/\bar{\lambda} - \theta^*(\bar{\lambda}, \alpha) \rangle \geq 1/\bar{\lambda}\|\mathbf{y}\|^2 - \|\mathbf{y}\|\|\theta^*(\bar{\lambda}, \alpha)\| \geq 0. \tag{102}$$

Consider the second term on the right hand side of Eq. (100). It is easy to see that

$$\langle \mathbf{y}/\bar{\lambda} - \theta^*(\bar{\lambda}, \alpha), \mathbf{n}_{\alpha}(\bar{\lambda}) \rangle = \begin{cases} 0, & \text{if } \bar{\lambda} = \lambda_{\max}^{\alpha}, \\ \|\mathbf{n}_{\alpha}(\bar{\lambda})\|^2, & \text{if } \bar{\lambda} < \lambda_{\max}^{\alpha}. \end{cases} \tag{103}$$

Combining (101), (102) and Eq. (103), we have $\langle \mathbf{v}_{\alpha}(\lambda, \bar{\lambda}), \mathbf{n}_{\alpha}(\bar{\lambda}) \rangle \geq 0$, which completes the proof.

■

Appendix F. Proof of Theorem 18

Proof To simplify notations, let $\mathbf{o} = \mathbf{o}_{\alpha}(\lambda, \bar{\lambda})$, $r = \frac{1}{2}\|\mathbf{v}_{\alpha}^{\perp}(\lambda, \bar{\lambda})\|$ and $t_{g_k}^* = t_{g_k}^*(\lambda, \bar{\lambda}; \alpha)$. Therefore, the set Θ in Eq. (31) can be written as

$$\Theta = \{\mathbf{o} + \mathbf{v} : \|\mathbf{v}\| \leq r\}.$$

Then, problem (34) becomes

$$t_{g_k}^* = \sup_{\mathbf{v}} \{|\mathbf{x}_{g_k}^T(\mathbf{o} + \mathbf{v})| : \|\mathbf{v}\| \leq r\}.$$

We can see that

$$|\mathbf{x}_{g_k}^T(\mathbf{o} + \mathbf{v})| \leq |\mathbf{x}_{g_k}^T\mathbf{o}| + |\mathbf{x}_{g_k}^T\mathbf{v}| \leq |\mathbf{x}_{g_k}^T\mathbf{o}| + \|\mathbf{x}_{g_k}\|\|\mathbf{v}\| \leq |\mathbf{x}_{g_k}^T\mathbf{o}| + \|\mathbf{x}_{g_k}\|r.$$

Thus, we have

$$t_{g_k}^* \leq |\mathbf{x}_{g_k}^T\mathbf{o}| + \|\mathbf{x}_{g_k}\|r.$$

Consider $\mathbf{v}_1^* = r\mathbf{x}_{g_k}/\|\mathbf{x}_{g_k}\|$ and $\mathbf{v}_2^* = -r\mathbf{x}_{g_k}/\|\mathbf{x}_{g_k}\|$. It is easy to see that $\mathbf{o} + \mathbf{v}_1^* \in \Theta$ and $\mathbf{o} + \mathbf{v}_2^* \in \Theta$. Then,

$$|\mathbf{x}_{g_k}^T(\mathbf{o} + \mathbf{v}_i^*)| = |\mathbf{x}_{g_k}^T\mathbf{o}| + \|\mathbf{x}_{g_k}\|r, \quad \text{for } i = 1, 2,$$

which leads to

$$t_{g_k}^* = |\mathbf{x}_{g_k}^T\mathbf{o}| + \|\mathbf{x}_{g_k}\|r.$$

This completes the proof. ■

Appendix G. Efficiency of sgLeastR, nnLeastR, and APCG

We show that sgLeastR and nnLeastR are more appropriate solvers than APCG (Lin et al., 2014)—which is an accelerated coordinate descent method—in terms of the efficiency in solving SGL and nonnegative Lasso problems, respectively. Indeed, sgLeastR and nnLeastR are among the state-of-the-arts (Zhang et al., 2018a).

From the theoretical perspective, sgLeastR has a convergence rate of $\mathcal{O}(1/k^2)$ (k is the number of iterations) for the SGL problem, and so does nnLeastR for nonnegative Lasso (Liu et al., 2009). However, as pointed out by (Lin et al., 2014), without strong convexity, APCG recovers a special case of APPROX (Fercoq and Richtrik, 2013) and has a sublinear convergence rate of $\mathcal{O}(\frac{m}{m+k})$, where m is the number of groups. Notice that, sparse learning techniques usually deal with problems with $p \gg N$ (Tibshirani et al., 2015) (p is the number of features and N is the number of samples), in which strong convexity does not hold. Thus, sgLeastR and nnLeastR converge faster than APCG in solving SGL and nonnegative Lasso, respectively.

Synthetic 1 This data set is the same as that in Section 6.1.1, which consists of 1000 samples with 160000 features. The entries of the data matrix are i.i.d. standard Gaussian with pairwise correlation zero. For SGL, we randomly divide the features into 16000 groups.

E2006-tfidf (Chang and Lin, 2011) The E2006-tfidf data set consists of 3308 samples with 150360 features. The features include the volitility in twelve months and tf-idf of unigrams. For SGL, we randomly divide the 150360 features into 15036 groups.

Tables 5 and 6 show that sgLeastR and nnLeastR significantly outperform APCG in terms of the running time. Thus, we use sgLeastR and nnLeastR to solve SGL and nonnegative Lasso, respectively, in this paper.

Table 5: Running time (in seconds) for solving SGL along a sequence of 100 tuning parameter values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}^{\alpha}$ from 1.0 to 0.01 by (a): APCG (Lin et al., 2014); (b): sgLeastR Liu et al. (2009).

