A Safe Screening Rule for Sparse Logistic Regression

Jie Wang[1], Jiayu Zhou[1], Jun Liu[2], Peter Wonka[1], and Jieping Ye[1]

[1]Computer Science and Engineering, Arizona State University, USA
[2]SAS Institute Inc., USA

July 22, 2013

Abstract

The ℓ_1-regularized logistic regression (or sparse logistic regression) is a widely used method for simultaneous classification and feature selection. Although many recent efforts have been devoted to its efficient implementation, its application to high dimensional data still poses significant challenges. In this paper, we present a fast and effective **s**parse **lo**gistic **re**gression **s**creening rule (Slores) to identify the "0" components in the solution vector, which may lead to a substantial reduction in the number of features to be entered to the optimization. An appealing feature of Slores is that the data set needs to be scanned only once to run the screening and its computational cost is negligible compared to that of solving the sparse logistic regression problem. Moreover, Slores is independent of solvers for sparse logistic regression, thus Slores can be integrated with any existing solver to improve the efficiency. We have evaluated Slores using high-dimensional data sets from different applications. Extensive experimental results demonstrate that Slores outperforms the existing state-of-the-art screening rules and the efficiency of solving sparse logistic regression is improved by one magnitude in general.

1 Introduction

Logistic regression (LR) is a popular and well established classification method that has been widely used in many domains such as machine learning [5, 8], text mining [4, 9], image processing [10, 17], bioinformatics [1, 15, 22, 30, 31], medical and social sciences [2, 19] etc. When the number of feature variables is large compared to the number of training samples, logistic regression is prone to over-fitting. To reduce over-fitting, regularization has been shown to be a promising approach. Typical examples include ℓ_2 and ℓ_1 regularization. Although ℓ_1 regularized LR is more challenging to solve compared to ℓ_2 regularized LR, it has received much attention in the last few years and the interest in it is growing [23, 27, 31] due to the increasing prevalence of high-dimensional data. The most appealing property of ℓ_1 regularized LR is the sparsity of the resulting models, which is equivalent to feature selection.

In the past few years, many algorithms have been proposed to efficiently solve the ℓ_1 regularized LR [6, 14, 13, 20]. However, for large-scale problems, solving the ℓ_1 regularized LR with higher accuracy remains challenging. One promising solution is by "screening", that is, we first identify the "*inactive*" features, which have 0 coefficients in the solution and then discard them from the optimization. This would result in a reduced feature matrix and substantial savings in computational cost and memory size. In [7], El Ghaoui *et al.* proposed novel screening rules, called "SAFE", to accelerate the optimization for a class of ℓ_1 regularized problems, including LASSO [25], ℓ_1 regularized LR and ℓ_1 regularized support vector machines. Inspired by SAFE, Tibshirani *et al.* [24] proposed "strong rules" for a large class of ℓ_1 regularized problems, including LASSO, elastic net, ℓ_1 regularized LR and more general convex problems. In [29, 28], Xiang et al. proposed "DOME" rules to further improve SAFE rules for LASSO based on the observation that SAFE rules can be understood as a special case of the general "sphere test". Although both strong rules and the sphere tests are more effective in discarding features than SAFE for solving

LASSO, it is worthwhile to mention that strong rules may mistakenly discard features that have non-zero coefficients in the solution and the sphere tests are not easy to be generalized to handle the ℓ_1 regularized LR. To the best of our knowledge, the SAFE rule is the only screening test for the ℓ_1 regularized LR that is "safe", that is, it only discards features that are guaranteed to be absent from the resulting models.

In this paper, we develop novel screening rules, called "Slores", for the ℓ_1 regularized LR. The proposed screening tests detect inactive features by estimating an upper bound of the inner product between each feature vector and the "dual optimal solution" of the ℓ_1 regularized LR, which is unknown. The more accurate the estimation is, the more inactive features can be detected. An accurate estimation of such an upper bound turns out to be quite challenging. Indeed most of the key ideas/insights behind existing "safe" screening rules for LASSO heavily rely on the least square loss, which are not applicable for the ℓ_1 regularized LR case due to the presence of the logistic loss. To this end, we propose a novel framework to accurately estimate an upper bound. Our key technical contribution is to formulate the estimation of an upper bound of the inner product as a constrained convex optimization problem and show that it admits a closed form solution. Therefore, the estimation of the inner product can be computed efficiently.

Figure 1: Comparison of Slores, strong rule and SAFE on the prostate cancer data set.

Our extensive experiments have shown that Slores discards far more features than SAFE yet requires much less computational efforts. In contrast with strong rules, Slores is "safe", i.e., it never discards features which have non-zero coefficients in the solution. To illustrate the effectiveness of Slores, we compare Slores, strong rule and SAFE on a data set of prostate cancer along a sequence of 86 parameters equally spaced on the λ/λ_{max} scale from 0.1 to 0.95, where λ is the parameter for the ℓ_1 penalty and λ_{max} is the smallest tuning parameter [12] such that the solution of the ℓ_1 regularized LR is 0 [please refer to Eq. (1)]. The data matrix contains 132 patients with 15154 features. To measure the performance of different screening rules, we compute the rejection ratio which is the ratio between the number of features discarded by screening rules and the number of features with 0 coefficients in the solution. Therefore, the larger the rejection ratio is, the more effective the screening rule is. The results are shown in Fig. 1. Clearly, Slores discards far more features than SAFE especially when λ/λ_{max} is large while the strong rule is not applicable when $\lambda/\lambda_{max} \leq 0.5$. We present more experimental results and discussions to demonstrate the effectiveness of Slores in Section 6.
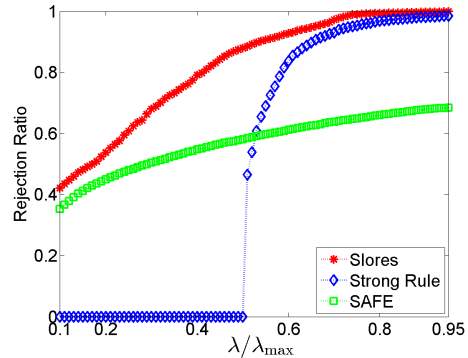
2 Basics and Motivations

In this section, we briefly review the basics of the ℓ_1 regularized LR and then motivate the general screening rules via the KKT conditions. Suppose we are given a set of training samples $\{\mathbf{x}_i\}_{i=1}^m$ and the associate labels $\mathbf{b} \in \Re^m$, where $\mathbf{x}_i \in \Re^p$ and $b_i \in \{1, -1\}$ for all $i \in \{1, \ldots, m\}$. The ℓ_1 regularized logistic regression is:

$$\min_{\beta, c} \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-\langle \beta, \bar{\mathbf{x}}_i \rangle - b_i c)) + \lambda \|\beta\|_1, \quad \text{(LRP}_\lambda\text{)}$$

where $\beta \in \Re^p$ and $c \in \Re$ are the model parameters to be estimated, $\bar{\mathbf{x}}_i = b_i \mathbf{x}_i$, and $\lambda > 0$ is the tuning parameter. Let the data matrix be $\overline{\mathbf{X}} \in \Re^{m \times p}$ with the ith row being $\bar{\mathbf{x}}_i$ and the jth column being $\bar{\mathbf{x}}^j$.

Let $\mathcal{C} = \{\theta \in \Re^m : \theta_i \in (0,1), i = 1, \ldots, m\}$ and $f(y) = y \log(y) + (1-y) \log(1-y)$ for $y \in (0,1)$. The dual problem of (LRP$_\lambda$) (please refer to the supplement) is given by

$$\min_\theta \left\{ g(\theta) = \frac{1}{m} \sum_{i=1}^m f(\theta_i) : \|\overline{\mathbf{X}}^T \theta\|_\infty \leq m\lambda, \langle \theta, \mathbf{b} \rangle = 0, \theta \in \mathcal{C} \right\}. \quad \text{(LRD}_\lambda\text{)}$$

2

To simplify notations, we denote the feasible set of problem (LRD_λ) as $\mathcal{F}_\lambda$, and let (β^*_λ, c^*) and θ^*_λ be the optimal solutions of problems (LRP_λ) and (LRD_λ) respectively. In [12], the authors have shown that for some special choice of the tuning parameter λ, both of (LRP_λ) and (LRD_λ) have closed form solutions. In fact, let $\mathcal{P} = \{i : b_i = 1\}$, $\mathcal{N} = \{i : b_i = -1\}$, and m^+ and m^- be the cardinalities of $\mathcal{P}$ and $\mathcal{N}$ respectively. We define

$$\lambda_{max} = \tfrac{1}{m}\|\bar{\mathbf{X}}^T \theta^*_{\lambda_{max}}\|_\infty, \tag{1}$$

where

$$[\theta^*_{\lambda_{max}}]_i = \begin{cases} \frac{m^-}{m}, & \text{if } i \in \mathcal{P}, \\ \frac{m^+}{m}, & \text{if } i \in \mathcal{N}, \end{cases} \quad i = 1, \dots, m. \tag{2}$$

($[\cdot]_i$ denotes the ith component of a vector.) Then, it is known [12] that $\beta^*_\lambda = 0$ and $\theta^*_\lambda = \theta^*_{\lambda_{max}}$ whenever $\lambda \geq \lambda_{max}$. When $\lambda \in (0, \lambda_{max}]$, it is known that (LRD_λ) has a unique optimal solution. (For completeness, we include the proof in the supplement.) We can now write the KKT conditions of problems (LRP_λ) and (LRD_λ) as

$$\langle \theta^*_\lambda, \bar{\mathbf{x}}^j \rangle \in \begin{cases} m\lambda, & \text{if } [\beta^*_\lambda]_j > 0, \\ -m\lambda, & \text{if } [\beta^*_\lambda]_j < 0, \quad j = 1, \dots, p. \\ [-m\lambda, m\lambda], & \text{if } [\beta^*_\lambda]_j = 0. \end{cases} \tag{3}$$

In view of Eq. (3), we can see that

$$|\langle \theta^*_\lambda, \bar{\mathbf{x}}^j \rangle| < m\lambda \Rightarrow [\beta^*_\lambda]_j = 0. \tag{R1}$$

In other words, if $|\langle \theta^*_\lambda, \bar{\mathbf{x}}^j \rangle| < m\lambda$, then the KKT conditions imply that the coefficient of $\bar{\mathbf{x}}^j$ in the solution β^*_λ is 0 and thus the jth feature can be safely removed from the optimization of (LRP_λ). However, for the general case in which $\lambda < \lambda_{max}$, (R1) is not applicable since it assumes the knowledge of θ^*_λ. Although it is unknown, we can still estimate a region $\mathcal{A}_\lambda$ which contains θ^*_λ. As a result, if $\max_{\theta \in \mathcal{A}} |\langle \theta, \bar{\mathbf{x}}^j \rangle| < m\lambda$, we can also conclude that $[\beta^*_\lambda]_j = 0$ by (R1). In other words, (R1) can be relaxed as

$$T(\theta^*_\lambda, \bar{\mathbf{x}}^j) := \max_{\theta \in \mathcal{A}_\lambda} |\langle \theta, \bar{\mathbf{x}}^j \rangle| < m\lambda \Rightarrow [\beta^*_\lambda]_j = 0. \tag{R1$'$}$$

In this paper, (R1$'$) serves as the foundation for constructing our screening rules, Slores. From (R1$'$), it is easy to see that screening rules with smaller $T(\theta^*_\lambda, \bar{\mathbf{x}}^j)$ are more aggressive in discarding features. To give a tight estimation of $T(\theta^*_\lambda, \bar{\mathbf{x}}^j)$, we need to restrict the region $\mathcal{A}_\lambda$ which includes θ^*_λ as small as possible. In Section 3, we show that the estimation of the upper bound $T(\theta^*_\lambda, \bar{\mathbf{x}}^j)$ can be obtained via solving a convex optimization problem. We show in Section 4 that the convex optimization problem admits a closed form solution and derive Slores in Section 5 based on (R1$'$).

3   Estimating the Upper Bound via Solving a Convex Optimization Problem

In this section, we present a novel framework to estimate an upper bound $T(\theta^*_\lambda, \bar{\mathbf{x}}^j)$ of $|\langle \theta^*_\lambda, \bar{\mathbf{x}}^j \rangle|$. In the subsequent development, we assume a parameter λ_0 and the corresponding dual optimal $\theta^*_{\lambda_0}$ are given. In our Slores rule to be presented in Section 5, we set λ_0 and $\theta^*_{\lambda_0}$ to be λ_{max} and $\theta^*_{\lambda_{max}}$ given in Eqs. (1) and (2). We formulate the estimation of $T(\theta^*_\lambda, \bar{\mathbf{x}}^j)$ as a constrained convex optimization problem in this section, which will be shown to admit a closed form solution in Section 4.

For the dual function $g(\theta)$, it follows that $[\nabla g(\theta)]_i = \frac{1}{m}\log(\frac{\theta_i}{1-\theta_i})$, $[\nabla^2 g(\theta)]_{i,i} = \frac{1}{m}\frac{1}{\theta_i(1-\theta_i)} \geq \frac{4}{m}$. Since $\nabla^2 g(\theta)$ is a diagonal matrix, it follows that $\nabla^2 g(\theta) \succeq \frac{4}{m}I$, where I is the identity matrix. Thus, $g(\theta)$ is strongly convex with modulus $\mu = \frac{4}{m}$ [18]. Rigorously, we have the following lemma.

3

Lemma 1. *Let $\lambda > 0$ and $\theta_1, \theta_2 \in \mathcal{F}_\lambda$, then*

$$a). \qquad g(\theta_2) - g(\theta_1) \geq \langle \nabla g(\theta_1), \theta_2 - \theta_1 \rangle + \frac{2}{m} \|\theta_2 - \theta_1\|_2^2. \tag{4}$$

b). If $\theta_1 \neq \theta_2$, the inequality in (4) becomes a strict inequality, i.e., "$\geq$" becomes "$>$".

Given $\lambda \in (0, \lambda_0]$, it is easy to see that both of θ_λ^* and $\theta_{\lambda_0}^*$ belong to $\mathcal{F}_{\lambda_0}$. Therefore, Lemma 1 can be a useful tool to bound θ_λ^* with the knowledge of $\theta_{\lambda_0}^*$. In fact, we have the following theorem.

