Journal of Machine Learning Research 16 (2015) 1063-1101 Submitted 1/14; Revised 9/14; Published 5/15

Lasso Screening Rules via Dual Polytope Projection

Jie Wang JWANGUMI@UMICH.EDU
Department of Computational Medicine and Bioinformatics
University of Michigan
Ann Arbor, MI 48109-2218, USA
Peter Wonka PETER.WONKA@ASU.EDU
Department of Computer Science and Engineering
Arizona State University
Tempe, AZ 85287-8809, USA
Jieping Ye JPYE@UMICH.EDU
Department of Computational Medicine and Bioinformatics
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109-2218, USA

Editor: Hui Zou

Abstract

Lasso is a widely used regression technique to find sparse representations. When the dimension of the feature space and the number of samples are extremely large, solving the Lasso problem remains challenging. To improve the efficiency of solving large-scale Lasso problems, El Ghaoui and his colleagues have proposed the SAFE rules which are able to quickly identify the inactive predictors, i.e., predictors that have 0 components in the solution vector. Then, the inactive predictors or features can be removed from the optimization problem to reduce its scale. By transforming the standard Lasso to its dual form, it can be shown that the inactive predictors include the set of inactive constraints on the optimal dual solution. In this paper, we propose an efficient and effective screening rule via Dual Polytope Projections (DPP), which is mainly based on the uniqueness and nonexpansiveness of the optimal dual solution due to the fact that the feasible set in the dual space is a convex and closed polytope. Moreover, we show that our screening rule can be extended to identify inactive groups in group Lasso. To the best of our knowledge, there is currently no exact screening rule for group Lasso. We have evaluated our screening rule using synthetic and real data sets. Results show that our rule is more effective in identifying inactive predictors than existing state-of-the-art screening rules for Lasso.

Keywords: lasso, safe screening, sparse regularization, polytope projection, dual formulation, large-scale optimization

1. Introduction

Data with various structures and scales comes from almost every aspect of daily life. To effectively extract patterns in the data and build interpretable models with high prediction accuracy is always desirable. One popular technique to identify important explanatory features is by sparse regularization. For instance, consider the widely used ℓ_1-regularized

©2015 Jie Wang, Peter Wonka and Jieping Ye.

least squares regression problem known as Lasso (Tibshirani, 1996). The most appealing property of Lasso is the sparsity of the solutions, which is equivalent to feature selection. Suppose we have N observations and p features. Let $\mathbf{y}$ denote the N dimensional response vector and $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$ be the $N \times p$ feature matrix. Let $\lambda \geq 0$ be the regularization parameter. The Lasso problem is formulated as the following optimization problem:

$$\inf_{\beta \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1. \tag{1}$$

Lasso has achieved great success in a wide range of applications (Chen et al., 2001; Candès, 2006; Zhao and Yu, 2006; Bruckstein et al., 2009; Wright et al., 2010) and in recent years many algorithms have been developed to efficiently solve the Lasso problem (Efron et al., 2004; Kim et al., 2007; Park and Hastie, 2007; Donoho and Tsaig, 2008; Friedman et al., 2007; Becker et al., 2010; Friedman et al., 2010). However, when the dimension of feature space and the number of samples are very large, solving the Lasso problem remains challenging because we may not even be able to load the data matrix into main memory. The idea of *screening* has been shown very promising in solving Lasso for large-scale problems. Essentially, screening aims to quickly identify the *inactive features* that have 0 components in the solution and then remove them from the optimization. Therefore, we can work on a reduced feature matrix to solve the Lasso problem, which may lead to substantial savings in computational cost and memory usage.

Existing screening methods for Lasso can be roughly divided into two categories: the *Heuristic Screening Methods* and the *Safe Screening Methods*. As the name indicated, the heuristic screening methods can not guarantee that the discarded features have zero coefficients in the solution vector. In other words, they may mistakenly discard the active features which have nonzero coefficients in the sparse representations. Well-known heuristic screening methods for Lasso include SIS (Fan and Lv, 2008) and strong rules (Tibshirani et al., 2012). SIS is based on the associations between features and the prediction task, but not from an optimization point of view. Strong rules rely on the assumption that the absolute values of the inner products between features and the residue are *nonexpansive* (Bauschke and Combettes, 2011) with respect to the parameter values. Notice that, in real applications, this assumption is not always true. In order to ensure the correctness of the solutions, strong rules check the KKT conditions for violations. In case of violations, they weaken the screened set and repeat this process. In contrast to the heuristic screening methods, the safe screening methods for Lasso can guarantee that the discarded features are absent from the resulting sparse models. Existing safe screening methods for Lasso includes SAFE (El Ghaoui et al., 2012) and DOME (Xiang et al., 2011), which are based on an estimation of the dual optimal solution. The key challenge of searching for effective safe screening rules is how to accurately estimate the dual optimal solution. The more accurate the estimation is, the more effective the resulting screening rule is in discarding the inactive features. Moreover, Xiang et al. (2011) have shown that the SAFE rule for Lasso can be read as a special case of their testing rules.

In this paper, we develop novel efficient and effective screening rules for the Lasso problem; our screening rules are safe in the sense that no active features will be discarded. As the name indicated (DPP), the proposed approaches heavily rely on the geometric properties of the Lasso problem. Indeed, the dual problem of problem (1) can be formulated as

a projection problem. More specifically, the **d**ual optimal solution of the Lasso problem is the **p**rojection of the scaled response vector onto a nonempty closed and convex **p**olytope (the feasible set of the dual problem). This nice property provides us many elegant approaches to accurately estimate the dual optimal solutions, e.g., nonexpansiveness, firmly nonexpansiveness (Bauschke and Combettes, 2011). In fact, the estimation of the dual optimal solution in DPP is a direct application of the nonexpansiveness of the projection operators. Moreover, by further exploiting the properties of the projection operators, we can significantly improve the estimation of the dual optimal solution. Based on this estimation, we develop the so called *enhanced DPP* (EDPP) rules which are able to detect far more inactive features than DPP. Therefore, the speedup gained by EDPP is much higher than the one by DPP.