	α	tan(5°)	tan(15°)	tan(30°)	tan(45°)	tan(60°)	tan(75°)	tan(85°)
Synthetic 1	APCG	123088.60	126410.46	126301.35	126536.18	127348.69	127533.82	127419.43
	sgLeastR	15555.28	16124.08	16106.24	16293.04	16426.44	16836.16	16862.36
E2006-tfidf	APCG	116522.32	116546.41	115324.36	115982.47	119432.35	115433.98	115453.62
	sgLeastR	1303.25	1438.62	1458.32	1498.63	1506.35	1462.98	1408.38

Table 6: Running time (in seconds) for solving nonnegative Lasso along a sequence of 100 tuning parameter values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}$ from 1.0 to 0.01 by (a): APCG (Lin et al., 2014); (b): nnLeastR Liu et al. (2009).

	Synthetic 1	E2006-tfidf
APCG	92405 .15	106446.16
nnLeastR	13140.84	634.66

References

S. Armstrong, J. Staunton, L. Silverman, R. Pieters, M. den Boer, M. Minden, S. Sallan, E. Lander, T. Golub, and S. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30:41–47, 2002.

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

J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization, Second Edition*. Canadian Mathematical Society, 2006.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

D. Cai, X. He, and J. Han. Efficient kernel discriminant analysis via spectral regression. In *ICDM*, 2007.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. *Pacific Journal of Optimization*, 8:667–698, 2012.

O Fercoq and P Richtrik. Accelerated, parallel and proximal coordinate descent. *Journal on Optimization, 2013*, 25(4), 2013.

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. arXiv:1001.0736.

N. Halko, P. Martinsson, and J. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53:217–288, 2011.

J.-B. Hiriart-Urruty. From convex optimization to nonconvex optimization. necessary and sufficient conditions for global optimality. In *Nonsmooth optimization and related topics*. Springer, 1988.

J.-B. Hiriart-Urruty. A note on the Legendre-Fenchel transform of convex composite functions. In *Nonsmooth Mechanics and Analysis*. Springer, 2006.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998.

Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25(4):52–58, 2014.

J. Liu and J. Ye. Moreau-Yosida regularization for grouped tree structure learning. In *Advances in neural information processing systems 23*, 2010.

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

J. Liu, Z. Zhao, J. Wang, and J. Ye. Safe screening with variational inequalities and its application to lasso. In *International Conference on Machine Learning*, 2014.

Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Ng. Reading digits in nature images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2001.

K. Ogawa, Y. Suzuki, and I. Takeuchi. Safe screening of non-support vectors in pathwise SVM computation. In *International Conference on Machine Learning*, 2013.

K. Ogawa, Y. Suzuki, S. Suzumura, and I. Takeuchi. Safe sample screening for Support Vector Machine. *arXiv:1401.6740*, 2014.

J. Peng, J. Zhu, A. Bergamaschi, W. Han, D. Noh, J. Pollack, and P. Wang. Regularized multivariate regression for indentifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Appliced Statistics*, 4:53–77, 2010.

E. Petricoin, D. Ornstein, C. Paweletz, A. Ardekani, P. Hackett, B. Hitt, A. Velassco, C. Trucco, L. Wiegand, K. Wood, C. Simone, P. Levine, W. Linehan, M. Emmert-Buck, S. Steinberg, E. Kohn, and L. Liotta. Serum proteomic patterns for detection of prostate cancer. *Journal of National Cancer Institute*, 94:1576–1578, 2002.

A. Ruszczyński. *Nonlinear Optimization.* Princeton University Press, 2006.

S. Shevade and S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19:2246–2253, 2003.

T. Sim, B. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1615–1618, 2003.

N. Simon, J. Friedman., T. Hastie., and R. Tibshirani. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22:231–245, 2013.

P. Sprechmann, I. Ramírez, G. Sapiro., and Y. Eldar. C-HiLasso: a collaborative hierarchical sparse modeling framework. *IEEE Transactions on Signal Processing*, 59:4183–4198, 2011.

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

R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society Series B*, 74:245–266, 2012.

Robert Tibshirani, Martin Wainwright, and Trevor Hastie. *Statistical learning with sparsity: the Lasso and generalizations.* Chapman and Hall/CRC, 2015.

M. Vidyasagar. Machine learning methods in the cocomputation biology of cancer. In *Proceedings of the Royal Society A*, 2014.

M. Vincent and N. Hansen. Sparse group lasso and high dimensional multinomial classification. *Computational Statistics and Data Analysis*, 71:771–786, 2014.

J. Wang, J. Jun, and J. Ye. Efficient mixed-norm regularization: Algorithms and safe screening methods. *arXiv:1307.4156v1*.

J. Wang, J. Zhou, P. Wonka, and J. Ye. Lasso screening rules via dual polytope projection. In *Advances in neural information processing systems 26*, 2013.

J. Wang, P. Wonka, and J. Ye. Scaling svm and least absolute deviations via exact data reduction. In *International Conference on Machine Learning*, 2014.

M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. Olson, J. Marks, and J. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98:11462–11467, 2001.

Z. J. Xiang and P. J. Ramadge. Fast lasso screening tests based on correlations. In *IEEE ICASSP*, 2012.

D. Yogatama and N. Smith. Linguistic structured sparsity in text categorization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2014.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68:49–67, 2006.

Y.J. Zhang, N. Zhang, D.F. Sun, and K.C. Toh. An efficient hessian based algorithm for solving large-scale sparse group lasso problems. *Mathematical Programming*, 2018a.

Y.J. Zhang, N. Zhang, D.F. Sun, and K.C. Toh. An efficient hessian based algorithm for solving large-scale sparse group lasso problems. *Mathematical Programming*, 2018b.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67:301–320, 2005.