Theorem 2. *Let $\lambda_{max} \geq \lambda_0 > \lambda > 0$, then the following holds:*

$$a). \qquad \|\theta_\lambda^* - \theta_{\lambda_0}^*\|_2^2 \leq \frac{m}{2} \left[g\left(\frac{\lambda}{\lambda_0}\theta_{\lambda_0}^*\right) - g(\theta_{\lambda_0}^*) + \left(1 - \frac{\lambda}{\lambda_0}\right) \langle \nabla g(\theta_{\lambda_0}^*), \theta_{\lambda_0}^* \rangle \right] \tag{5}$$

b). If $\theta_\lambda^* \neq \theta_{\lambda_0}^*$, the inequality in (5) becomes a strict inequality, i.e., "$\leq$" becomes "$<$".

Proof. a). It is easy to see that $\mathcal{F}_\lambda \subseteq \mathcal{F}_{\lambda_0}$, $\theta_\lambda^* \in \mathcal{F}_\lambda$ and $\theta_{\lambda_0}^* \in \mathcal{F}_{\lambda_0}$. Therefore, both of $\theta_{\lambda_0}^*$ and θ_λ^* belong to the set $\mathcal{F}_{\lambda_0}$. By Lemma 1, we have

$$\|\theta_\lambda^* - \theta_{\lambda_0}^*\|_2^2 \leq \frac{m}{2} \left[g(\theta_\lambda^*) - g(\theta_{\lambda_0}^*) + \langle \nabla g(\theta_{\lambda_0}^*), \theta_{\lambda_0}^* - \theta_\lambda^* \rangle \right]. \tag{6}$$

Let $\theta_\lambda = \frac{\lambda}{\lambda_0}\theta_{\lambda_0}^*$. It is easy to see that

$$\theta_\lambda \in \mathcal{C}, \ \|\bar{\mathbf{X}}^T \theta_\lambda\|_\infty = \frac{\lambda}{\lambda_0}\|\bar{\mathbf{X}}^T \theta_{\lambda_0}^*\|_\infty \leq m\lambda, \ \langle \theta_\lambda, \mathbf{b} \rangle = \frac{\lambda}{\lambda_0}\langle \theta_{\lambda_0}^*, \mathbf{b} \rangle = 0.$$

Therefore, we can see that $\theta_\lambda \in \mathcal{F}_\lambda$ and thus

$$g(\theta_\lambda^*) = \min_{\theta \in \mathcal{F}_\lambda} g(\theta) \leq g(\theta_\lambda) = g\left(\frac{\lambda}{\lambda_0}\theta_{\lambda_0}^*\right).$$

Then the inequality in (6) becomes

$$\|\theta_\lambda^* - \theta_{\lambda_0}^*\|_2^2 \leq \frac{m}{2} \left[g\left(\frac{\lambda}{\lambda_0}\theta_{\lambda_0}^*\right) - g(\theta_{\lambda_0}^*) + \langle \nabla g(\theta_{\lambda_0}^*), \theta_{\lambda_0}^* - \theta_\lambda^* \rangle \right]. \tag{7}$$

On the other hand, by noting that (LRD_λ) is feasible, we can see that the Slater's conditions holds and thus the KKT conditions [21] lead to:

$$0 \in \nabla g(\theta_\lambda^*) + \sum_{j=1}^{p} \eta_j^+ \bar{\mathbf{x}}^j + \sum_{j'=1}^{p} \eta_i^- (-\bar{\mathbf{x}}^i) + \gamma \mathbf{b} + N_\mathcal{C}(\theta_\lambda^*), \tag{8}$$

where $\eta^+, \eta^- \in \Re_+^p$, $\gamma \in \Re$ and $N_\mathcal{C}(\theta_\lambda^*)$ is the normal cone of $\mathcal{C}$ at θ_λ^* [21]. Because $\theta_\lambda^* \in \mathcal{C}$ and $\mathcal{C}$ is an open set, θ_λ^* is an interior point of $\mathcal{C}$ and thus $N_\mathcal{C}(\theta_\lambda^*) = \emptyset$ [21]. Therefore, Eq. (8) becomes:

$$\nabla g(\theta_\lambda^*) + \sum_{j=1}^{p} \eta_j^+ \bar{\mathbf{x}}^j + \sum_{j'=1}^{p} \eta_{j'}^- (-\bar{\mathbf{x}}^{j'}) + \gamma \mathbf{b} = 0. \tag{9}$$

Let $\mathcal{I}_{\lambda_0}^+ = \{j : \langle \theta_{\lambda_0}^*, \bar{\mathbf{x}}^j \rangle = m\lambda_0, j = 1, \ldots, p\}$, $\mathcal{I}_{\lambda_0}^- = \{j' : \langle \theta_{\lambda_0}^*, \bar{\mathbf{x}}^{j'} \rangle = -m\lambda_0, j = 1, \ldots, p\}$ and $\mathcal{I}_{\lambda_0} = \mathcal{I}_{\lambda_0}^+ \cup \mathcal{I}_{\lambda_0}^-$. We can see that $\mathcal{I}_{\lambda_0}^+ \cap \mathcal{I}_{\lambda_0}^- = \emptyset$. By the complementary slackness condition, if $k \notin \mathcal{I}_{\lambda_0}$, we have $\eta_k^+ = \eta_k^- = 0$. Therefore,

$$\langle \nabla g(\theta_{\lambda_0}^*), \theta_{\lambda_0}^* \rangle + \sum_{j \in \mathcal{I}_{\lambda_0}^+} \eta_j^+ \langle \theta_{\lambda_0}^*, \bar{\mathbf{x}}^j \rangle + \sum_{j' \in \mathcal{I}_{\lambda_0}^-} \eta_{j'}^- \langle \theta_{\lambda_0}^*, -\bar{\mathbf{x}}^{j'} \rangle + \gamma \langle \theta_{\lambda_0}^*, \mathbf{b} \rangle = 0$$

$$\Leftrightarrow -\frac{1}{m\lambda_0}\langle \nabla g(\theta_{\lambda_0}^*), \theta_{\lambda_0}^* \rangle = \sum_{j \in \mathcal{I}_{\lambda_0}^+} \eta_j^+ + \sum_{j' \in \mathcal{I}_{\lambda_0}^-} \eta_{j'}^-$$

4

Similarly, we have

$$\langle \nabla g(\theta^*_{\lambda_0}), \theta^*_\lambda \rangle + \sum_{j \in \mathcal{I}^+_{\lambda_0}} \eta^+_j \langle \theta^*_\lambda, \bar{\mathbf{x}}^j \rangle + \sum_{j' \in \mathcal{I}^-_{\lambda_0}} \eta^-_{j'} \langle \theta^*_\lambda, -\bar{\mathbf{x}}^{j'} \rangle + \gamma \langle \theta^*_\lambda, \mathbf{b} \rangle = 0$$

$$\Leftrightarrow - \langle \nabla g(\theta^*_{\lambda_0}), \theta^*_\lambda \rangle = \sum_{j \in \mathcal{I}^+_{\lambda_0}} \eta^+_j \langle \theta^*_\lambda, \bar{\mathbf{x}}^j \rangle + \sum_{j' \in \mathcal{I}^-_{\lambda_0}} \eta^-_{j'} \langle \theta^*_\lambda, -\bar{\mathbf{x}}^{j'} \rangle$$

$$\leq \sum_{j \in \mathcal{I}^+_{\lambda_0}} \eta^+_j |\langle \theta^*_\lambda, \bar{\mathbf{x}}^j \rangle| + \sum_{j' \in \mathcal{I}^-_{\lambda_0}} \eta^-_{j'} |\langle \theta^*_\lambda, -\bar{\mathbf{x}}^{j'} \rangle|$$

$$\leq m\lambda \left\{ \sum_{j \in \mathcal{I}^+_{\lambda_0}} \eta^+_j + \sum_{j' \in \mathcal{I}^-_{\lambda_0}} \eta^-_{j'} \right\}$$

$$= -\tfrac{\lambda}{\lambda_0} \langle g(\theta^*_{\lambda_0}), \theta^*_{\lambda_0} \rangle$$

Recall (7), the inequality in (5) follows.

  b). The proof is the same as part a) by noting part b) of Lemma 1. □

Theorem 2 implies that θ^*_λ is inside a ball centred at $\theta^*_{\lambda_0}$ with radius

$$r = \sqrt{\tfrac{m}{2} \left[g\left(\tfrac{\lambda}{\lambda_0} \theta^*_{\lambda_0} \right) - g(\theta^*_{\lambda_0}) + (1 - \tfrac{\lambda}{\lambda_0}) \langle \nabla g(\theta^*_{\lambda_0}), \theta^*_{\lambda_0} \rangle \right]}. \tag{10}$$

Recall that to make our screening rules more aggressive in discarding features, we need to get a tight upper bound $T(\theta^*_\lambda, \bar{\mathbf{x}}^j)$ of $|\langle \theta^*_\lambda, \bar{\mathbf{x}}^j \rangle|$ [please see (R1′)]. Thus, it is desirable to further restrict the possible region $\mathcal{A}_\lambda$ of θ^*_λ. Clearly, we can see that

$$\langle \theta^*_\lambda, \mathbf{b} \rangle = 0 \tag{11}$$

since θ^*_λ is feasible for problem (LRD$_\lambda$). On the other hand, we call the set $\mathcal{I}_{\lambda_0}$ defined in the proof of Theorem 2 the "*active set*" of $\theta^*_{\lambda_0}$. In fact, we have the following lemma for the active set.

Lemma 3. *Given the optimal solution θ^*_λ of problem (LRD$_\lambda$), the active set $\mathcal{I}_\lambda = \{j : |\langle \theta^*_\lambda, \bar{\mathbf{x}}^j \rangle| = m\lambda, j = 1, \ldots, p\}$ is not empty if $\lambda \in (0, \lambda_{max}]$.*

Since $\lambda_0 \in (0, \lambda_{max}]$, we can see that $\mathcal{I}_{\lambda_0}$ is not empty by Lemma 3. We pick $j_0 \in \mathcal{I}_{\lambda_0}$ and set

$$\bar{\mathbf{x}}^* = \operatorname{sign}(\langle \theta^*_{\lambda_0}, \bar{\mathbf{x}}^{j_0} \rangle) \bar{\mathbf{x}}^{j_0}. \tag{12}$$

It follows that $\langle \bar{\mathbf{x}}^*, \theta^*_{\lambda_0} \rangle = m\lambda_0$. Due to the feasibility of θ^*_λ for problem (LRD$_\lambda$), θ^*_λ satisfies

$$\langle \theta^*_\lambda, \bar{\mathbf{x}}^* \rangle \leq m\lambda. \tag{13}$$

As a result, Theorem 2, Eq. (11) and (13) imply that θ^*_λ is contained in the following set:

$$\mathcal{A}^\lambda_{\lambda_0} := \{\theta : \|\theta - \theta^*_{\lambda_0}\|^2_2 \leq r^2, \langle \theta, \mathbf{b} \rangle = 0, \langle \theta, \bar{\mathbf{x}}^* \rangle \leq m\lambda\}.$$

Since $\theta^*_\lambda \in \mathcal{A}^\lambda_{\lambda_0}$, we can see that $|\langle \theta^*_\lambda, \bar{\mathbf{x}}^j \rangle| \leq \max_{\theta \in \mathcal{A}^\lambda_{\lambda_0}} |\langle \theta, \bar{\mathbf{x}}^j \rangle|$. Therefore, (R1′) implies that if

$$T(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}) := \max_{\theta \in \mathcal{A}^\lambda_{\lambda_0}} |\langle \theta, \bar{\mathbf{x}}^j \rangle| \tag{UBP}$$

is smaller than $m\lambda$, we can conclude that $[\beta^*_\lambda]_j = 0$ and $\bar{\mathbf{x}}^j$ can be discarded from the optimization of (LRP$_\lambda$). Notice that, we replace the notations $\mathcal{A}_\lambda$ and $T(\theta^*_\lambda, \bar{\mathbf{x}}^j)$ with $T(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0})$ and $\mathcal{A}^\lambda_{\lambda_0}$ to emphasize their dependence on $\theta^*_{\lambda_0}$. Clearly, as long as we can solve for $T(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0})$, (R1′) would be an applicable screening rule to discard features which have 0 coefficients in β^*_λ. We give a closed form solution of problem (42) in the next section.

5

4 Solving the Convex Optimization Problem (UBP)

In this section, we show how to solve the convex optimization problem (42) based on the standard Lagrangian multiplier method. We first transform problem (42) into a pair of convex minimization problem (UBP′) via Eq. (15) and then show that the strong duality holds for (UBP′) in Lemma 6. The strong duality guarantees the applicability of the Lagrangian multiplier method. We then give the closed form solution of (UBP′) in Theorem 8. After we solve problem (UBP′), it is straightforward to compute the solution of problem (42) via Eq. (15).

Before we solve (42) for the general case, it is worthwhile to mention a special case in which $\mathbf{P}\bar{\mathbf{x}}^j = \bar{\mathbf{x}}^j - \frac{\langle \bar{\mathbf{x}}^j, \mathbf{b} \rangle}{\|\mathbf{b}\|_2^2}\mathbf{b} = 0$. Clearly, $\mathbf{P}$ is the projection operator which projects a vector onto the orthogonal complement of the space spanned by $\mathbf{b}$. In fact, we have the following theorem.

Theorem 4. *Let $\lambda_{max} \geq \lambda_0 > \lambda > 0$, and assume $\theta^*_{\lambda_0}$ is known. For $j \in \{1, \ldots, p\}$, if $\mathbf{P}\bar{\mathbf{x}}^j = 0$, then $T(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}) = 0$.*

Because of (R1′), we immediately have the following corollary.

Corollary 5. *Let $\lambda \in (0, \lambda_{max})$ and $j \in \{1, \ldots, p\}$. If $\mathbf{P}\bar{\mathbf{x}}^j = 0$, then $[\beta^*_\lambda]_j = 0$.*

For the general case in which $\mathbf{P}\bar{\mathbf{x}}^j \neq 0$, let

$$T_+(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}) := \max_{\theta \in \mathcal{A}^\lambda_{\lambda_0}} \langle \theta, +\bar{\mathbf{x}}^j \rangle, \ T_-(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}) := \max_{\theta \in \mathcal{A}^\lambda_{\lambda_0}} \langle \theta, -\bar{\mathbf{x}}^j \rangle. \tag{14}$$

Clearly, we have

$$T(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}) = \max\{T_+(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}), T_-(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0})\}. \tag{15}$$

Therefore, we can solve problem (42) by solving the two sub-problems in (14).