In real applications, the optimal parameter value of λ is generally unknown and needs to be estimated. To determine an appropriate value of λ, commonly used approaches such as cross validation and stability selection involve solving the Lasso problems over a grid of tuning parameters $\lambda_1 > \lambda_2 > \ldots > \lambda_{\mathcal{K}}$. Thus, the process can be very time consuming. To address this challenge, we develop the sequential version of the DPP families. Briefly speaking, for the Lasso problem, suppose we are given the solution $\beta^*(\lambda_{k-1})$ at λ_{k-1}. We then apply the screening rules to identify the inactive features of problem (1) at λ_k by making use of $\beta^*(\lambda_{k-1})$. The idea of the sequential screening rules is proposed by El Ghaoui et al. (2012) and Tibshirani et al. (2012) and has been shown to be very effective for the aforementioned scenario. In Tibshirani et al. (2012), the authors demonstrate that the sequential strong rules are very effective in discarding inactive features especially for very small parameter values and achieve the state-of-the-art performance. However, in contrast to the recursive SAFE (the sequential version of SAFE rules) and the sequential version of DPP rules, it is worthwhile to mention that the sequential strong rules may mistakenly discard active features because they are heuristic methods. Moreover, it is worthwhile to mention that, for the existing screening rules including SAFE and strong rules, the basic versions are usually special cases of their sequential versions, and the same applies to our DPP and EDPP rules. For the DOME rule (Xiang et al., 2011), it is unclear whether its sequential version exists.

The rest of this paper is organized as follows. We present the family of DPP screening rules, i.e., DPP and EDPP, in detail for the Lasso problem in Section 2. Section 3 extends the idea of DPP screening rules to identify inactive groups in group Lasso (Yuan and Lin, 2006). We evaluate our screening rules on synthetic and real data sets from many different applications. In Section 4, the experimental results demonstrate that our rules are more effective in discarding inactive features than existing state-of-the-art screening rules. We show that the efficiency of the solver can be improved by *several orders of magnitude* with the enhanced DPP rules, especially for the high-dimensional data sets (notice that, the screening methods can be integrated with any existing solvers for the Lasso problem). Some concluding remarks are given in Section 5.

2. Screening Rules for Lasso via Dual Polytope Projections

In this section, we present the details of the proposed DPP and EDPP screening rules for the Lasso problem. We first review some basics of the dual problem of Lasso including its

geometric properties in Section 2.1; we also briefly discuss some basic guidelines for developing safe screening rules for Lasso. Based on the geometric properties discussed in Section 2.1, we then develop the basic DPP screening rule in Section 2.2. As a straightforward extension in dealing with the model selection problems, we also develop the sequential version of DPP rules. In Section 2.3, by exploiting more geometric properties of the dual problem of Lasso, we further improve the DPP rules by developing the so called *enhanced DPP* (EDPP) rules. The EDPP screening rules significantly outperform DPP rules in identifying the inactive features for the Lasso problem.

2.1 Basics

Different from Xiang et al. (2011), we do not assume $\mathbf{y}$ and all $\mathbf{x}_i$ have unit length.The dual problem of problem (1) takes the form of (to make the paper self-contained, we provide the detailed derivation of the dual form in the appendix):

$$\sup_{\theta} \quad \left\{ \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{\lambda^2}{2}\left\|\theta - \frac{\mathbf{y}}{\lambda}\right\|_2^2 : \ |\mathbf{x}_i^T\theta| \leq 1, \, i = 1, 2, \ldots, p \right\}, \tag{2}$$

where θ is the dual variable. For notational convenience, let the optimal solution of problem (2) be $\theta^*(\lambda)$ [recall that the optimal solution of problem (1) with parameter λ is denoted by $\beta^*(\lambda)$]. Then, the KKT conditions are given by:

$$\mathbf{y} = \mathbf{X}\beta^*(\lambda) + \lambda\theta^*(\lambda), \tag{3}$$

$$\mathbf{x}_i^T\theta^*(\lambda) \in \begin{cases} \text{sign}([\beta^*(\lambda)]_i), & \text{if } [\beta^*(\lambda)]_\text{i} \neq 0, \\ [-1, 1], & \text{if } [\beta^*(\lambda)]_\text{i} = 0, \end{cases} \quad i = 1, \ldots, p, \tag{4}$$

where $[\cdot]_k$ denotes the kth component. In view of the KKT condition in (4), we have

$$|\mathbf{x}_i^T(\theta^*(\lambda))^T| < 1 \Rightarrow [\beta^*(\lambda)]_i = 0 \Rightarrow \mathbf{x}_i \text{ is an inactive feature.} \tag{R1}$$

In other words, we can potentially make use of (R1) to identify the inactive features for the Lasso problem. However, since $\theta^*(\lambda)$ is generally unknown, we can not directly apply (R1) to identify the inactive features. Inspired by the SAFE rules (El Ghaoui et al., 2012), we can first estimate a region $\mathbf{\Theta}$ which contains $\theta^*(\lambda'')$. Then, (R1) can be relaxed as follows:

$$\sup_{\theta \in \mathbf{\Theta}} |\mathbf{x}_i^T\theta| < 1 \Rightarrow [\beta^*(\lambda)]_i = 0 \Rightarrow \mathbf{x}_i \text{ is an inactive feature.} \tag{R1'}$$

Clearly, as long as we can find a region $\mathbf{\Theta}$ which contains $\theta^*(\lambda)$, (R1') will lead to a screening rule to detect the inactive features for the Lasso problem. Moreover, in view of (R1) and (R1'), we can see that the smaller the region $\mathbf{\Theta}$ is, the more accurate the estimation of $\theta^*(\lambda)$ is. As a result, more inactive features can be identified by the resulting screening rules.

The dual problem has interesting geometric interpretations. By a closer look at the dual problem (2), we can observe that the dual optimal solution is the feasible point which is closest to $\mathbf{y}/\lambda$. For notational convenience, let the feasible set of problem (2) be F. Clearly, F is the intersection of $2p$ closed half-spaces, and thus a closed and convex polytope. (Notice that, F is also nonempty since $0 \in F$.) In other words, $\theta^*(\lambda)$ is the projection of $\mathbf{y}/\lambda$ onto

the polytope F. Mathematically, for an arbitrary vector $\mathbf{w}$ and a convex set C in a Hilbert space $\mathcal{H}$, let us define the projection operator as

$$P_C(\mathbf{w}) = \underset{\mathbf{u}\in C}{\operatorname{argmin}} \|\mathbf{u} - \mathbf{w}\|_2. \tag{5}$$

Then, the dual optimal solution $\theta^*(\lambda)$ can be expressed by

$$\theta^*(\lambda) = P_F(\mathbf{y}/\lambda) = \underset{\theta\in F}{\operatorname{argmin}} \left\|\theta - \frac{\mathbf{y}}{\lambda}\right\|_2. \tag{6}$$