Let $\xi \in \{+1, -1\}$. Then problems in (14) can be written uniformly as

$$T_\xi(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}) = \max_{\theta \in \mathcal{A}^\lambda_{\lambda_0}} \langle \theta, \xi\bar{\mathbf{x}}^j \rangle. \tag{UBP$_s$}$$

To make use of the standard Lagrangian multiplier method, we transform problem (UBP$_s$) to the following minimization problem:

$$-T_\xi(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}) = \min_{\theta \in \mathcal{A}^\lambda_{\lambda_0}} \langle \theta, -\xi\bar{\mathbf{x}}^j \rangle \tag{UBP′}$$

by noting that $\max_{\theta \in \mathcal{A}^\lambda_{\lambda_0}} \langle \theta, \xi\bar{\mathbf{x}}^j \rangle = -\min_{\theta \in \mathcal{A}^\lambda_{\lambda_0}} \langle \theta, -\xi\bar{\mathbf{x}}^j \rangle$.

Lemma 6. *Let $\lambda_{max} \geq \lambda_0 > \lambda > 0$ and assume $\theta^*_{\lambda_0}$ is known. The strong duality holds for problem (UBP′). Moreover, problem (UBP′) admits an optimal solution in $\mathcal{A}^\lambda_{\lambda_0}$.*

Because the strong duality holds for problem (UBP′) by Lemma 6, the Lagrangian multiplier method is applicable for (UBP′). In general, we need to first solve the dual problem and then recover the optimal solution of the primal problem via KKT conditions. Recall that r and $\bar{\mathbf{x}}^*$ are defined by Eq. (10) and (12) respectively. Lemma 7 derives the dual problems of (UBP′) for different cases.

Lemma 7. *Let $\lambda_{max} \geq \lambda_0 > \lambda > 0$ and assume $\theta^*_{\lambda_0}$ is known. For $j \in \{1, \ldots, p\}$ and $\mathbf{P}\bar{\mathbf{x}}^j \neq 0$, let $\bar{\mathbf{x}} = -\xi\bar{\mathbf{x}}^j$. Denote*

$$\mathcal{U}_1 = \{(u_1, u_2) : u_1 > 0, u_2 \geq 0\} \ and \ \mathcal{U}_2 = \left\{(u_1, u_2) : u_1 = 0, u_2 = -\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2}\right\}.$$

a). If $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} \in (-1, 1]$, the dual problem of (UBP′) is equivalent to:

$$\max_{(u_1, u_2) \in \mathcal{U}_1} \bar{g}(u_1, u_2) = -\frac{1}{2u_1}\|\mathbf{P}\bar{\mathbf{x}} + u_2 \mathbf{P}\bar{\mathbf{x}}^*\|_2^2 + u_2 m(\lambda_0 - \lambda) + \langle \theta_{\lambda_0}^*, \bar{\mathbf{x}} \rangle - \frac{1}{2}u_1 r^2. \qquad \text{(UBD′)}$$

Moreover, $\bar{g}(u_1, u_2)$ attains its maximum in $\mathcal{U}_1$.

b). If $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} = -1$, the dual problem of (UBP′) is equivalent to:

$$\max_{(u_1, u_2) \in \mathcal{U}_1 \cup \mathcal{U}_2} \bar{\bar{g}}(u_1, u_2) = \begin{cases} \bar{g}(u_1, u_2), & \text{if } (u_1, u_2) \in \mathcal{U}_1, \\ -\frac{\|\mathbf{P}\bar{\mathbf{x}}\|_2}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2} m\lambda, & \text{if } (u_1, u_2) \in \mathcal{U}_2. \end{cases} \qquad \text{(UBD″)}$$

We can now solve problem (UBP′) in the following theorem.

Theorem 8. *Let $\lambda_{max} \geq \lambda_0 > \lambda > 0$, $d = \frac{m(\lambda_0 - \lambda)}{r\|\mathbf{P}\bar{x}^*\|_2}$ and assume $\theta_{\lambda_0}^*$ is known. For $j \in \{1, \dots, p\}$ and $\mathbf{P}\bar{\mathbf{x}}^j \neq 0$, let $\bar{\mathbf{x}} = -\xi \bar{\mathbf{x}}^j$.*

a). *If $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} \geq d$, then*

$$T_\xi(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = r\|\mathbf{P}\bar{\mathbf{x}}\|_2 - \langle \theta_{\lambda_0}^*, \bar{\mathbf{x}} \rangle; \qquad (16)$$

b). *If $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} < d$, then*

$$T_\xi(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = r\|\mathbf{P}\bar{\mathbf{x}} + u_2^* \mathbf{P}\bar{\mathbf{x}}^*\|_2 - u_2^* m(\lambda_0 - \lambda) - \langle \theta_{\lambda_0}^*, \bar{\mathbf{x}} \rangle, \qquad (17)$$

where

$$u_2^* = \frac{-a_1 + \sqrt{\Delta}}{2a_2}, \qquad (18)$$

$$a_2 = \|\mathbf{P}\bar{\mathbf{x}}^*\|_2^4 (1 - d^2),$$
$$a_1 = 2\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle \|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2 (1 - d^2),$$
$$a_0 = \langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle^2 - d^2 \|\mathbf{P}\bar{\mathbf{x}}\|_2^2 \|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2,$$
$$\Delta = a_1^2 - 4a_2 a_0 = 4d^2(1 - d^2)\|\mathbf{P}\bar{\mathbf{x}}^*\|_2^4 (\|\mathbf{P}\bar{\mathbf{x}}\|_2^2 \|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2 - \langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle^2).$$

Notice that, although the dual problems of (UBP′) in Lemma 7 are different, the resulting upper bound $T_\xi(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*)$ can be given by Theorem 8 in a uniform way. The tricky part is how to deal with the extremal cases in which $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} \in \{-1, +1\}$. To avoid the lengthy discussion of Theorem 8, we omit the proof in the main text and include the details in the supplement.

5 The proposed Slores Rule for ℓ_1 Regularized Logistic Regression

Using (R1′), we are now ready to construct the screening rules for the ℓ_1 Regularized Logistic Regression. By Corollary 5, we can see that the orthogonality between the jth feature and the response vector $\mathbf{b}$ implies the absence of $\bar{\mathbf{x}}^j$ from the resulting model. For the general case in which $\mathbf{P}\bar{\mathbf{x}}^j \neq 0$, (R1′) implies that if $T(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = \max\{T_+(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*), T_-(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*)\} < m\lambda$, then the jth feature can be discarded from the optimization of (LRP$_\lambda$). Notice that, letting $\xi = \pm 1$, $T_+(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*)$ and $T_-(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*)$ have been solved by Theorem 8. Rigorously, we have the following theorem.

Theorem 9 (Slores). *Let $\lambda_0 > \lambda > 0$ and assume $\theta_{\lambda_0}^*$ is known.*

1. *If $\lambda \geq \lambda_{max}$, then $\beta_\lambda^* = 0$;*

2. *If $\lambda_{max} \geq \lambda_0 > \lambda > 0$ and either of the following holds:*

(a) $\mathbf{P}\bar{\mathbf{x}}^j = 0$,

(b) $\max\{T_\xi(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) : \xi = \pm 1\} < m\lambda$,

then $[\beta_\lambda^*]_j = 0$.

Based on Theorem 9, we construct the Slores rule as summarized below in Algorithm 1.

Notice that, the output $\mathcal{R}$ of Slores is the indices of the features that need to be entered to the optimization. As a result, suppose the output of Algorithm 1 is $\mathcal{R} = \{j_1, \ldots, j_k\}$, we can substitute the full matrix $\overline{\mathbf{X}}$ in problem (LRP_λ) with the submatrix $\overline{\mathbf{X}}_{\mathcal{R}} = (\bar{\mathbf{x}}^{j_1}, \ldots, \bar{\mathbf{x}}^{j_k})$ and just solve for $[\beta_\lambda^*]_{\mathcal{R}}$ and c_λ^*.

On the other hand, Algorithm 1 implies that Slores needs five inputs. Since $\overline{\mathbf{X}}$ and $\mathbf{b}$ come with the data and λ is chosen by the user, we only need to specify $\theta_{\lambda_0}^*$ and λ_0. In other words, we need to provide Slores with an dual optimal solution of problem (LRD_λ) for an arbitrary parameter. A natural choice is by setting $\lambda_0 = \lambda_{max}$ and $\theta_{\lambda_0}^* = \theta_{\lambda_{max}}^*$ given in Eq. (1) and Eq. (2).

Algorithm 1 $\mathcal{R} = \mathrm{Slores}(\overline{\mathbf{X}}, \mathbf{b}, \lambda, \lambda_0, \theta_{\lambda_0}^*)$

Initialize $\mathcal{R} := \{1, \ldots, p\}$;
if $\lambda \geq \lambda_{max}$ **then**
  set $\mathcal{R} = \emptyset$;
else
  for $j = 1$ **to** p **do**
    if $\mathbf{P}\bar{\mathbf{x}}^j = 0$ **then**
      remove j from $\mathcal{R}$;
    else if $\max\{T_\xi(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) : \xi = \pm 1\} < m\lambda$ **then**
      remove j from $\mathcal{R}$;
    end if
  end for
end if
Return: $\mathcal{R}$

---

6 Experiments

We evaluate our screening rules using the newgroup data set [12] and Yahoo web pages data sets [26]. The newgroup data set is cultured from the data by Koh et al. [12]. The Yahoo data sets include 11 top-level categories, each of which is further divided into a set of subcategories. In our experiment we construct five balanced binary classification datasets from the topics of Computers, Education, Health, Recreation, and Science. For each topic, we choose samples from one subcategory as the positive class and randomly sample an equal number of samples from the rest of subcategories as the negative class. The statistics of the data sets are given in Table 1.

We compare the performance of Slores and the strong rule which achieves state-of-the-art performance for ℓ_1 regularized LR. We do not include SAFE because it is less effective in discarding features than strong rules and requires much higher computational time [24]. Fig. 1 has shown the performance of Slores, strong rule and SAFE. We compare the efficiency of the three screening rules on the same prostate cancer data set in Table 2. All of the screening rules are tested along a sequence of 86 parameter values equally spaced on the λ/λ_{max} scale from 0.1 to 0.95. We repeat the procedure 100 times and during each time we undersample 80% of the data. We report the total running time of the three screening rules over the 86 values of λ/λ_{max} in Table 2. For reference, we also report the total

Table 1: Statistics of the test data sets.

Data set	m	p	no. nonzeros
newsgroup	11269	61188	1467345
Computers	216	25259	23181
Education	254	20782	28287
Health	228	18430	40145
Recreation	370	25095	49986
Science	222	24002	37227

Table 2: Running time (in seconds) of Slores, strong rule, SAFE and the solver.

Slores	Strong Rule	SAFE	Solver
0.37	0.33	1128.65	10.56

running time of the solver[1]. We observe that the running time of Slores and strong rule is negligible compared to that of the solver. However, SAFE takes much longer time even than the solver.

---

[1] In this paper, the ground truth is computed by SLEP [16].

In Section 6.1, we evaluate the performance of Slores and strong rule. Recall that we use the rejection ratio, i.e., the ratio between the number of features discarded by the screening rules and the number of features with 0 coefficients in the solution, to measure the performance of screening rules. Note that, because no features with non-zero coefficients in the solution would be mistakenly discarded by Slores, its rejection ratio is no larger than one. We then compare the efficiency of Slores and strong rule in Section 6.2.

The experiment settings are as follows. For each data set, we undersample 80% of the date and run Slores and strong rules along a sequence of 86 parameter values equally spaced on the λ/λ_{max} scale from 0.1 to 0.95. We repeat the procedure 100 times and report the average performance and running time at each of the 86 values of λ/λ_{max}. Slores, strong rules and SAFE are all implemented in Matlab. All of the experiments are carried out on a Intel(R) (i7-2600) 3.4Ghz processor.

6.1 Comparison of Performance

In this experiment, we evaluate the performance of the Slores and the strong rule via the rejection ratio. Fig. 2 shows the rejection ratio of Slores and strong rule on six real data sets. When $\lambda/\lambda_{max} > 0.5$, we can see that both Slores and strong rule are able to identify almost 100% of the inactive features, i.e., features with 0 coefficients in the solution vector. However, when $\lambda/\lambda_{max} \leq 0.5$, strong rule can not detect the inactive features. In contrast, we observe that Slores exhibits much stronger capability in discarding inactive features for small λ, even when λ/λ_{max} is close to 0.1. Taking the data point at which $\lambda/\lambda_{max} = 0.1$ for example, Slores discards about 99% inactive features for the newsgroup data set. For the other data sets, more than 80% inactive features are identified by Slores. Therefore, in terms of rejection ratio, Slores significantly outperforms the strong rule. It is also worthwhile to mention that the discarded features by Slores are guaranteed to have 0 coefficients in the solution. But strong rule may mistakenly discard features which have non-zero coefficients in the solution.

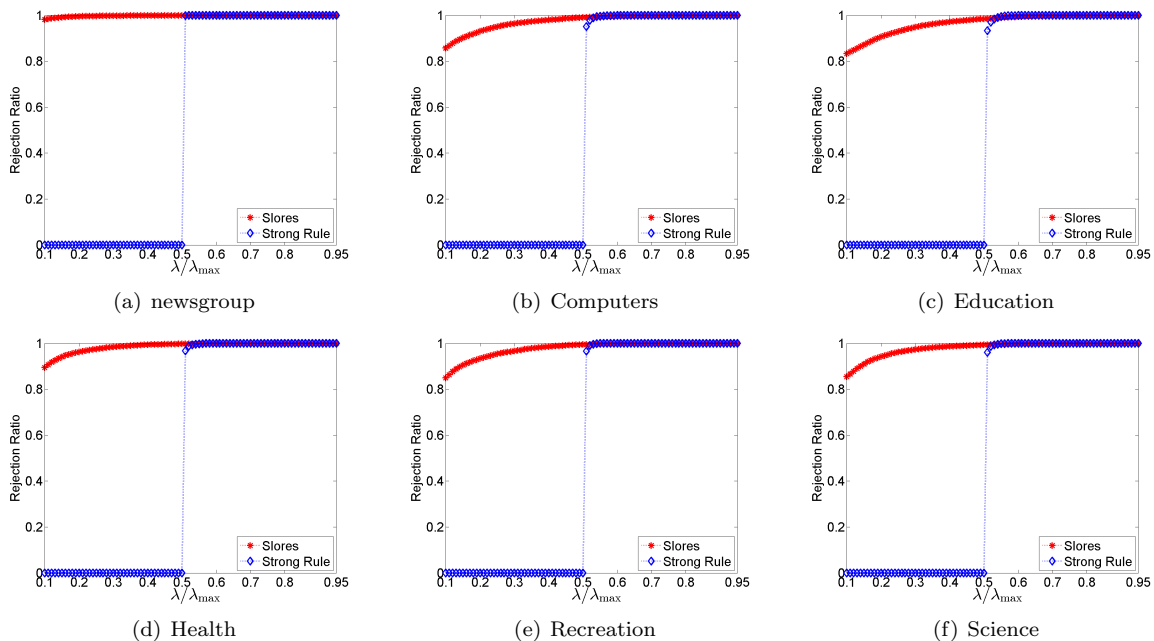| (a) newsgroup | (b) Computers | (c) Education |

| (d) Health | (e) Recreation | (f) Science |

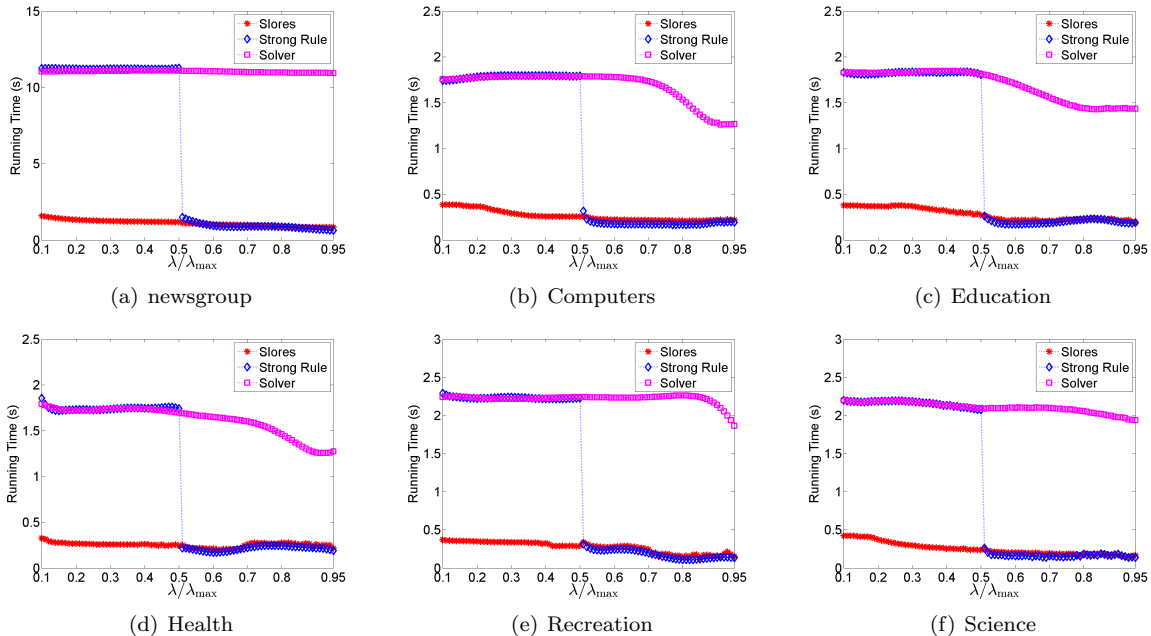Figure 2: Comparison of the performance of Slores and strong rules on six real data sets.

Figure 3: Comparison of the efficiency of Slores and strong rule on six real data sets.

6.2 Comparison of Efficiency

We compare efficiency of Slores and the strong rule in this experiment. The data sets for evaluating the rules are the same as Section 6.1. The running time of the screening rules reported in Fig. 3 includes the computational cost of the rules themselves and that of the solver after screening. We plot the running time of the screening rules against that of the solver without screening. As indicated by Fig. 2, when $\lambda/\lambda_{max} > 0.5$, Slores and strong rule discards almost 100% of the inactive features. As a result, the size of the feature matrix involved in the optimization of problem (LRP_λ) is greatly reduced. From Fig. 3, we can observe that the efficiency is improved by about one magnitude on average compared to that of the solver without screening. However, when $\lambda/\lambda_{max} < 0.5$, strong rule can not identify any inactive features and thus the running time is almost the same as that of the solver without screening. In contrast, Slores is still able to identify more than 80% of the inactive features for the data sets cultured from the Yahoo web pages data sets and thus the efficiency is improved by roughly 5 times. For the newgroup data set, about 99% inactive features are identified by Slores which leads to about 10 times savings in running time. These results demonstrate the power of the proposed Slores rule in improving the efficiency of solving the ℓ_1 regularized LR.

7 Conclusions

In this paper, we propose novel screening rules to effectively discard features for ℓ_1 regularized LR. Extensive numerical experiments on real data demonstrate that Slores outperforms the existing state-of-the-art screening rules. We plan to extend the framework of Slores to more general sparse formulations, including convex ones, like group Lasso, fused Lasso, ℓ_1 regularized SVM, and non-convex ones, like ℓ_p regularized problems where $0 < p < 1$.

References

[1] M. Asgary, S. Jahandideh, P. Abdolmaleki, and A. Kazemnejad. Analysis and identification of β-turn types using multinomial logistic regression and artificial neural network. *Bioinformatics*, 23(23):3125–3130, 2007. URL `http://dblp.uni-trier.de/db/journals/bioinformatics/bioinformatics23.html#AsgaryJAK07`.

[2] C. Boyd, M. Tolson, and W. Copes. Evaluating trauma care: The TRISS method, trauma score and the injury severity score. *Journal of Trauma*, 27:370–378, 1987.

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

[4] Jack R. Brzezinski and George J. Knafl. Logistic regression modeling for context-based classification. In *DEXA Workshop*, pages 755–759, 1999. URL `http://dblp.uni-trier.de/db/conf/dexaw/dexaw99.html#BrzezinskiK99`.

[5] K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *NIPS*, 2008.

[6] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32:407–499, 2004.

[7] L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. arXiv:1009.4219v2.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 38(2), 2000.

[9] A. Genkin, D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49:291–304(14), 2007. doi: doi:10.1198/004017007000000245. URL `http://www.ingentaconnect.com/content/asa/tech/2007/00000049/00000003/art00007`.

[10] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, 80(3):300–316, 2008. URL `http://dblp.uni-trier.de/db/journals/ijcv/ijcv80.html#GouldRCEK08`.

[11] O. Güler. *Foundations of Optimization*. Springer, 2010.

[12] K. Koh, S. J. Kim, and S. Boyd. An interior-point method for large scale l1-regularized logistic regression. *J. Mach. Learn. Res.*, 8:1519–1555, 2007.

[13] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:957–968, 2005.

[14] S. Lee, H. Lee, P. Abbeel, and A. Ng. Efficient l1 regularized logistic regression. In *In AAAI-06*, 2006.

[15] J. Liao and K. Chin. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics*, 23(15):1945–1951, 2007. URL `http://dblp.uni-trier.de/db/journals/bioinformatics/bioinformatics23.html#LiaoC07`.

[16] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009. URL `http://www.public.asu.edu/~jye02/Software/SLEP`.

[17] S. Martins, L. Sousa, and J. Martins. Additive logistic regression applied to retina modelling. In *ICIP (3)*, pages 309–312. IEEE, 2007. URL `http://dblp.uni-trier.de/db/conf/icip/icip2007-3.html#MartinsSM07`.

[18] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

[19] S. Palei and S. Das. Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: An approach. *Safety Science*, 47:88–96, 2009.

[20] M. Park and T. Hastie. ℓ_1 regularized path algorithm for generalized linear models. *J. R. Statist. Soc. B*, 69:659–677, 2007.

[21] A. Ruszczyński. *Nonlinear Optimization*. Princeton university Press, 2006.

[22] M. Sartor, G. Leikauf, and M. Medvedovic. LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, 25(2):211–217, 2009. URL `http://dblp.uni-trier.de/db/journals/bioinformatics/bioinformatics25.html#SartorLM09`.

[23] D. Sun, T. Erp, P. Thompson, C. Bearden, M. Daley, L. Kushan, M. Hardt, K. Nuechterlein, A. Toga, and T. Cannon. Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: classification analysis using probabilistic brain atlas and machine learning algorithms. *Biological Psychiatry*, 66:1055–1–60, 2009.

[24] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *J. R. Statist. Soc. B*, 74:245–266, 2012.

[25] Robert Tibshirani. Regression shringkage and selection via the lasso. *J. R. Statist. Soc. B*, 58:267–288, 1996.

[26] Naonori Ueda and Kazumi Saito. Parametric mixture models for multi-labeled text. *Advances in neural information processing systems*, 15:721–728, 2002.

[27] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25:714–721, 2009.

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

[29] Z. J. Xiang, H. Xu, and P. J. Ramadge. Learning sparse representation of high dimensional data on large scale dictionaries. In *NIPS*, 2011.

[30] J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *NIPS*, pages 1081–1088. MIT Press, 2001. URL `http://dblp.uni-trier.de/db/conf/nips/nips2001.html#ZhuH01`.

[31] J. Zhu and T. Hastie. Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5:427–443, 2004.

Appendix

In this appendix, we will provide detailed proofs of the theorems, lemmas and corollaries in the main text.

A Deviation of the Dual Problem of the Sparse Logistic Regression

Suppose we are given a set of training samples $\{\mathbf{x}_i\}_{i=1}^m$ and the associate labels $\mathbf{b} \in \Re^m$, where $\mathbf{x}_i \in \Re^p$ and $b_i \in \{1, -1\}$ for all $i \in \{1, \ldots, m\}$. The logistic regression problem takes the form as follows:

$$\min_{\beta, c} \ \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-\langle \beta, \bar{\mathbf{x}}_i \rangle - b_i c)) + \lambda \|\beta\|_1,$$

where $\beta \in \Re^p$ and $c \in \Re$ are the model parameters to be estimated, $\bar{\mathbf{x}}_i = b_i \mathbf{x}_i$ and $\lambda > 0$. Let $\bar{\mathbf{X}}$ denote the data matrix whose rows consist of $\bar{\mathbf{x}}_i$. Denote the columns of $\bar{\mathbf{X}}$ as $\bar{\mathbf{x}}^j$, $j \in \{1, \ldots, p\}$.

A.1 Dual Formulation

By introducing the slack variables $q_i = \frac{1}{m}(-\langle \beta, \bar{x}_i \rangle - b_i c)$ for all $i \in \{1, \ldots, m\}$, problem (LRP$_\lambda$) can be formulated as:

$$\min_{\beta, c} \ \frac{1}{m} \sum_{i=1}^m \log\left[1 + \exp(m q_i)\right] + \lambda \|\beta\|_1, \tag{19}$$

$$\text{s.t. } q_i = \frac{1}{m}(-\langle \beta, \bar{\mathbf{x}}_i \rangle - b_i c), \ i \in \{1, \ldots, m\}.$$

The Lagrangian is

$$L(\mathbf{q}, \beta, c; \theta) = \frac{1}{m} \sum_{i=1}^m \log\left[1 + \exp(m q_i)\right] + \lambda \|\beta\|_1 + \sum_{i=1}^m \theta_i \left[\frac{1}{m}(-\langle \beta, \bar{\mathbf{x}}_i \rangle - b_i c) - q_i\right] \tag{20}$$

$$= \frac{1}{m} \sum_{i=1}^m \log\left[1 + \exp(m q_i)\right] - \langle \theta, \mathbf{q} \rangle + \lambda \|\beta\|_1 - \frac{1}{m}\langle \sum_{i=1}^m \theta_i \bar{\mathbf{x}}_i, \beta \rangle - \frac{c}{m}\langle \theta, \mathbf{b} \rangle$$

In order to find the dual function, we need to solve the following subproblems:

$$\min_{\mathbf{q}} f_1(\mathbf{q}) = \frac{1}{m} \sum_{i=1}^m \log\left[1 + \exp(m q_i)\right] - \langle \theta, \mathbf{q} \rangle, \tag{21}$$

$$\min_{\beta} f_2(\beta) = \lambda \|\beta\|_1 - \frac{1}{m}\langle \sum_{i=1}^m \theta_i \bar{\mathbf{x}}_i, \beta \rangle, \tag{22}$$

$$\min_{c} f_3(c) = -\frac{c}{m}\langle \theta, \mathbf{b} \rangle. \tag{23}$$

Consider $f_1(\mathbf{q})$. It is easy to see

$$[\nabla f_1(\mathbf{q})]_i = \frac{\exp(m q_i)}{1 + \exp(m q_i)} - \theta_i.$$

By setting $[\nabla f_1(\mathbf{q})]_i = 0$, we get

$$\exp(m q_i') = \frac{\theta_i}{1 - \theta_i}, \ q_i' = \frac{1}{m} \log\left(\frac{\theta_i}{1 - \theta_i}\right).$$

Clearly, we can see that $\theta_i \in (0,1)$ for all $i \in \{1, \ldots, m\}$. Therefore,

$$\min_{\mathbf{q}} f_1(\mathbf{q}) = f_1(\mathbf{q}') = -\frac{1}{m} \sum_{i=1}^{m} \left[\theta_i \log(\theta_i) + (1 - \theta_i) \log(1 - \theta_i)\right].$$

Consider $f_2(\beta)$ and let $\beta' = \operatorname{argmin}_\beta f_2(\beta)$. The optimality condition is

$$0 \in \lambda \mathbf{v} - \frac{1}{m} \theta_i \bar{\mathbf{x}}_i, \ \|\mathbf{v}\|_\infty \le 1, \ \langle \mathbf{v}, \beta' \rangle = \|\beta'\|_1.$$

It is easy to see that

$$\langle \frac{1}{m} \sum_{i=1}^{m} \theta_i \bar{\mathbf{x}}_i, \beta' \rangle = \|\beta'\|_1,$$

and thus

$$\langle \theta, \bar{\mathbf{x}}^j \rangle \in \begin{cases} m\lambda, & \text{if } \beta'_j > 0, \\ -m\lambda, & \text{if } \beta'_j < 0, \\ [-m\lambda, m\lambda], & \text{if } \beta'_j = 0. \end{cases}$$

Moreover, it follows that

$$\min_\beta f_2(\beta) = f_2(\beta') = 0.$$

For $f_3(c)$, we can see that

$$f_3'(c) = -\frac{1}{m} \langle \theta, \mathbf{b} \rangle.$$

Therefore, we have

$$\langle \theta, \mathbf{b} \rangle = 0,$$

since otherwise $\inf_c f_3(c) = -\infty$ and the dual problem is infeasible. Clearly, $\min_c f_3(c) = 0$.

All together, the dual problem is

$$\min_\theta g(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[\theta_i \log(\theta_i) + (1 - \theta_i) \log(1 - \theta_i)\right], \tag{LRD$_\lambda$}$$
$$\text{s.t. } \|\bar{\mathbf{X}}^T \theta\|_\infty \le m\lambda,$$
$$\langle \theta, \mathbf{b} \rangle = 0,$$
$$\theta \in \mathcal{C},$$

where $\mathcal{C} = \{\theta \in \Re^m : \theta_i \in (0,1), \ i = 1, \ldots, m\}$.