Indeed, the nice property of problem (2) illustrated by (6) leads to many interesting results. For example, it is easy to see that $\mathbf{y}/\lambda$ would be an *interior point* of F when λ is large enough. If this is the case, we immediately have the following assertions: 1) $\mathbf{y}/\lambda$ is an interior point of F implies that none of the constraints of problem (2) would be *active* on $\mathbf{y}/\lambda$, i.e., $|\mathbf{x}_i^T(\mathbf{y}/(\lambda)|) < 1$ for all $i = 1, \ldots, p$; 2) $\theta^*(\lambda)$ is an interior point of F as well since $\theta^*(\lambda) = P_F(\mathbf{y}/\lambda) = \mathbf{y}/\lambda$ by (6) and the fact $\mathbf{y}/\lambda \in F$. Combining the results in 1) and 2), it is easy to see that $|\mathbf{x}_i^T\theta^*(\lambda)| < 1$ for all $i = 1, \ldots, p$. By (R1), we can conclude that $\beta^*(\lambda) = 0$, under the assumption that λ is large enough.

The above analysis may naturally lead to a question: does there exist a specific parameter value $\lambda_{\max}$ such that the optimal solution of problem (1) is 0 whenever $\lambda > \lambda_{\max}$? The answer is affirmative. Indeed, let us define

$$\lambda_{\max} = \max_i |\mathbf{x}_i^T\mathbf{y}|. \tag{7}$$

It is well known (Tibshirani et al., 2012) that $\lambda_{\max}$ defined by (7) is the smallest parameter such that problem (1) has a trivial solution, i.e.,

$$\beta^*(\lambda) = 0, \ \ \forall\, \lambda \in [\lambda_{\max}, \infty). \tag{8}$$

Combining the results in (8) and (3), we immediately have

$$\theta^*(\lambda) = \frac{\mathbf{y}}{\lambda}, \ \ \forall\, \lambda \in [\lambda_{\max}, \infty). \tag{9}$$

Therefore, through out the rest of this paper, we will focus on the cases with $\lambda \in (0, \lambda_{\max})$.

In the subsequent sections, we will follow (R1') to develop our screening rules. More specifically, the derivation of the proposed screening rules can be divided into the following three steps:

1. We first estimate a region $\mathbf{\Theta}$ which contains the dual optimal solution $\theta^*(\lambda)$.
2. We solve the maximization problem in (R1'), i.e., $\sup_{\theta\in\mathbf{\Theta}} |\mathbf{x}_i^T\theta|$.
3. By plugging in the upper bound we find in 2, it is straightforward to develop the screening rule based on (R1').

The geometric property of the dual problem illustrated by (6) serves as a fundamentally important role in developing our DPP and EDPP screening rules.

2.2 Fundamental Screening Rules via Dual Polytope Projections (DPP)

In this Section, we propose the so called DPP screening rules for discarding the inactive features for Lasso. As the name indicates, the idea of DPP heavily relies on the properties of projection operators, e.g., the *nonexpansiveness* (Bertsekas, 2003). We will follow the three steps stated in Section 2.1 to develop the DPP screening rules.

First, we need to find a region $\mathbf{\Theta}$ which contains the dual optimal solution $\theta^*(\lambda)$. Indeed, the result in (9) provides us an important clue. That is, we may be able to estimate a possible region for $\theta^*(\lambda)$ in terms of a known $\theta^*(\lambda_0)$ with $\lambda < \lambda_0$. Notice that, we can always set $\lambda_0 = \lambda_{\max}$ and make use of the fact that $\theta^*(\lambda_{\max}) = \mathbf{y}/\lambda_{\max}$ implied by (9). Another key ingredient comes from (6), i.e., the dual optimal solution $\theta^*(\lambda)$ is the projection of $\mathbf{y}/\lambda$ onto the feasible set F, which is nonempty closed and convex. A nice property of the projection operators defined in a Hilbert space with respect to a nonempty closed and convex set is the so called *nonexpansiveness*. For convenience, we restate the definition of nonexpansiveness in the following theorem.

Theorem 1 *Let C be a nonempty closed convex subset of a Hilbert space $\mathcal{H}$. Then the projection operator defined in (5) is continuous and nonexpansive, i.e.,*

$$\|P_C(\mathbf{w}_2) - P_C(\mathbf{w}_1)\|_2 \leq \|\mathbf{w}_2 - \mathbf{w}_1\|_2, \ \forall \mathbf{w}_2, \mathbf{w}_1 \in \mathcal{H}. \tag{10}$$

In view of (6), a direct application of Theorem 1 leads to the following result:

Theorem 2 *For the Lasso problem, let $\lambda, \lambda_0 > 0$ be two regularization parameters. Then,*

$$\|\theta^*(\lambda) - \theta^*(\lambda_0)\|_2 \leq \left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \|\mathbf{y}\|_2. \tag{11}$$

For notational convenience, let a ball centered at $\mathbf{c}$ with radius ρ be denoted by $B(\mathbf{c}, \rho)$. Theorem 2 actually implies that the dual optimal solution must be inside a ball centered at $\theta^*(\lambda_0)$ with radius $|1/\lambda - 1/\lambda_0| \, \|\mathbf{y}\|_2$, i.e.,

$$\theta^*(\lambda) \in B\left(\theta^*(\lambda_0), \left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \|\mathbf{y}\|_2\right). \tag{12}$$

We thus complete the first step for developing DPP. Because it is easy to find the upper bound of a linear functional over a ball, we combine the remaining two steps as follows.

Theorem 3 *For the Lasso problem, assume that the dual optimum at λ_0, i.e., $\theta^*(\lambda_0)$, is known. Let λ be a positive value different from λ_0. Then $[\beta^*(\lambda)]_i = 0$ if*

$$\left|\mathbf{x}_i^T \theta^*(\lambda)\right| < 1 - \|\mathbf{x}_i\|_2 \|\mathbf{y}\|_2 \left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right|.$$

Proof The dual optimal solution $\theta^*(\lambda)$ is estimated to be inside the ball given by (12). To simplify notations, let $\mathbf{c} = \theta^*(\lambda_0)$ and $\rho = |1/\lambda - 1/\lambda_0| \, \|\mathbf{y}\|_2$. To develop a screening rule based on (R1'), we need to solve the optimization problem: $\sup_{\theta \in B(\mathbf{c}, \rho)} |\mathbf{x}_i^T \theta|$.