B Proof of the Existence of the Optimal Solution of (LRD$_\lambda$)

In this section, we prove that problem (LRD$_\lambda$) has a unique optimal solution for all $\lambda > 0$.

Therefore, in Lemma A, we first show that problem (LRD$_\lambda$) is feasible for all $\lambda > 0$. Then Lemma B confirms the existence of the dual optimal solution θ_λ^*.

Lemma 10. *A 1. For $\lambda > 0$, problem (LRD$_\lambda$) is feasible, i.e., $\mathcal{F}_\lambda \ne \emptyset$.*
2. The Slater's condition holds for problem (LRD$_\lambda$) in which $\lambda > 0$.

Proof. 1. When $\lambda \geq \lambda_{max}$, the feasibility of (LRD_λ) is trivial because of the existence of $\theta^*_{\lambda_{max}}$. We focus on the case in which $\lambda \in (0, \lambda_{max}]$ below.

Recall that [12] λ_{max} is the smallest tuning parameter such that $\beta^*_\lambda = 0$ and $\theta^*_\lambda = \theta^*_{\lambda_{max}}$ whenever $\lambda \geq \lambda_{max}$. For convenience, we rewrite the definition of λ_{max} and $\theta^*_{\lambda_{max}}$ as follows.

$$\lambda_{max} = \max_{j \in \{1,\ldots,p\}} \frac{1}{m} |\langle \theta^*_{\lambda_{max}}, \bar{\mathbf{x}}^j \rangle|,$$

$$[\theta^*_{\lambda_{max}}]_i = \begin{cases} \frac{m^-}{m}, & \text{if } i \in \mathcal{P}, \\ \frac{m^+}{m}, & \text{if } i \in \mathcal{N}, \end{cases} \quad i = 1,\ldots,m.$$

Clearly, we have $\theta^*_{\lambda_{max}} \in \mathcal{F}_{\lambda_{max}}$, i.e., $\theta^*_{\lambda_{max}} \in \mathcal{C}$, $\|\bar{\mathbf{X}}^T \theta^*_{\lambda_{max}}\|_\infty \leq m\lambda_{max}$ and $\langle \theta^*_{\lambda_{max}}, \mathbf{b} \rangle = 0$.

Let us define:

$$\theta_\lambda = \frac{\lambda}{\lambda_{max}} \theta^*_{\lambda_{max}}.$$

Since $0 < \frac{\lambda}{\lambda_{max}} \leq 1$, we can see that $\theta_\lambda \in \mathcal{C}$. Moreover, it is easy to see that

$$\|\bar{\mathbf{X}}^T \theta_\lambda\|_\infty = \frac{\lambda}{\lambda_{max}} \|\bar{\mathbf{X}}^T \theta^*_{\lambda_{max}}\|_\infty \leq m\lambda,$$

and

$$\langle \theta_\lambda, \mathbf{b} \rangle = \frac{\lambda}{\lambda_{max}} \langle \theta^*_{\lambda_{max}}, \mathbf{b} \rangle = 0.$$

Therefore, $\theta_\lambda \in \mathcal{F}_\lambda$, i.e., $\mathcal{F}_\lambda$ is not empty and thus (LRD_λ) is feasible.

2. The constraints of (LRD_λ) are all affine. Therefore, the Slater's condition reduces to the feasibility of (LRD_λ) [3]. When $\lambda \geq \lambda_{max}$, $\theta^*_{\lambda_{max}}$ is clearly a feasible solution of (LRD_λ). When $\lambda \in (0, \lambda_{max}]$, we have shown the feasibility of (LRD_λ) in part 1. Therefore, the Slater's condition always holds for (LRD_λ) in which $\lambda > 0$.

$\square$

Lemma 11. *B Given $\lambda \in (0, \lambda_{max}]$, problem (LRD_λ) has a unique optimal solution, i.e., there exists a unique $\theta^*_\lambda \in \mathcal{F}_\lambda$ such that $g(\theta)$ achieves its minimum over $\mathcal{F}_\lambda$ at θ^*_λ.*

Proof. Let $\widetilde{\mathcal{C}} := \{\theta \in \Re^m : \theta_i \in [0,1], i = 1,\ldots,m\}$ and

$$\widetilde{\mathcal{F}}_\lambda := \{\theta \in \Re^m : \theta \in \widetilde{\mathcal{C}}, \|\bar{\mathbf{X}}^T \theta\|_\infty \leq m\lambda, \langle \theta, \mathbf{b} \rangle = 0\}.$$

Clearly, $\mathcal{F}_\lambda \subseteq \widetilde{\mathcal{F}}_\lambda$ and $\widetilde{\mathcal{F}}_\lambda$ is compact.

We show that $g(\theta)$ can be continuously extended to $\widetilde{F}_\lambda$. For $y \in (0,1)$, define

$$f(y) = y \log(y) + (1-y) \log(1-y).$$

By L'Hôpital's rule, it is easy to see that $\lim_{y\downarrow 0} f(y) = \lim_{y\uparrow 1} f(y) = 0$. Therefore, let

$$\widetilde{f}(y) = \begin{cases} f(y), & y \in (0,1), \\ 0, & y \in \{0,1\}, \end{cases}$$

and

$$\widetilde{g}(\theta) = \frac{1}{m} \sum_{i=1}^m \widetilde{f}(\theta_i).$$

15

Clearly, when $\theta \in \mathcal{F}_\lambda$, we have $\widetilde{g}(\theta) = g(\theta)$, i.e., $g(\theta)$ is the restriction of $\widetilde{g}(\theta)$ over $\mathcal{F}_\lambda$.

On the other hand, the definition of $\widetilde{g}(\theta)$ implies that $\widetilde{g}(\theta)$ is continuous over $\widetilde{\mathcal{F}}_\lambda$. Together with the fact that $\widetilde{\mathcal{F}}_\lambda$ is compact, we can see that there exists $\theta_\lambda^* \in \widetilde{\mathcal{F}}_\lambda$ and $\widetilde{g}(\theta_\lambda^*) = \min_{\theta \in \widetilde{\mathcal{F}}_\lambda} \widetilde{g}(\theta)$.

Consider the optimization problem

$$\min_{\theta \in \widetilde{\mathcal{F}}_\lambda} \widetilde{g}(\theta) \tag{LRD$_\lambda'$}$$

Because of Lemma A, we know that $\mathcal{F}_\lambda \neq \emptyset$ and thus $\widetilde{\mathcal{F}}_\lambda \neq \emptyset$ either. Therefore, problem (LRD$_\lambda'$) is feasible. By noting that the constraints of problem (LRD$_\lambda'$) are all linear, the Slater's condition is satisfied. Hence, there exists a set of Lagrangian multipliers $\eta^+, \eta^- \in \Re_+^p$, $\xi^+, \xi^- \in \Re_+^m$ and $\gamma \in \Re$ such that

$$\nabla \widetilde{g}(\theta_\lambda^*) + \sum_{i=1}^m \eta_i^+ \bar{\mathbf{x}}^i + \sum_{i=1}^m \eta_i^- (-\bar{\mathbf{x}}^i) + \xi^+ - \xi^- + \gamma \mathbf{b} = 0. \tag{24}$$

We can see that if there is an i_0 such that $[\theta_\lambda^*]_{i_0} \in \{0, 1\}$, i.e., $\theta_\lambda^* \notin \mathcal{F}_\lambda$, Eq. (24) does not hold since $|[\nabla \widetilde{g}(\theta_\lambda^*)]_{i_0}| = \infty$. Therefore, we can conclude that $\theta_\lambda^* \in \mathcal{F}_\lambda$. Moreover, it is easy to see that $\theta_\lambda^* = \operatorname{argmin}_{\theta \in \widetilde{\mathcal{F}}_\lambda} \widetilde{g}(\theta) = \operatorname{argmin}_{\theta \in \mathcal{F}_\lambda} g(\theta)$, i.e. θ_λ^* is a minimum of $g(\theta)$ over $\mathcal{F}_\lambda$. The uniqueness of θ_λ^* is due to the strict convexity of $g(\theta)$ (strong convexity implies strict convexity), which completes the proof. $\square$

C   Proof of Lemma 1

Proof. Recall that the domain of g is $\mathcal{C} = \{\theta \in \Re^m : [\theta]_i \in (0, 1), i = 1, \ldots, m\}$.

a. It is easy to see that

$$[\nabla g(\theta)]_i = \frac{1}{m} \log(\frac{[\theta]_i}{1 - [\theta]_i}), \ [\nabla^2 g(\theta)]_{i,i} = \frac{1}{m} \frac{1}{[\theta]_i(1 - [\theta]_i)} \geq \frac{4}{m}, \ [\nabla^2 g(\theta)]_{i,j} = 0, \ i \neq j.$$

Clearly, $\nabla^2 g(\theta)$ is a diagonal matrix and

$$\nabla^2 g(\theta) \geq \frac{4}{m} I. \tag{25}$$

Therefore, g is a strong convex function with convexity parameter $\mu = \frac{4}{m}$ [18]. The claim then follows directly from the definition of strong convex functions.

b. If $\theta_1 \neq \theta_2$, then there exists at least one $i' \in \{1, \ldots, m\}$ such that $[\theta_1]_{i'} \neq [\theta_2]_{i'}$. Moreover, at most one of $[\theta_1]_{i'}$ and $[\theta_2]_{i'}$ can be $\frac{1}{2}$. Without loss of generality, assume $[\theta_1]_{i'} \neq \frac{1}{2}$.

For $t \in (0, 1)$, let $\theta(t) = t\theta_2 + (1 - t)\theta_1$. Since $[\theta_1]_{i'} \neq \frac{1}{2}$, we can find $t' \in (0, 1)$ such that $[\theta(t')]_{i'} \neq \frac{1}{2}$. Therefore, we can see that

$$\langle \nabla^2 g(\theta(t'))(\theta_2 - \theta_1), \theta_2 - \theta_1 \rangle = \sum_{i=1}^m [\nabla^2 g(\theta(t'))]_{i,i}([\theta_2]_i - [\theta_1]_i)^2 \tag{26}$$

$$= \sum_{\substack{i \in \{1,\ldots,m\}, \\ i \neq i'}} [\nabla^2 g(\theta(t'))]_{i,i}([\theta_2]_i - [\theta_1]_i)^2$$

$$+ [\nabla^2 g(\theta(t'))]_{i',i'}([\theta_2]_{i'} - [\theta_1]_{i'})^2.$$

Clearly, when $i \in \{1, \ldots, m\} \setminus i'$, we have $[\nabla^2 g(\theta(t'))]_{i,i} \geq \frac{4}{m}$ and thus

$$\sum_{\substack{i \in \{1,\ldots,m\}, \\ i \neq i'}} [\nabla^2 g(\theta(t'))]_{i,i}([\theta_2]_i - [\theta_1]_i)^2 \geq \sum_{\substack{i \in \{1,\ldots,m\}, \\ i \neq i'}} \frac{4}{m}([\theta_2]_i - [\theta_1]_i)^2. \tag{27}$$

16

Since $[\theta(t')]_{i'} \neq \frac{1}{2}$, we have $[\nabla^2 g(\theta(t'))]_{i,i} > \frac{4}{m}$ which results in

$$[\nabla^2 g(\theta(t'))]_{i',i'}([\theta_2]_{i'} - [\theta_1]_{i'})^2 > \frac{4}{m}([\theta_2]_{i'} - [\theta_1]_{i'})^2. \tag{28}$$

Moreover, because $[\theta_1]_{i'} \neq [\theta_2]_{i'}$, we can see that

$$([\theta_2]_{i'} - [\theta_1]_{i'})^2 \neq 0. \tag{29}$$

Therefore, Eq. (26) and the inequalities in (27), (28) and (29) imply that

$$\langle \nabla^2 g(\theta(t'))(\theta_2 - \theta_1), \theta_2 - \theta_1 \rangle > \frac{4}{m}\|\theta_2 - \theta_1\|_2^2. \tag{30}$$

Due to the continuity of $\nabla^2 g(\theta(t))$, there exist a small number $\epsilon > 0$ such that for all $t \in (t' - \frac{1}{2}\epsilon, t' + \frac{1}{2}\epsilon)$, we have

$$\langle \nabla^2 g(\theta(t))(\theta_2 - \theta_1), \theta_2 - \theta_1 \rangle > \frac{4}{m}\|\theta_2 - \theta_1\|_2^2, \tag{31}$$

It follows that

$$g(\theta_2) - g(\theta_1) = \int_0^1 \nabla g(\theta(t))(\theta_2 - \theta_1)dt \tag{32}$$

$$= \int_0^1 \left[\nabla g(\theta(0)) + \int_0^t \nabla^2 g(\theta(\tau))(\theta_2 - \theta_1)d\tau \right] (\theta_2 - \theta_1)dt$$

$$= \langle \nabla g(\theta_1), \theta_2 - \theta_1 \rangle + \int_0^1 \int_0^t \langle \nabla^2 g(\theta(\tau))(\theta_2 - \theta_1), \theta_2 - \theta_1 \rangle d\tau dt.$$

Let

$$\Delta_{t_1}^{t_2} := \int_{t_1}^{t_2} \int_0^t \langle \nabla^2 g(\theta(\tau))(\theta_2 - \theta_1), \theta_2 - \theta_1 \rangle d\tau dt.$$

Clearly, the second term on the right hand side of Eq. (32) is Δ_0^1, which can be written as

$$\Delta_0^1 = \Delta_0^{t' - \frac{1}{2}\epsilon} + \Delta_{t' - \frac{1}{2}\epsilon}^{t' + \frac{1}{2}\epsilon} + \Delta_{t' + \frac{1}{2}\epsilon}^1.$$

Due to the inequality in (31), we can see that

$$\Delta_{t' - \frac{1}{2}\epsilon}^{t' + \frac{1}{2}\epsilon} = \int_{t' - \frac{1}{2}\epsilon}^{t' + \frac{1}{2}\epsilon} \int_0^t \langle \nabla^2 g(\theta(\tau))(\theta_2 - \theta_1), \theta_2 - \theta_1 \rangle d\tau dt \tag{33}$$

$$> \int_{t' - \frac{1}{2}\epsilon}^{t' + \frac{1}{2}\epsilon} \frac{4}{m}\|\theta_2 - \theta_1\|_2^2 t dt$$

$$= \frac{4}{m}\|\theta_2 - \theta_1\|_2^2 \cdot (t'\epsilon).$$

Due to the inequality in (25), it is easy to see that

$$\langle \nabla^2 g(\theta(t))(\theta_2 - \theta_1), \theta_2 - \theta_1 \rangle \geq \frac{4}{m}\|\theta_2 - \theta_1\|_2^2 \tag{34}$$

for all $t \in (0, 1)$. Therefore,

$$\Delta_0^{t' - \frac{1}{2}\epsilon} = \int_{t' - \frac{1}{2}\epsilon}^{t' + \frac{1}{2}\epsilon} \int_0^t \langle \nabla^2 g(\theta(\tau))(\theta_2 - \theta_1), \theta_2 - \theta_1 \rangle d\tau dt \tag{35}$$

$$\geq \int_0^{t' - \frac{1}{2}\epsilon} \frac{4}{m}\|\theta_2 - \theta_1\|_2^2 t dt$$

$$= \frac{4}{m}\|\theta_2 - \theta_1\|_2^2 \cdot \left[\frac{1}{2}\left(t' - \frac{1}{2}\epsilon \right)^2 \right].$$

17

Similarly,

$$\Delta^1_{t'+\frac{1}{2}\epsilon} = \int^1_{t'+\frac{1}{2}\epsilon} \int^t_0 \langle \nabla^2 g(\theta(\tau))(\theta_2-\theta_1),\theta_2-\theta_1\rangle d\tau dt \tag{36}$$

$$\geq \int^1_{t'+\frac{1}{2}\epsilon} \frac{4}{m}\|\theta_2-\theta_1\|^2_2 t dt$$

$$= \frac{4}{m}\|\theta_2-\theta_1\|^2_2 \cdot \frac{1}{2}\left[1-\left(t'+\frac{1}{2}\epsilon\right)^2\right].$$

The inequalites in (33), (35) and (36) imply that

$$\Delta^1_0 > \frac{2}{m}\|\theta_2-\theta_1\|^2_2. \tag{37}$$

Therefore, the inequality in (37) and Eq. (32) lead to the following strict inequality

$$g(\theta_2) - g(\theta_1) > \langle \nabla g(\theta_1), \theta_2-\theta_1\rangle + \frac{2}{m}\|\theta_2-\theta_1\|^2_2, \tag{38}$$

which completes the proof.

$\square$

D   Proof of Lemma 3

Proof. According to the definition of λ_{max} in Eq. (1), there must be $j_0 \in \{1,\ldots,p\}$ such that $\lambda_{max} = \frac{1}{m}|\langle \theta^*_{\lambda_{max}}, \bar{\mathbf{x}}^{j_0}\rangle|$. Clearly, $j_0 \in \mathcal{I}_{\lambda_{max}}$ and thus $\mathcal{I}_{\lambda_{max}}$ is not empty.

For $\lambda \in (0, \lambda_{max})$, we prove the statement by contradiction. Suppose $\mathcal{I}_\lambda$ is empty, then the KKT condition for (LRD_λ) at θ^*_λ can be written as:

$$0 \in \nabla g(\theta^*_\lambda) + \sum^p_{j=1}\eta^+_j\bar{\mathbf{x}}^j + \sum^p_{j'=1}\eta^-_i(-\bar{\mathbf{x}}^i) + \gamma\mathbf{b} + N_\mathcal{C}(\theta^*_\lambda),$$

where $\eta^+, \eta^- \in \Re^p_+$, $\gamma \in \Re$ and $N_\mathcal{C}(\theta^*_\lambda)$ is the normal cone of $\mathcal{C}$ at θ^*_λ. Because $\theta^*_\lambda \in \mathcal{C}$ and $\mathcal{C}$ is an open set, θ^*_λ is an interior point of $\mathcal{C}$ and thus $N_\mathcal{C}(\theta^*_\lambda) = \emptyset$. Therefore, the above equation becomes:

$$\nabla g(\theta^*_\lambda) + \sum^p_{j=1}\eta^+_j\bar{\mathbf{x}}^j + \sum^p_{j'=1}\eta^-_{j'}(-\bar{\mathbf{x}}^{j'}) + \gamma\mathbf{b} = 0. \tag{39}$$

Moreover, by the complementary slackness condition [3], we have $\eta^+_j = \eta^-_j = 0$ for $j = 1,\ldots,p$ since $\mathcal{I}_\lambda$ is empty. Then, Eq. (39) becomes:

$$\nabla g(\theta^*_\lambda) + \gamma\mathbf{b} = 0. \tag{40}$$

By the similar argument, the KKT condition for $(\text{LRD}_{\lambda_{max}})$ at $\theta^*_{\lambda_{max}}$ is:

$$\nabla g(\theta^*_{\lambda_{max}}) + \sum^p_{j=1}\bar{\eta}^+_j\bar{\mathbf{x}}^j + \sum^p_{j'=1}\bar{\eta}^-_{j'}(-\bar{\mathbf{x}}^{j'}) + \gamma'\mathbf{b} = 0, \tag{41}$$

where $\bar{\eta}^+, \bar{\eta}^- \in \Re^p_+$ and $\gamma' \in \Re$.

Since $\mathcal{I}_\lambda = \emptyset$, we can see that $|\langle \theta^*_\lambda, \bar{\mathbf{x}}^j\rangle| < m\lambda < m\lambda_{max}$ for all $j \in \{1,\ldots,m\}$. Therefore, θ^*_λ also satisfies Eq. (41) by setting $\eta^+ = \bar{\eta}^+$, $\eta^- = \bar{\eta}^-$ and $\gamma = \gamma'$ without violating the complementary slackness conditions. As a result, θ^*_λ is an optimal solution of problem $(\text{LRD}_{\lambda_{max}})$ as well.

Moreover, it is easy to see that $\theta^*_\lambda \neq \theta^*_{\lambda_{max}}$ because $|\langle \theta^*_\lambda, \bar{\mathbf{x}}^{j_0}\rangle| < m\lambda < m\lambda_{max} = |\langle \theta^*_{\lambda_{max}}, \bar{\mathbf{x}}^{j_0}\rangle|$. Consequently, $(\text{LRD}_{\lambda_{max}})$ has at least two distinct optimal solutions, which contradicts with Lemma B. Therefore, $\mathcal{I}_\lambda$ must be an nonempty set. Because λ is arbitrary in $(0, \lambda_{max})$, the proof is complete.

$\square$

E   Proof of Theorem 4

Proof. Recall that we need to solve the following optimization proble:

$$T(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) := \max_\theta \ \left\{ |\langle \theta, \bar{\mathbf{x}}^j \rangle| : \|\theta - \theta_{\lambda_0}^*\|_2^2 \leq r^2, \langle \theta, \mathbf{b} \rangle = 0, \langle \theta, \bar{\mathbf{x}}^* \rangle \leq m\lambda \right\}. \tag{42}$$

The feasibility of θ implies that

$$\langle \theta, \mathbf{b} \rangle = 0,$$

i.e., θ belongs to the orthogonal complement of the space spanned by $\mathbf{b}$. As a result, $\theta = \mathbf{P}\theta$ and

$$|\langle \theta, \bar{\mathbf{x}}^j \rangle| = |\langle \mathbf{P}\theta, \bar{\mathbf{x}}^j \rangle| = |\langle \theta, \mathbf{P}\bar{\mathbf{x}}^j \rangle| = 0.$$

Therefore, we can see that $T(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = 0$, which completes the proof.

$\square$

F   Proof of Corollary 5

Proof. We set $\lambda_0 = \lambda_{max}$ and $\theta_{\lambda_0}^* = \theta_{\lambda_{max}}^*$. Clearly, we have $\lambda_0 > \lambda > 0$ and $\theta_{\lambda_0}^*$ is known. Therefore, the assumptions in Theorem 4 are satisfied and thus $T(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = 0$.

By the rule in (R1$'$), it is straightforward to see that $[\beta_\lambda^*]_j = 0$, which completes the proof. $\square$

G   Proof of Lemma 6

Proof. We show that the Slater's condition holds for (UBP$'$).

Recall that the feasible set of problem (UBP$'$) is

$$\mathcal{A}_{\lambda_0}^\lambda = \left\{ \theta : \|\theta - \theta_{\lambda_0}^*\|_2^2 \leq r^2, \langle \theta, \mathbf{b} \rangle = 0, \langle \theta, \bar{\mathbf{x}}^* \rangle \leq m\lambda \right\}.$$

To show that the Slater's condition holds, we need to seek a point θ' such that

$$\|\theta' - \theta_{\lambda_0}^*\|_2^2 < r^2, \langle \theta', \mathbf{b} \rangle = 0, \langle \theta', \bar{\mathbf{x}}^* \rangle \leq m\lambda.$$

Consider θ_λ^*, since $\theta_\lambda^* \in \mathcal{A}_{\lambda_0}^\lambda$, the last two constraints of $\mathcal{A}_{\lambda_0}^\lambda$ are satisfied by θ_λ^* automatically.

On the other hand, because $\lambda_0 \in (0, \lambda_{max}]$, Lemma 3 leads to the fact that the active set $\mathcal{I}_{\lambda_0}$ is not empty. Let $j_0 \in \mathcal{I}_{\lambda_0}$, we can see that

$$|\langle \theta_{\lambda_0}^*, \bar{\mathbf{x}}^{j_0} \rangle| = m\lambda_0.$$

However, because $\theta_\lambda^* \in \mathcal{F}_\lambda$, we have

$$|\bar{\mathbf{X}}^T \theta_\lambda^*|_\infty \leq m\lambda,$$

and thus

$$|\langle \theta_\lambda^*, \bar{\mathbf{x}}^{j_0} \rangle| \leq m\lambda < m\lambda_0.$$

Therefore, we can see that $\theta_\lambda^* \neq \theta_{\lambda_0}^*$. By part b of Theorem 2, it follows that

$$\|\theta_\lambda^* - \theta_{\lambda_0}^*\|_2^2 < r^2.$$

As a result, the Slater's condition holds for (UBP$'$) which implies the strong duality of (UBP$'$).

Moreover, it is easy to see that problem (UBP$'$) admits optimal solution in $\mathcal{A}_{\lambda_0}^\lambda$ because the objective function of (UBP$'$) is continuous and $\mathcal{A}_{\lambda_0}^\lambda$ is compact.

$\square$

H Proof of Lemma 7

Proof. The Lagrangian of (UBP′) is

$$L(\theta; u_1, u_2, v) = \langle \theta, \bar{\mathbf{x}} \rangle + \frac{u_1}{2} \left(\|\theta - \theta^*_{\lambda_0}\|_2^2 - r^2 \right) + u_2(\langle \theta, \bar{\mathbf{x}}^* \rangle - m\lambda_2) + v\langle \theta, \mathbf{b} \rangle \tag{43}$$

$$= \langle \theta, \bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b} \rangle + \frac{u_1}{2} \left(\|\theta - \theta^*_{\lambda_0}\|_2^2 - r^2 \right) - u_2 m\lambda_2.$$

where $u_1, u_2 \geq 0$ and $v \in \Re$ are the Lagrangian multipliers. To derive the dual function $\hat{g}(u_1, u_2, v) = \min_\theta L(\theta; u_1, u_2, v)$, we can simply set $\nabla_\theta L(\theta; u_1, u_2, v) = 0$, i.e.,

$$\nabla_\theta L(\theta; u_1, u_2, v) = \bar{\mathbf{x}} + u_1(\theta - \theta^*_{\lambda_1}) + u_2\bar{\mathbf{x}}^* + v\mathbf{b} = 0.$$

When $u_1 \neq 0$, we set

$$\theta = -\frac{1}{u_1}(\bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b}) + \theta^*_{\lambda_0} \tag{44}$$

to minimize $L(\theta; \mathbf{u}, v)$ with $(\mathbf{u}, v)$ fixed. By plugging Eq. (44) to Eq. (43), we can see that the dual function $\hat{g}(u_1, u_2, v)$ is:

$$\hat{g}(u_1, u_2, v) = -\frac{1}{2u_1}\|\bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b}\|_2^2 + \langle \theta^*_{\lambda_0}, \bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b} \rangle - \frac{1}{2}u_1 r^2 - u_2 m\lambda_2 \tag{45}$$

$$= -\frac{1}{2u_1}\|\bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b}\|_2^2 + u_2 m(\lambda_1 - \lambda_2) + \langle \theta^*_{\lambda_0}, \bar{\mathbf{x}} \rangle - \frac{1}{2}u_1 r^2.$$

The second line of (45) is due to the fact that $\langle \theta^*_{\lambda_0}, \bar{\mathbf{x}}^* \rangle = m\lambda_1$ and $\langle \theta^*_{\lambda_0}, \mathbf{b} \rangle = 0$.

When $u_1 = 0$, it is not straightforward to derive the dual function. The difficulty comes from the fact that $\bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b}$ might be 0 for some $u_2' \geq 0$ and v'.

a). Suppose $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} \in (-1, 1]$, we show that $\bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b} \neq 0$ for all $u_2 \geq 0$ and v.

Suppose for contradiction that

$$\bar{\mathbf{x}} + u_2'\bar{\mathbf{x}}^* + v'\mathbf{b} = 0,$$

in which $u_2' \geq 0$ and v'. Then we can see that

$$\mathbf{P}\left(\bar{\mathbf{x}} + u_2'\bar{\mathbf{x}}^* + v'\mathbf{b} \right) = 0.$$

Since $\mathbf{P}\mathbf{b} = 0$, it follows that

$$\mathbf{P}\bar{\mathbf{x}} + u_2'\mathbf{P}\bar{\mathbf{x}}^* = 0, \tag{46}$$

and thus

$$u_2' = -\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2}. \tag{47}$$

Clearly, Eq. (46) implies that $\mathbf{P}\bar{\mathbf{x}}$ and $\mathbf{P}\bar{\mathbf{x}}^*$ are collinear, i.e., $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} \in \{-1, 1\}$. Moreover, by Eq. (47), we know that $\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle \leq 0$. Therefore, it is easy to see that $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} = -1$, which contradicts the assumption. Hence, $\bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b} \neq 0$ for all $u_2 \geq 0$ and v.

We then show that $\hat{g}(u_1, u_2, v) = -\infty$ when $u_1 = 0$ for all $u_2 \geq 0$ and v.