Indeed, for any $\theta \in B(\mathbf{c}, \rho)$, it can be expressed by:

$$\theta = \theta^*(\lambda_0) + \mathbf{v}, \quad \|\mathbf{v}\|_2 \leq \rho.$$

Therefore, the optimization problem can be easily solved as follows:

$$\sup_{\theta \in B(\mathbf{c},\rho)} \left|\mathbf{x}_i^T \theta\right| = \sup_{\|\mathbf{v}\|_2 \leq \rho} \left|\mathbf{x}_i^T \left(\theta^*(\lambda_0) + \mathbf{v}\right)\right| = \left|\mathbf{x}_i^T \theta^*(\lambda_0)\right| + \rho \|\mathbf{x}_i\|_2. \tag{13}$$

By plugging the upper bound in (13) to (R1'), we obtain the statement in Theorem 3, which completes the proof. ■

Theorem 3 implies that we can develop applicable screening rules for Lasso as long as the dual optimal solution $\theta^*(\cdot)$ is known for a certain parameter value λ_0. By simply setting $\lambda_0 = \lambda_{\max}$ and noting that $\theta^*(\lambda_{\max}) = \mathbf{y}/\lambda_{\max}$ from (9), Theorem 3 immediately leads to the following result.

Corollary 4 Basic DPP: *For the Lasso problem (1), let $\lambda_{\max} = \max_i |\mathbf{x}_i^T \mathbf{y}|$. If $\lambda \geq \lambda_{\max}$, then $[\beta^*]_i = 0, \forall i \in \mathcal{I}$. Otherwise, $[\beta^*(\lambda)]_i = 0$ if*

$$\left|\mathbf{x}_i^T \frac{\mathbf{y}}{\lambda_{\max}}\right| < 1 - \left(\frac{1}{\lambda} - \frac{1}{\lambda_{\max}}\right) \|\mathbf{x}_i\|_2 \|\mathbf{y}\|_2.$$

Remark 5 *Notice that, DPP is not the same as ST1 Xiang et al. (2011) and SAFE El Ghaoui et al. (2012), which discards the ith feature if*

$$|\mathbf{x}_i^T \mathbf{y}| < \lambda - \|\mathbf{x}_i\|_2 \|\mathbf{y}\|_2 \frac{\lambda_{\max} - \lambda}{\lambda_{\max}}. \tag{14}$$

From the perspective of the sphere test, the radius of ST1/SAFE and DPP are the same. But the centers of ST1 and DPP are $\mathbf{y}/\lambda$ and $\mathbf{y}/\lambda_{\max}$ respectively, which leads to different formulas, i.e., (14) and Corollary 4.

In real applications, the optimal parameter value of λ is generally unknown and needs to be estimated. To determine an appropriate value of λ, commonly used approaches such as cross validation and stability selection involve solving the Lasso problem over a grid of tuning parameters $\lambda_1 > \lambda_2 > \ldots > \lambda_{\mathcal{K}}$, which is very time consuming. Motivated by the ideas of Tibshirani et al. (2012) and El Ghaoui et al. (2012), we develop a sequential version of DPP rules. We first apply the DPP screening rule in Corollary 4 to discard inactive features for the Lasso problem (1) with parameter being λ_1. After solving the reduced optimization problem for λ_1, we obtain the exact solution $\beta^*(\lambda_1)$. Hence by (3), we can find $\theta^*(\lambda_1)$. According to Theorem 3, once we know the optimal dual solution $\theta^*(\lambda_1)$, we can construct a new screening rule by setting $\lambda_0 = \lambda_1$ to identify inactive features for problem (1) with parameter being λ_2. By repeating the above process, we obtain the sequential version of the DPP rule as in the following corollary.

Corollary 6 Sequential DPP: *For the Lasso problem (1), suppose we are given a sequence of parameter values $\lambda_{\max} = \lambda_0 > \lambda_1 > \ldots > \lambda_m$. Then for any integer $0 \leq k < m$, we have $[\beta^*(\lambda_{k+1})]_i = 0$ if $\beta^*(\lambda_k)$ is known and the following holds:*

$$\left|\mathbf{x}_i^T \frac{\mathbf{y} - \mathbf{X}\beta^*(\lambda_k)}{\lambda_k}\right| < 1 - \left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k}\right) \|\mathbf{x}_i\|_2 \|\mathbf{y}\|_2.$$

Remark 7 *From Corollaries 4 and 6, we can see that both of the DPP and sequential DPP rules discard the inactive features for the Lasso problem with a smaller parameter value by assuming a known dual optimal solution at a larger parameter value. This is in fact a standard way to construct screening rules for Lasso (Tibshirani et al., 2012; El Ghaoui et al., 2012; Xiang et al., 2011).*

Remark 8 *For illustration purpose, we present both the basic and sequential version of the DPP screening rules. However, it is easy to see that the basic DPP rule can be easily derived from its sequential version by simply setting $\lambda_k = \lambda_{\max}$ and $\lambda_{k+1} = \lambda$. Therefore, in this paper, we will focus on the development and evaluation of the sequential version of the proposed screening rules. To avoid any confusions, DPP and EDPP all refer to the corresponding sequential versions.*

2.3 Enhanced DPP Rules for Lasso

In this section, we further improve the DPP rules presented in Section 2.2 by a more careful analysis of the projection operators. Indeed, from the three steps by which we develop the DPP rules, we can see that the first step is a key. In other words, the estimation of the dual optimal solution serves as a fundamentally important role in developing the DPP rules. Moreover, (R1') implies that the more accurate the estimation is, the more effective the resulting screening rule is in discarding the inactive features. The estimation of the dual optimal solution in DPP rules is in fact a direct consequence of the nonexpansiveness of the projection operators. Therefore, in order to improve the performance of the DPP rules in discarding the inactive features, we propose two different approaches to find more accurate estimations of the dual optimal solution. These two approaches are presented in detail in Sections 2.3.1 and 2.3.2 respectively. By combining the ideas of these two approaches, we can further improve the estimation of the dual optimal solution. Based on this estimation, we develop the enhanced DPP rules (EDPP) in Section 2.3.3. Again, we will follow the three steps in Section 2.1 to develop the proposed screening rules.

2.3.1 Improving the DPP rules via Projections of Rays

In the DPP screening rules, the dual optimal solution $\theta^*(\lambda)$ is estimated to be inside the ball $B\left(\theta^*(\lambda_0), |1/\lambda - 1/\lambda_0|\|\mathbf{y}\|_2\right)$ with $\theta^*(\lambda_0)$ given. In this section, we show that $\theta^*(\lambda)$ lies inside a ball centered at $\theta^*(\lambda_0)$ with a smaller radius.