Since $\hat{g}(u_1, u_2, v) = \min_\theta L(\theta; u_1, u_2, v)$, for $u_1 = 0$, $u_2 \geq 0$ and v, we have

$$\hat{g}(0, u_2, v) = \min_\theta L(\theta; 0, u_2, v) \tag{48}$$

$$= \min_\theta \langle \theta, \bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b} \rangle - u_2 m\lambda_2.$$

20

Let $\theta = -t(\bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b})$, it follows that

$$\langle \theta, \bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b} \rangle - u_2 m\lambda_2 = -t\|\bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b}\|_2^2 - u_2 m\lambda_2, \tag{49}$$

and thus

$$\lim_{t\to\infty} -t\|\bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b}\|_2^2 - u_2 m\lambda_2 = -\infty. \tag{50}$$

Therefore, $\hat{g}(u_1, u_2, v) = \min_\theta L(\theta; u_1, u_2, v)$, for $u_1 = 0$, $u_2 \geq 0$ and v.

All together, the dual function is

$$\hat{g}(u_1, u_2, v) = \begin{cases} -\frac{1}{2u_1}\|\bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b}\|_2^2 + u_2 m(\lambda_1 - \lambda_2) + \langle \theta_{\lambda_0}^*, \bar{\mathbf{x}} \rangle - \frac{1}{2}u_1 r^2, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \text{if } u_1 > 0, u_2 \geq 0, \\ -\infty, \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{if } u_1 = 0, u_2 \geq 0. \end{cases} \tag{51}$$

Because the dual problem is to maximize the dual function, we can write the dual problem as:

$$\max_{u_1 > 0, u_2 \geq 0, v} \hat{g}(u_1, u_2, v) = -\frac{1}{2u_1}\|\bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b}\|_2^2 + u_2 m(\lambda_1 - \lambda_2) + \langle \theta_{\lambda_0}^*, \bar{\mathbf{x}} \rangle - \frac{1}{2}u_1 r^2. \tag{52}$$

Moreover, it is easy to see that problem (52) is an unconstrained optimization problem with respect to v. Therefore, we set

$$\frac{\partial \hat{g}(u_1, u_2, v)}{\partial v} = -\frac{1}{u_1}\langle \bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b}, \mathbf{b} \rangle = 0,$$

and thus

$$v = -\frac{\langle \bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^*, \mathbf{b} \rangle}{\|\mathbf{b}\|_2^2}. \tag{53}$$

By plugging (53) into $\hat{g}(u_1, u_2, v)$ and noting that $\mathcal{U}_1 = \{(u_1, u_2) : u_1 > 0, u_2 \geq 0\}$, problem (52) is equivalent to

$$\max_{(u_1, u_2) \in \mathcal{U}_1} \bar{g}(u_1, u_2) = -\frac{1}{2u_1}\|\mathbf{P}\bar{\mathbf{x}} + u_2\mathbf{P}\bar{\mathbf{x}}^*\|_2^2 + u_2 m(\lambda_1 - \lambda_2) + \langle \theta_{\lambda_1}^*, \bar{\mathbf{x}} \rangle - \frac{1}{2}u_1 r^2. \tag{UBD'}$$

By Lemma 6, we know that the Slater's conditions holds for (UBP'). Therefore, the strong duality holds for (UBP') and (UBD'). By the strong duality theorem [11], there exists $u_1^* \geq 0$, $u_2^* \geq 0$ and v^* such that $\hat{g}(u_1^*, u_2^*, v^*) = \max_{u_1 \geq 0, u_2 \geq 0, v} \hat{g}(u_1, u_2, v)$. By Eq. (51), it is easy to see that $u_1^* > 0$. Therefore, $\bar{g}(u_1, u_2)$ attains its maximum in $\mathcal{U}_1$.

b). Suppose $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} = -1$, it is easy to see that

$$\bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b} = 0 \Leftrightarrow u_2 = -\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2} \text{ and } v = -\frac{\langle \bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^*, \mathbf{b} \rangle}{\|\mathbf{b}\|_2^2}.$$

Let $u_2' = -\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2}$ and $v' = -\frac{\langle \bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^*, \mathbf{b} \rangle}{\|\mathbf{b}\|_2^2}$, we have

$$L(\theta; 0, u_2', v') = -u_2' m\lambda_2 = \frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2} m\lambda = -\frac{\|\mathbf{P}\bar{\mathbf{x}}\|_2}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2} m\lambda,$$

and thus

$$\hat{g}(0, u_2', v') = \min_\theta L(\theta; 0, u_2', v') = -\frac{\|\mathbf{P}\bar{\mathbf{x}}\|_2}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2} m\lambda.$$

By a similar argument as in the proof of part 1, we can see that the dual function is

$$\hat{g}(u_1, u_2, v) = \begin{cases} -\frac{1}{2u_1}\|\bar{\mathbf{x}} + u_2\bar{\mathbf{x}}^* + v\mathbf{b}\|_2^2 + u_2 m(\lambda_1 - \lambda_2) + \langle\theta_{\lambda_0}^*, \bar{\mathbf{x}}\rangle - \frac{1}{2}u_1 r^2, \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } u_1 > 0, u_2 \geq 0, \\ -\infty, \qquad\qquad\qquad\qquad\quad \text{if } u_1 = 0, u_2 \geq 0, u_2 \neq u_2', v \neq v', \\ -\frac{\|\mathbf{P}\bar{\mathbf{x}}\|_2}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2}m\lambda, \qquad\qquad \text{if } u_1 = 0, u_2 = u_2', v = v', \end{cases} \tag{54}$$

and the dual problem is equivalent to

$$\max_{(u_1, u_2) \in \mathcal{U}_1 \cup \mathcal{U}_2} \bar{\bar{g}}(u_1, u_2) = \begin{cases} \bar{g}(u_1, u_2), & \text{if } (u_1, u_2) \in \mathcal{U}_1, \\ -\frac{\|\mathbf{P}\bar{\mathbf{x}}\|_2}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2}m\lambda, & \text{if } (u_1, u_2) \in \mathcal{U}_2. \end{cases} \tag{UBD''}$$

$\square$

I Proof of Theorem 8

I.1 Proof for the Non-Collinear Case

In this section, we show that the results in Theorem 8 holds for the case in which $\frac{\langle\mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^*\rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2\|\mathbf{P}\bar{x}^*\|_2} \in (-1, 1)$, i.e., $\mathbf{P}\bar{\mathbf{x}}$ and $\mathbf{P}\bar{x}^*$ are not collinear.

Proof. As shown by part a) of Lemma 7, the dual problem of (UBP′) is equivalent to (UBD′). Since the strong duality holds by Lemma 6, we have

$$-T_\xi(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = \max_{u_1 > 0, u_2 \geq 0} \bar{g}(u_1, u_2). \tag{55}$$

Clearly, problem (UBD′) can be solved via the following minimization problem

$$T_\xi(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = \min_{u_1 > 0, u_2 \geq 0} -\bar{g}(u_1, u_2). \tag{56}$$

Let u_1^* and u_2^* be the optimal solution of problem (56). By part a) of Lemma 7, the existence of $u_1^* > 0$ and $u_2^* \geq 0$ is guaranteed. By introducing the slack variables $s_1 \geq 0$ and $s_2 \geq 0$, the KKT conditions of problem (56) can be written as follows:

$$-\frac{1}{2(u_1^*)^2}\|\mathbf{P}\bar{\mathbf{x}} + u_2^*\mathbf{P}\bar{\mathbf{x}}^*\|_2^2 + \frac{1}{2}r^2 - s_1 = 0, \tag{57}$$

$$\frac{1}{u_1^*}\langle\mathbf{P}\bar{\mathbf{x}} + u_2^*\mathbf{P}\bar{\mathbf{x}}^*, \mathbf{P}\bar{\mathbf{x}}^*\rangle - m(\lambda_0 - \lambda) - s_2 = 0, \tag{58}$$

$$s_1 u_1^* = 0, \quad s_2 u_2^* = 0. \tag{59}$$

Since $u_1^* > 0$, then $s_1 = 0$ and Eq. (64) results in

$$u_1^* = \|\mathbf{P}\bar{\mathbf{x}} + u_2^*\mathbf{P}\bar{\mathbf{x}}^*\|_2/r. \tag{60}$$

By plugging u_1^* into (65), we have

$$\varphi(u_2^*) := \frac{\langle\mathbf{P}\bar{\mathbf{x}} + u_2^*\mathbf{P}\bar{\mathbf{x}}^*, \mathbf{P}\bar{\mathbf{x}}^*\rangle}{\|\mathbf{P}\bar{\mathbf{x}} + u_2^*\mathbf{P}\bar{\mathbf{x}}^*\|_2\|\mathbf{P}\bar{\mathbf{x}}^*\|_2} = \frac{m(\lambda_0 - \lambda) + s_2}{r\|\mathbf{P}\bar{\mathbf{x}}^*\|_2} = d + \frac{s_2}{r\|\mathbf{P}\bar{\mathbf{x}}^*\|_2}. \tag{61}$$

It is easy to observe that

m1). since $\frac{\langle\mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^*\rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2\|\mathbf{P}\bar{x}^*\|_2} \in (-1, 1)$, i.e., $\mathbf{P}\bar{\mathbf{x}}$ and $\mathbf{P}\bar{x}^*$ are not collinear, $\varphi(u_2^*)$ monotonically increases when $u_2^* \to \infty$ and $\lim_{u_2^* \to \infty} \varphi(u_2^*) = 1$;

m2). $d \in (0,1]$ due to the fact that $\lambda_0 > \lambda$ and Eq. (68) (note $s_2 \geq 0$).

a): Assume $\frac{\langle \mathbf{P\bar{x}}, \mathbf{P\bar{x}^*} \rangle}{\|\mathbf{P\bar{x}}\|_2 \|\mathbf{P\bar{x}^*}\|_2} \geq d$. We divide the proof into two cases.

a1) Suppose $\frac{\langle \mathbf{P\bar{x}}, \mathbf{P\bar{x}^*} \rangle}{\|\mathbf{P\bar{x}}\|_2 \|\mathbf{P\bar{x}^*}\|_2} > d$, Eq. (68) and the monotonicity of φ imply that

$$d + \frac{s_2}{r\|\mathbf{P\bar{x}^*}\|_2} = \varphi(u_2^*) \geq \varphi(0) = \frac{\langle \mathbf{P\bar{x}}, \mathbf{P\bar{x}^*} \rangle}{\|\mathbf{P\bar{x}}\|_2 \|\mathbf{P\bar{x}^*}\|_2} > d.$$

Therefore, we can see that $s_2 > 0$ and thus $u_2^* = 0$ due to the complementary slackness condition. By plugging $u_2^* = 0$ into Eq. (67), we can get $u_1^* = \frac{\|\mathbf{P\bar{x}}\|_2}{r}$. The result in part a) of Theorem 8 follows by noting that $T_\xi(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = -\bar{g}(u_1^*, u_2^*)$.

a2) Suppose $\frac{\langle \mathbf{P\bar{x}}, \mathbf{P\bar{x}^*} \rangle}{\|\mathbf{P\bar{x}}\|_2 \|\mathbf{P\bar{x}^*}\|_2} = d$. If $u_2^* > 0$, then $s_2 = 0$ by the complementary slackness condition. In view of Eq. (68) and **m1)**, we can see that

$$d = \varphi(u_2^*) > \varphi(0) = \frac{\langle \mathbf{P\bar{x}}, \mathbf{P\bar{x}^*} \rangle}{\|\mathbf{P\bar{x}}\|_2 \|\mathbf{P\bar{x}^*}\|_2} = d,$$

which leads to a contradiction. Therefore $u_2^* = 0$ and the result in part a) of Theorem 8 follows by a similar argument as in the proof of a1).

b): Assume $\frac{\langle \mathbf{P\bar{x}}, \mathbf{P\bar{x}^*} \rangle}{\|\mathbf{P\bar{x}}\|_2 \|\mathbf{P\bar{x}^*}\|_2} < d$. If $u_2^* = 0$, Eq. (68) results in

$$d + \frac{s_2}{r\|\mathbf{P\bar{x}^*}\|_2} = \varphi(0) = \frac{\langle \mathbf{P\bar{x}}, \mathbf{P\bar{x}^*} \rangle}{\|\mathbf{P\bar{x}}\|_2 \|\mathbf{P\bar{x}^*}\|_2} < d,$$

which implies that $s_2 < 0$, a contradiction. Thus, we have $u_2^* > 0$ and $s_2 = 0$ by the complementary slackness condition. Eq. (68) becomes:

$$\frac{\langle \mathbf{P\bar{x}} + u_2^* \mathbf{P\bar{x}^*}, \mathbf{P\bar{x}^*} \rangle}{\|\mathbf{P\bar{x}} + u_2^* \mathbf{P\bar{x}^*}\|_2 \|\mathbf{P\bar{x}^*}\|_2} = \frac{m(\lambda_0 - \lambda)}{r\|\mathbf{P\bar{x}^*}\|_2} = d. \tag{62}$$

Expanding the terms in Eq. (62) yields the following quadratic equation:

$$a_2(u_2^*)^2 + a_1 u_2^* + a_0 = 0, \tag{63}$$

where a_0, a_1 and a_2 are given by Eq. (86).

On the other hand, Eq. (62) implies that $d \leq 1$. In view of the assumptions $\frac{\langle \mathbf{P\bar{x}}, \mathbf{P\bar{x}^*} \rangle}{\|\mathbf{P\bar{x}}\|_2 \|\mathbf{P\bar{x}^*}\|_2} < d$ and $\frac{\langle \mathbf{P\bar{x}}, \mathbf{P\bar{x}^*} \rangle}{\|\mathbf{P\bar{x}}\|_2 \|\mathbf{P\bar{x}^*}\|_2} \neq -1$, we can see that $\mathbf{P\bar{x}^*}$ and $\mathbf{P\bar{x}}$ are not collinear. Therefore, **m1)** and Eq. (62) imply that $d < 1$. Moreover, the assumption $\lambda_0 > \lambda$ leads to $d > 0$. As a result, we have $(1 - d^2) > 0$ and thus

$$a_2 a_0 < 0, \Delta = a_1^2 - 4a_2 a_0 = 4d^2(1 - d^2)\|\mathbf{P\bar{x}^*}\|_2^4(\|\mathbf{P\bar{x}}\|_2^2\|\mathbf{P\bar{x}^*}\|_2^2 - \langle \mathbf{P\bar{x}}, \mathbf{P\bar{x}^*} \rangle^2) > 0.$$

Consequently, (63) has only one positive solution which can be computed by the formula in Eq. (86). The result in Eq. (85) follows by a similar argument as in the proof of a1). $\square$

I.2 Proof for the Collinear and Positive Correlated Case

We prove for the case in which $\frac{\langle \mathbf{P\bar{x}}, \mathbf{P\bar{x}^*} \rangle}{\|\mathbf{P\bar{x}}\|_2 \|\mathbf{P\bar{x}^*}\|_2} = 1$.