Indeed, it is well known that the projection of an arbitrary point onto a nonempty closed convex set C in a Hilbert space $\mathcal{H}$ always exists and is unique (Bauschke and Combettes, 2011). However, the converse is not true, i.e., there may exist $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{H}$ such that $\mathbf{w}_1 \neq \mathbf{w}_2$ and $P_C(\mathbf{w}_1) = P_C(\mathbf{w}_2)$. In fact, it is known that the following result holds:

Lemma 9 *(Bauschke and Combettes, 2011) Let C be a nonempty closed convex subset of a Hilbert space $\mathcal{H}$. For a point $\mathbf{w} \in \mathcal{H}$, let $\mathbf{w}(t) = P_C(\mathbf{w}) + t(\mathbf{w} - P_C(\mathbf{w}))$. Then, the projection of the point $\mathbf{w}(t)$ is $P_C(\mathbf{w})$ for all $t \geq 0$, i.e.,*

$$P_C(\mathbf{w}(t)) = P_C(\mathbf{w}), \forall t \geq 0.$$

Clearly, when $\mathbf{w} \neq P_C(\mathbf{w})$, i.e., $\mathbf{w} \notin C$, $\mathbf{w}(t)$ with $t \geq 0$ is the ray starting from $P_C(\mathbf{w})$ and pointing in the same direction as $\mathbf{w} - P_C(\mathbf{w})$. By Lemma 9, we know that the projection

of the ray $\mathbf{w}(t)$ with $t \geq 0$ onto the set C is a single point $P_C(\mathbf{w})$. [When $\mathbf{w} = P_C(\mathbf{w})$, i.e., $\mathbf{w} \in C$, $\mathbf{w}(t)$ with $t \geq 0$ becomes a single point and the statement in Lemma 9 is trivial.]

By making use of Lemma 9 and the nonexpansiveness of the projection operators, we can improve the estimation of the dual optimal solution in DPP [please refer to Theorem 2 and (12)]. More specifically, we have the following result:

Theorem 10 *For the Lasso problem, suppose that the dual optimal solution $\theta^*(\cdot)$ at $\lambda_0 \in (0, \lambda_{\max}]$ is known. For any $\lambda \in (0, \lambda_0]$, let us define*

$$\mathbf{v}_1(\lambda_0) = \begin{cases} \dfrac{\mathbf{y}}{\lambda_0} - \theta^*(\lambda_0), & \text{if } \lambda_0 \in (0, \lambda_{\max}), \\ \text{sign}(\mathbf{x}_*^T \mathbf{y})\mathbf{x}_*, & \text{if } \lambda_0 = \lambda_{\max}, \end{cases} \quad \text{where } \mathbf{x}_* = \text{argmax}_{\mathbf{x}_i} |\mathbf{x}_i^T \mathbf{y}|, \tag{15}$$

$$\mathbf{v}_2(\lambda, \lambda_0) = \frac{\mathbf{y}}{\lambda} - \theta^*(\lambda_0), \tag{16}$$

$$\mathbf{v}_2^{\perp}(\lambda, \lambda_0) = \mathbf{v}_2(\lambda, \lambda_0) - \frac{\langle \mathbf{v}_1(\lambda_0), \mathbf{v}_2(\lambda, \lambda_0) \rangle}{\|\mathbf{v}_1(\lambda_0)\|_2^2} \mathbf{v}_1(\lambda_0). \tag{17}$$

Then, the dual optimal solution $\theta^*(\lambda)$ can be estimated as follows:

$$\theta^*(\lambda) \in B\left(\theta^*(\lambda_0), \|\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\|_2\right) \subseteq B\left(\theta^*(\lambda_0), \left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \|\mathbf{y}\|_2\right).$$

Proof By making use of Lemma 9, we present the proof of the statement for the cases with $\lambda_0 \in (0, \lambda_{\max})$. We postpone the proof of the statement for the case with $\lambda_0 = \lambda_{\max}$ after we introduce more general technical results.

In view of the assumption $\lambda_0 \in (0, \lambda_{\max})$, it is easy to see that

$$\frac{\mathbf{y}}{\lambda_0} \notin F \Rightarrow \frac{\mathbf{y}}{\lambda_0} \neq P_F\left(\frac{\mathbf{y}}{\lambda_0}\right) = \theta^*(\lambda_0) \Rightarrow \frac{\mathbf{y}}{\lambda_0} - \theta^*(\lambda_0) \neq 0. \tag{18}$$

For each $\lambda_0 \in (0, \lambda_{\max})$, let us define

$$\theta_{\lambda_0}(t) = \theta^*(\lambda_0) + t\mathbf{v}_1(\lambda_0) = \theta^*(\lambda_0) + t\left(\frac{\mathbf{y}}{\lambda_0} - \theta^*(\lambda_0)\right), \quad t \geq 0. \tag{19}$$

By the result in (18), we can see that $\theta_{\lambda_0}(\cdot)$ defined by (19) is a ray which starts at $\theta^*(\lambda_0)$ and points in the same direction as $\mathbf{y}/\lambda_0 - \theta^*(\lambda_0)$. In view of (6), a direct application of Lemma 9 leads to that:

$$P_F(\theta_{\lambda_0}(t)) = \theta^*(\lambda_0), \quad \forall\, t \geq 0. \tag{20}$$

By applying Theorem 1 again, we have

$$\begin{aligned} \|\theta^*(\lambda) - \theta^*(\lambda_0)\|_2 &= \left\| P_F\left(\frac{\mathbf{y}}{\lambda}\right) - P_F(\theta_{\lambda_0}(t)) \right\|_2 \\ &\leq \left\| \frac{\mathbf{y}}{\lambda} - \theta_{\lambda_0}(t) \right\|_2 = \left\| t\left(\frac{\mathbf{y}}{\lambda_0} - \theta^*(\lambda_0)\right) - \left(\frac{\mathbf{y}}{\lambda} - \theta^*(\lambda_0)\right) \right\|_2 \\ &= \|t\mathbf{v}_1(\lambda_0) - \mathbf{v}_2(\lambda, \lambda_0)\|_2, \quad \forall\, t \geq 0. \end{aligned} \tag{21}$$

Because the inequality in (21) holds for all $t \geq 0$, it is easy to see that

$$\begin{aligned} \|\theta^*(\lambda) - \theta^*(\lambda_0)\|_2 &\leq \min_{t\geq 0} \ \|t\mathbf{v}_1(\lambda_0) - \mathbf{v}_2(\lambda, \lambda_0)\|_2 \\ &= \begin{cases} \|\mathbf{v}_2(\lambda, \lambda_0)\|_2, & \text{if } \langle \mathbf{v}_1(\lambda_0), \mathbf{v}_2(\lambda, \lambda_0)\rangle < 0, \\ \left\|\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\right\|_2, & \text{otherwise.} \end{cases} \end{aligned} \tag{22}$$

The inequality in (22) implies that, to prove the first half of the statement, i.e., $\theta^*(\lambda) \in B(\theta^*(\lambda_0), \|\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\|_2)$, we only need to show that $\langle \mathbf{v}_1(\lambda_0), \mathbf{v}_2(\lambda, \lambda_0)\rangle \geq 0$.

Indeed, it is easy to see that $0 \in F$. Therefore, in view of (20), the distance between $\theta_{\lambda_0}(t)$ and $\theta^*(\lambda_0)$ must be shorter than the one between $\theta_{\lambda_0}(t)$ and 0 for all $t \geq 0$, i.e.,

$$\begin{aligned} &\|\theta_{\lambda_0}(t) - \theta^*(\lambda_0)\|_2^2 \leq \|\theta_{\lambda_0}(t) - 0\|_2^2 \\ \Rightarrow\ & 0 \leq \|\theta^*(\lambda_0)\|_2^2 + 2t\left(\left\langle \theta^*(\lambda_0), \frac{\mathbf{y}}{\lambda_0}\right\rangle - \|\theta^*(\lambda_0)\|_2^2\right), \ \ \forall\, t \geq 0. \end{aligned} \tag{23}$$

Since the inequality in (23) holds for all $t \geq 0$, we can conclude that:

$$\left\langle \theta^*(\lambda_0), \frac{\mathbf{y}}{\lambda_0}\right\rangle - \|\theta^*(\lambda_0)\|_2^2 \geq 0 \Rightarrow \frac{\|\mathbf{y}\|_2}{\lambda_0} \geq \|\theta^*(\lambda_0)\|_2. \tag{24}$$

Therefore, we can see that:

$$\begin{aligned} \langle \mathbf{v}_1(\lambda_0), \mathbf{v}_2(\lambda, \lambda_0)\rangle &= \left\langle \frac{\mathbf{y}}{\lambda_0} - \theta^*(\lambda_0), \frac{\mathbf{y}}{\lambda} - \frac{\mathbf{y}}{\lambda_0} + \frac{\mathbf{y}}{\lambda_0} - \theta^*(\lambda_0)\right\rangle \\ &\geq \left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right)\left\langle \frac{\mathbf{y}}{\lambda_0} - \theta^*(\lambda_0), \mathbf{y}\right\rangle \\ &= \left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right)\left(\frac{\|\mathbf{y}\|_2^2}{\lambda_0} - \langle \theta^*(\lambda_0), \mathbf{y}\rangle\right) \\ &\geq \left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right)\left(\frac{\|\mathbf{y}\|_2^2}{\lambda_0} - \|\theta^*(\lambda_0)\|_2\|\mathbf{y}\|_2\right) \geq 0. \end{aligned} \tag{25}$$

The last inequality in (25) is due to the result in (24).

Clearly, in view of (22) and (25), we can see that the first half of the statement holds, i.e., $\theta^*(\lambda) \in B(\theta^*(\lambda_0), \|\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\|_2)$. The second half of the statement, i.e., $B(\theta^*(\lambda_0), \|\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\|_2) \subseteq B(\theta^*(\lambda_0), |1/\lambda - 1/\lambda_0|\|\mathbf{y}\|_2)$, can be easily obtained by noting that the inequality in (21) reduces to the one in (12) when $t = 1$. This completes the proof of the statement with $\lambda_0 \in (0, \lambda_{\max})$. ■

Before we present the proof of Theorem 10 for the case with $\lambda_0 = \lambda_{\max}$, let us briefly review some technical results from convex analysis first.

Definition 11 *(Ruszczyński, 2006) Let C be a nonempty closed convex subset of a Hilbert space $\mathcal{H}$ and $\mathbf{w} \in C$. The set*

$$N_C(\mathbf{w}) := \{\mathbf{v} : \langle \mathbf{v}, \mathbf{u} - \mathbf{w}\rangle \leq 0, \forall \mathbf{u} \in C\}$$

is called the normal cone to C at $\mathbf{w}$.

In terms of the normal cones, the following theorem provides an elegant and useful characterization of the projections onto nonempty closed convex subsets of a Hilbert space.

Theorem 12 *(Bauschke and Combettes, 2011) Let C be a nonempty closed convex subset of a Hilbert space $\mathcal{H}$. Then, for every $\mathbf{w} \in \mathcal{H}$ and $\mathbf{w}_0 \in C$, $\mathbf{w}_0$ is the projection of $\mathbf{w}$ onto C if and only if $\mathbf{w} - \mathbf{w}_0 \in N_C(\mathbf{w}_0)$, i.e.,*

$$\mathbf{w}_0 = P_C(\mathbf{w}) \Leftrightarrow \langle \mathbf{w} - \mathbf{w}_0, \mathbf{u} - \mathbf{w}_0 \rangle \leq 0, \forall \mathbf{u} \in C.$$

In view of the proof of Theorem 10, we can see that (20) is a key step. When $\lambda_0 = \lambda_{\max}$, similar to (19), let us define

$$\theta_{\lambda_{\max}}(t) = \theta^*(\lambda_{\max}) + t\mathbf{v}_1(\lambda_{\max}), \ \ \forall\, t \geq 0. \tag{26}$$

By Theorem 12, the following lemma shows that (20) also holds for $\lambda_0 = \lambda_{\max}$.