Proof. Because $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} = 1$, by part a) of Lemma 7, the dual problem of (UBP′) is given by (UBD′). Therefore, the following KKT conditions hold as well:

$$-\frac{1}{2(u_1^*)^2}\|\mathbf{P}\bar{\mathbf{x}} + u_2^* \mathbf{P}\bar{\mathbf{x}}^*\|_2^2 + \frac{1}{2}r^2 - s_1 = 0, \tag{64}$$

$$\frac{1}{u_1^*}\langle \mathbf{P}\bar{\mathbf{x}} + u_2^* \mathbf{P}\bar{\mathbf{x}}^*, \mathbf{P}\bar{\mathbf{x}}^* \rangle - m(\lambda_1 - \lambda_2) - s_2 = 0, \tag{65}$$

$$s_1 u_1^* = 0, \; s_2 u_2^* = 0 \tag{66}$$

where $u_1^* > 0$ and $u_2^* \geq 0$ are the optimal solution of (UBD′), and $s_1, s_2 \geq 0$ are the slack variables. Since $u_1^* > 0$, then $s_1 = 0$ and Eq. (64) results in

$$u_1^* = \|\mathbf{P}\bar{\mathbf{x}} + u_2^* \mathbf{P}\bar{\mathbf{x}}^*\|_2 / r. \tag{67}$$

By plugging u_1^* into (65), we have

$$\varphi(u_2^*) := \frac{\langle \mathbf{P}\bar{\mathbf{x}} + u_2^* \mathbf{P}\bar{\mathbf{x}}^*, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}} + u_2^* \mathbf{P}\bar{\mathbf{x}}^*\|_2 \|\mathbf{P}\bar{\mathbf{x}}^*\|_2} = \frac{m(\lambda_0 - \lambda) + s_2}{r\|\mathbf{P}\bar{\mathbf{x}}^*\|_2} = d + \frac{s_2}{r\|\mathbf{P}\bar{\mathbf{x}}^*\|_2}, \tag{68}$$

which implies

$$d \leq 1. \tag{69}$$

Moreover, we can see that the assumption $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} = 1$ actually implies $\mathbf{P}\bar{\mathbf{x}}$ and $\mathbf{P}\bar{x}^*$ are collinear. Therefore, there exists $\alpha > 0$ such that

$$\mathbf{P}\bar{\mathbf{x}} = \alpha \mathbf{P}\bar{\mathbf{x}}^*. \tag{70}$$

By plugging Eq. (70) into Eq. (68), we have

$$\varphi(u_2^*) = \frac{\langle \mathbf{P}\bar{\mathbf{x}} + u_2^* \mathbf{P}\bar{\mathbf{x}}^*, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}} + u_2^* \mathbf{P}\bar{\mathbf{x}}^*\|_2 \|\mathbf{P}\bar{\mathbf{x}}^*\|_2} = \frac{(\alpha + u_2^*)\langle \mathbf{P}\bar{\mathbf{x}}^*, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{(\alpha + u_2^*)\|\mathbf{P}\bar{\mathbf{x}}^*\|_2 \|\mathbf{P}\bar{\mathbf{x}}^*\|_2} = 1, \tag{71}$$

i.e., $\varphi(u_2^*)$ is a constant and does not depend on the value of u_2^*.

As a result, we can see that $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} = 1 \geq d$. Therefore, we need to show that the result in part a) of Theorem 8 holds. We divide the proof into two cases.

Case 1. Suppose $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} > d$, i.e., $d < 1$, Eq. (68) and Eq. (71) imply that

$$d + \frac{s_2}{r\|\mathbf{P}\bar{\mathbf{x}}^*\|_2} = \varphi(u_2^*) = 1 > d.$$

Therefore, we can see that $s_2 > 0$ and thus $u_2^* = 0$ due to the complementary slackness condition. By plugging $u_2^* = 0$ into Eq. (67), we can get $u_1^* = \frac{\|\mathbf{P}\bar{\mathbf{x}}\|_2}{r}$. The result in part a) of Theorem 8 follows by noting that $T_\xi(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = -\bar{g}(u_1^*, u_2^*)$.

Case 2. Suppose $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} = d = 1$. On the other hand, in view of Eq. (68) and Eq. (71), we can observe that $s_2 = 0$ because otherwise

$$\varphi(u_2^*) = 1 = d + \frac{s_2}{r\|\mathbf{P}\bar{\mathbf{x}}^*\|_2} > 1.$$

By plugging 70 into Eq. (67), we have $u_1^* = (\alpha + u_2^*)\|\mathbf{P}\bar{\mathbf{x}}^*\|_2 / r$ and thus

$$T_\xi(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = -\bar{g}(u_1^*, u_2^*) = r(\alpha + u_2^*)\|\mathbf{P}\bar{\mathbf{x}}^*\|_2 - u_2^* m(\lambda_0 - \lambda) - \langle \theta_{\lambda_0}^*, \bar{\mathbf{x}} \rangle. \tag{72}$$

By noting $d = 1$, it follows that

$$m(\lambda_0 - \lambda) = r\|\mathbf{P}\bar{\mathbf{x}}^*\|_2. \tag{73}$$

Therefore, in view of Eq. (70) and Eq. (73), Eq. (72) becomes

$$T_\xi(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = r(\alpha + u_2^*)\|\mathbf{P}\bar{\mathbf{x}}^*\|_2 - u_2^* r\|\mathbf{P}\bar{\mathbf{x}}^*\|_2 - \langle \theta_{\lambda_0}^*, \bar{\mathbf{x}} \rangle = r\|\mathbf{P}\bar{\mathbf{x}}\|_2 - \langle \theta_{\lambda_0}^*, \bar{\mathbf{x}} \rangle,$$

which is the same as the result in part a) of Theorem 8.

$$\square$$

I.3 Proof for the Collinear and Negative Correlated Case

Before we proceed to prove Theorem 8 for the case in which $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} = -1$, it is worthwhile to noting the following lemma.

Lemma 12. *C Let $\lambda_{max} \geq \lambda_0 > \lambda > 0$, $d = \frac{m(\lambda_0 - \lambda)}{r\|\mathbf{P}\bar{x}^*\|_2}$ and assume $\theta_{\lambda_0}^*$ is known. Then $d \in (0,1]$.*

Proof. Let $\bar{\mathbf{x}} := -\xi(-\bar{\mathbf{x}}^*)$. Clearly, we can see that

$$\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} = 1.$$

Therefore, the results in Section I.2 apply. As a result, in view of the inequality in (69),

Notice that, d is independent of $\bar{\mathbf{x}}$. In other words, as long as there exists a $j_0 \in \{1, \ldots, p\}$ such that $\mathbf{P}\bar{\mathbf{x}}^{j_0} \neq 0$ and $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} \in (-1, 1)$, one can see that

$$d \leq 1.$$

Moreover, since $\lambda_0 > \lambda$, it is easy to see that $d > 0$, which completes the proof. $\square$

We prove Theorem 8 for the case in which $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{x}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} = -1$.

Proof. Consider the case in which $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} = -1$. Clearly, $\mathbf{P}\bar{\mathbf{x}}$ and $\mathbf{P}\bar{\mathbf{x}}^*$ are collinear and there exists $\alpha = -\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2}$ such that

$$\mathbf{P}\bar{\mathbf{x}} = -\alpha \mathbf{P}\bar{\mathbf{x}}^*. \tag{74}$$

As shown by part b) of Lemma 7, the dual problem is given by (UBD″). Therefore, to find

$$T_\xi(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = \min_{(u_1, u_2) \in \mathcal{U}_1 \cup \mathcal{U}_2} -\bar{\bar{g}}(u_1, u_2), \tag{75}$$

we can first compute

$$T_\xi'(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = \inf_{(u_1, u_2) \in \mathcal{U}_1} -\bar{\bar{g}}(u_1, u_2) \tag{76}$$

and then compare $T_\xi'(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*)$ with $-\bar{\bar{g}}(u_1, u_2)|_{(u_1, u_2) \in \mathcal{U}_2} = \frac{\|\mathbf{P}\bar{\mathbf{x}}\|_2}{\mathbf{P}\bar{\mathbf{x}}^*} m\lambda$. Clearly, we have

$$T_\xi(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*) = \min\left\{ T_\xi'(\theta_\lambda^*, \bar{\mathbf{x}}^j; \theta_{\lambda_0}^*), \frac{\|\mathbf{P}\bar{\mathbf{x}}\|_2}{\mathbf{P}\bar{\mathbf{x}}^*} m\lambda \right\} \tag{77}$$

Let us consider the problem in (76). From problem (UBD″), we observe that

$$-\bar{\bar{g}}(u_1, u_2)|_{(u_1, u_2) \in \mathcal{U}_1} = -\bar{g}(u_1, u_2). \tag{78}$$

By noting Eq. (74), we have

$$-\bar{\bar{g}}(u_1, u_2)|_{(u_1, u_2) \in \mathcal{U}_1} = \frac{1}{2u_1}(u_2 - \alpha)^2 \|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2 - u_2 m(\lambda_0 - \lambda) + \frac{1}{2} u_1 r^2 - \langle \theta_{\lambda_0}^*, \bar{\mathbf{x}} \rangle. \tag{79}$$

Suppose u_1 is fixed, it is easy to see that

$$u_2^*(u_1) = \underset{u_2 \geq 0}{\operatorname{argmin}} -\bar{\bar{g}}(u_1, u_2)|_{(u_1, u_2) \in \mathcal{U}_1} = \frac{u_1 m(\lambda_0 - \lambda)}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2} + \alpha \tag{80}$$

and

$$h(u_1) := \min_{u_2} -\bar{\bar{g}}(u_1, u_2)|_{(u_1, u_2) \in \mathcal{U}_1} = -\bar{\bar{g}}(u_1, u_2^*(u_1)) \tag{81}$$

$$= \frac{1}{2} u_1 r^2 (1 - d^2) - \alpha m(\lambda_0 - \lambda) - \langle \theta_{\lambda_0}^*, \bar{\mathbf{x}} \rangle.$$

25

Clearly, the domain of $h(u_1)$ is $u_1 > 0$ and $T'_\xi(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}) = \inf_{u_1 > 0} h(u_1)$.

By Lemma C, one can see that that $d \in (0, 1]$.

If $d = 1$, then

$$\min_{u_1 > 0} h(u_1) = -\alpha m(\lambda_0 - \lambda) - \langle \theta^*_{\lambda_0}, \bar{\mathbf{x}} \rangle.$$

Otherwise, if $d \in (0, 1)$, it is easy to see that

$$\inf_{u_1 > 0} h(u_1) = \lim_{u_1 \downarrow 0} h(u_1) = -\alpha m(\lambda_0 - \lambda) - \langle \theta^*_{\lambda_0}, \bar{\mathbf{x}} \rangle.$$

All together, we have

$$T'_\xi(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}) = -\alpha m(\lambda_0 - \lambda) - \langle \theta^*_{\lambda_0}, \bar{\mathbf{x}} \rangle. \tag{82}$$

Moreover, we observe that

$$\langle \theta^*_{\lambda_0}, \bar{\mathbf{x}} \rangle = \langle \mathbf{P}\theta^*_{\lambda_0}, \bar{\mathbf{x}} \rangle = \langle \theta^*_{\lambda_0}, \mathbf{P}\bar{\mathbf{x}} \rangle = \langle \theta^*_{\lambda_0}, -\alpha \mathbf{P}\bar{\mathbf{x}}^* \rangle = -\alpha m \lambda_0.$$

Therefore, Eq. (82) can simplified as:

$$T'_\xi(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}) = -\alpha m(\lambda_0 - \lambda) - \langle \theta^*_{\lambda_0}, \bar{\mathbf{x}} \rangle = \alpha m \lambda = -\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2} m \lambda = \frac{\|\mathbf{P}\bar{\mathbf{x}}\|_2}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2} m \lambda. \tag{83}$$

By Eq. (77), we can see that

$$T_\xi(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}) = \frac{\|\mathbf{P}\bar{\mathbf{x}}\|_2}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2} m \lambda. \tag{84}$$

Since $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} = -1$ and $d \in (0, 1]$ by Lemma C, one can see that $\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}\|_2 \|\mathbf{P}\bar{x}^*\|_2} < d$. Therefore, we only need to show that part b) of Theorem 8 yields the same solution as Eq. (84) .

Recall that part b) of Theorem 8 says:

$$T_\xi(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}) = r\|\mathbf{P}\bar{\mathbf{x}} + u_2^* \mathbf{P}\bar{\mathbf{x}}^*\|_2 - u_2^* m(\lambda_0 - \lambda) - \langle \theta^*_{\lambda_0}, \bar{\mathbf{x}} \rangle, \tag{85}$$

where

$$\begin{aligned}
u_2^* &= \frac{-a_1 + \sqrt{\Delta}}{2a_2}, \\
a_2 &= \|\mathbf{P}\bar{\mathbf{x}}^*\|_2^4(1 - d^2), \\
a_1 &= 2\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle \|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2(1 - d^2), \\
a_0 &= \langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle^2 - d^2\|\mathbf{P}\bar{\mathbf{x}}\|_2^2\|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2, \\
\Delta &= a_1^2 - 4a_2 a_0 = 4d^2(1 - d^2)\|\mathbf{P}\bar{\mathbf{x}}^*\|_2^4(\|\mathbf{P}\bar{\mathbf{x}}\|_2^2\|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2 - \langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle^2).
\end{aligned} \tag{86}$$

In fact, if we plugging Eq. (74) into Eq. (86), we have

$$\Delta = 0,$$
$$u_2^* = \frac{-a_1}{2a_2} = -\frac{\langle \mathbf{P}\bar{\mathbf{x}}, \mathbf{P}\bar{\mathbf{x}}^* \rangle}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2^2}.$$

Therefore, we can see that $u_2^* = \alpha$, $\mathbf{P}\bar{\mathbf{x}} + u_2^* \mathbf{P}\bar{x}^* = 0$ and thus Eq. (85) results in

$$T_\xi(\theta^*_\lambda, \bar{\mathbf{x}}^j; \theta^*_{\lambda_0}) = -\alpha m(\lambda_0 - \lambda) - \langle \theta^*_{\lambda_0}, \bar{\mathbf{x}} \rangle = \alpha m \lambda = \frac{\|\mathbf{P}\bar{\mathbf{x}}\|_2}{\|\mathbf{P}\bar{\mathbf{x}}^*\|_2} m \lambda. \tag{87}$$

Clearly, Eq. (84) and Eq. (87) give the same result, which completes the proof. $\square$