Lemma 13 *For the Lasso problem, let $\mathbf{v}_1(\cdot)$ and $\theta_{\lambda_{\max}}(\cdot)$ be given by (15) and (26), then the following result holds:*

$$P_F(\theta_{\lambda_{\max}}(t)) = \theta^*(\lambda_{\max}), \ \forall\, t \geq 0. \tag{27}$$

Proof To prove the statement, Theorem 12 implies that we only need to show:

$$\langle \mathbf{v}_1(\lambda_{\max}), \theta - \theta^*(\lambda_{\max}) \rangle \leq 0, \ \ \forall\, \theta \in F. \tag{28}$$

Recall that $\mathbf{v}_1(\lambda_{\max}) = \text{sign}(\mathbf{x}_*^T\mathbf{y})\mathbf{x}_*$, $\mathbf{x}_* = \text{argmax}_{\mathbf{x}_i}|\mathbf{x}_i^T\mathbf{y}|$ from (15), and $\theta^*(\lambda_{\max}) = \mathbf{y}/\lambda_{\max}$ from (9). It is easy to see that

$$\langle \mathbf{v}_1(\lambda_{\max}), \theta^*(\lambda_{\max}) \rangle = \left\langle \text{sign}(\mathbf{x}_*^T\mathbf{y})\mathbf{x}_*, \frac{\mathbf{y}}{\lambda_{\max}} \right\rangle = \frac{|\mathbf{x}_*^T\mathbf{y}|}{\lambda_{\max}} = 1. \tag{29}$$

Moreover, assume θ is an arbitrary point of F. Then, we have $|\langle \mathbf{x}_*, \theta \rangle| \leq 1$, and thus

$$\langle \mathbf{v}_1(\lambda_{\max}), \theta \rangle = \langle \text{sign}(\mathbf{x}_*^T\mathbf{y})\mathbf{x}_*, \theta \rangle \leq |\langle \mathbf{x}_*, \theta \rangle| \leq 1. \tag{30}$$

Therefore, the inequality in (28) easily follows by combing the results in (29) and (30), which completes the proof. ■

We are now ready to give the proof of Theorem 10 for the case with $\lambda_0 = \lambda_{\max}$.

Proof In view of Theorem 1 and Lemma 13, we have

$$\begin{aligned} \|\theta^*(\lambda) - \theta^*(\lambda_{\max})\|_2 &= \left\| P_F\left(\frac{\mathbf{y}}{\lambda}\right) - P_F(\theta_{\lambda_{\max}}(t)) \right\|_2 \\ &\leq \left\| \frac{\mathbf{y}}{\lambda} - \theta_{\lambda_{\max}}(t) \right\|_2 = \left\| t\mathbf{v}_1(\lambda_{\max}) - \left(\frac{\mathbf{y}}{\lambda} - \theta^*(\lambda_{\max})\right) \right\|_2 \\ &= \|t\mathbf{v}_1(\lambda_{\max}) - \mathbf{v}_2(\lambda, \lambda_{\max})\|_2, \ \ \forall\, t \geq 0. \end{aligned} \tag{31}$$

Because the inequality in (31) holds for all $t \geq 0$, we can see that

$$\|\theta^*(\lambda) - \theta^*(\lambda_{\max})\|_2 \leq \min_{t \geq 0} \|t\mathbf{v}_1(\lambda_{\max}) - \mathbf{v}_2(\lambda, \lambda_{\max})\|_2 \tag{32}$$
$$= \begin{cases} \|\mathbf{v}_2(\lambda, \lambda_{\max})\|_2, & \text{if } \langle \mathbf{v}_1(\lambda_{\max}), \mathbf{v}_2(\lambda, \lambda_{\max})\rangle < 0, \\ \left\|\mathbf{v}_2^{\perp}(\lambda, \lambda_{\max})\right\|_2, & \text{otherwise.} \end{cases}$$

Clearly, we only need to show that $\langle \mathbf{v}_1(\lambda_{\max}), \mathbf{v}_2(\lambda, \lambda_{\max})\rangle \geq 0$.

Indeed, Lemma 13 implies that $\mathbf{v}_1(\lambda_{\max}) \in N_F(\theta^*(\lambda_{\max}))$ [please refer to the inequality in (28)]. By noting that $0 \in F$, we have

$$\left\langle \mathbf{v}_1(\lambda_{\max}), 0 - \frac{\mathbf{y}}{\lambda_{\max}} \right\rangle \leq 0 \Rightarrow \langle \mathbf{v}_1(\lambda_{\max}), \mathbf{y}\rangle \geq 0.$$

Moreover, because $\mathbf{y}/\lambda_{\max} = \theta^*(\lambda_{\max})$, it is easy to see that

$$\langle \mathbf{v}_1(\lambda_{\max}), \mathbf{v}_2(\lambda, \lambda_{\max})\rangle = \left\langle \mathbf{v}_1(\lambda_{\max}), \frac{\mathbf{y}}{\lambda} - \frac{\mathbf{y}}{\lambda_{\max}} \right\rangle \tag{33}$$
$$= \left(\frac{1}{\lambda} - \frac{1}{\lambda_{\max}}\right) \langle \mathbf{v}_1(\lambda_{\max}), \mathbf{y}\rangle \geq 0.$$

Therefore, in view of (32) and (33), we can see that the first half of the statement holds, i.e., $\theta^*(\lambda) \in B(\theta^*(\lambda_{\max}), \|\mathbf{v}_2^{\perp}(\lambda, \lambda_{\max})\|_2)$. The second half of the statement, i.e., $B(\theta^*(\lambda_{\max}), \|\mathbf{v}_2^{\perp}(\lambda, \lambda_{\max})\|_2) \subseteq B(\theta^*(\lambda_{\max}), |1/\lambda - 1/\lambda_{\max}|\|\mathbf{y}\|_2)$, can be easily obtained by noting that the inequality in (32) reduces to the one in (12) when $t = 0$. This completes the proof of the statement with $\lambda_0 = \lambda_{\max}$. Thus, the proof of Theorem 10 is completed. ■

Theorem 10 in fact provides a more accurate estimation of the dual optimal solution than the one in DPP, i.e., $\theta^*(\lambda)$ lies inside a ball centered at $\theta^*(\lambda_0)$ with a radius $\|\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\|_2$. Based on this improved estimation and (R1'), we can develop the following screening rule to discard the inactive features for Lasso.

Theorem 14 *For the Lasso problem, assume the dual optimal solution $\theta^*(\cdot)$ at $\lambda_0 \in (0, \lambda_{\max}]$ is known. Then, for each $\lambda \in (0, \lambda_0)$, we have $[\beta^*(\lambda)]_i = 0$ if*

$$|\mathbf{x}_i^T \theta^*(\lambda_0)| < 1 - \|\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\|_2 \|\mathbf{x}_i\|_2.$$

We omit the proof of Theorem 14 since it is very similar to the one of Theorem 3. By Theorem 14, we can easily develop the following sequential screening rule.

Improvement 1: *For the Lasso problem (1), suppose we are given a sequence of parameter values $\lambda_{\max} = \lambda_0 > \lambda_1 > \ldots > \lambda_{\mathcal{K}}$. Then for any integer $0 \leq k < \mathcal{K}$, we have $[\beta^*(\lambda_{k+1})]_i = 0$ if $\beta^*(\lambda_k)$ is known and the following holds:*

$$\left| \mathbf{x}_i^T \frac{\mathbf{y} - \mathbf{X}\beta^*(\lambda_k)}{\lambda_k} \right| < 1 - \|\mathbf{v}_2^{\perp}(\lambda_{k+1}, \lambda_k)\|_2 \|\mathbf{x}_i\|_2.$$

The screening rule in Improvement 1 is developed based on (R1') and the estimation of the dual optimal solution in Theorem 10, which is more accurate than the one in DPP. Therefore, in view of (R1'), the screening rule in Improvement 1 are more effective in discarding the inactive features than the DPP rule.

2.3.2 Improving the DPP rules via Firmly Nonexpansiveness

In Section 2.3.1, we improve the estimation of the dual optimal solution in DPP by making use of the projections of properly chosen rays. (R1') implies that the resulting screening rule stated in Improvement 1 is more effective in discarding the inactive features than DPP. In this Section, we present another approach to improve the estimation of the dual optimal solution in DPP by making use of the so called *firmly nonexpansiveness* of the projections onto nonempty closed convex subset of a Hilbert space.

Theorem 15 *(Bauschke and Combettes, 2011) Let C be a nonempty closed convex subset of a Hilbert space $\mathcal{H}$. Then the projection operator defined in (5) is continuous and firmly nonexpansive. In other words, for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{H}$, we have*

$$\|P_C(\mathbf{w}_1) - P_C(\mathbf{w}_2)\|_2^2 + \|(\mathrm{Id} - P_C)(\mathbf{w}_1) - (\mathrm{Id} - P_C)(\mathbf{w}_2)\|_2^2 \leq \|\mathbf{w}_1 - \mathbf{w}_2\|_2^2, \tag{34}$$

where Id *is the identity operator.*

In view of the inequalities in (34) and (10), it is easy to see that firmly nonexpansiveness implies nonexpansiveness. But the converse is not true. Therefore, firmly nonexpansiveness of the projection operators is a stronger property than the nonexpansiveness. A direct application of Theorem 15 leads to the following result.

Theorem 16 *For the Lasso problem, let $\lambda, \lambda_0 > 0$ be two parameter values. Then*

$$\theta^*(\lambda) \in B\left(\theta^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right)\mathbf{y}, \frac{1}{2}\left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \|\mathbf{y}\|_2\right) \subset B\left(\theta^*(\lambda_0), \left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \|\mathbf{y}\|_2\right). \tag{35}$$

Proof In view of (6) and the firmly nonexpansiveness in (34), we have

$$\begin{aligned} & \|\theta^*(\lambda) - \theta^*(\lambda_0)\|_2^2 + \left\|\left(\frac{\mathbf{y}}{\lambda} - \theta^*(\lambda)\right) - \left(\frac{\mathbf{y}}{\lambda_0} - \theta^*(\lambda_0)\right)\right\|_2^2 \leq \left\|\frac{\mathbf{y}}{\lambda} - \frac{\mathbf{y}}{\lambda_0}\right\|_2^2 \qquad (36) \\ \Leftrightarrow\ & \|\theta^*(\lambda) - \theta^*(\lambda_0)\|_2^2 \leq \left\langle \theta^*(\lambda) - \theta^*(\lambda_0), \frac{\mathbf{y}}{\lambda} - \frac{\mathbf{y}}{\lambda_0}\right\rangle \\ \Leftrightarrow\ & \left\|\theta^*(\lambda) - \left(\theta^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right)\mathbf{y}\right)\right\|_2 \leq \frac{1}{2}\left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \|\mathbf{y}\|_2, \end{aligned}$$

which completes the proof of the first half of the statement. The second half of the statement is trivial by noting that the first inequality in (36) (firmly nonexpansiveness) implies the inequality in (11) (nonexpansiveness) but not vice versa. Indeed, it is easy to see that the ball in the middle of (35) is inside the right one and has only a half radius. ■

Clearly, Theorem 16 provides a more accurate estimation of the dual optimal solution than the one in DPP, i.e., the dual optimal solution must be inside a ball which is a subset of the one in DPP and has only a half radius. Again, based on the estimation in Theorem 16 and (R1'), we have the following result.

Theorem 17 *For the Lasso problem, assume that the dual optimal solution $\theta^*(\cdot)$ at $\lambda_0 \in (0, \lambda_{\max}]$ is known. Then, for each $\lambda \in (0, \lambda_0)$, we have $[\beta^*(\lambda)]_i = 0$ if*

$$\left|\mathbf{x}_i^T\left(\theta^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right)\mathbf{y}\right)\right| < 1 - \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right)\|\mathbf{y}\|_2\|\mathbf{x}_i\|_2.$$

We omit the proof of Theorem 17 since it is very similar to the proof of Theorem 3. A direct application of Theorem 17 leads to the following sequential screening rule.

Improvement 2: *For the Lasso problem (1), suppose that we are given a sequence of parameter values $\lambda_{\max} = \lambda_0 > \lambda_1 > \ldots > \lambda_{\mathcal{K}}$. Then for any integer $0 \leq k < \mathcal{K}$, we have $[\beta^*(\lambda_{k+1})]_i = 0$ if $\beta^*(\lambda_k)$ is known and the following holds:*

$$\left|\mathbf{x}_i^T\left(\frac{\mathbf{y} - \mathbf{X}\beta^*(\lambda_k)}{\lambda_k} + \frac{1}{2}\left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k}\right)\mathbf{y}\right)\right| < 1 - \frac{1}{2}\left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k}\right)\|\mathbf{y}\|_2\|\mathbf{x}_i\|_2.$$

Because the screening rule in Improvement 2 is developed based on (R1') and the estimation in Theorem 16, it is easy to see that Improvement 2 is more effective in discarding the inactive features than DPP.

2.3.3 The Proposed Enhanced DPP Rules

In Sections 2.3.1 and 2.3.2, we present two different approaches to improve the estimation of the dual optimal solution in DPP. In view of (R1'), we can see that the resulting screening rules, i.e., Improvements 1 and 2, are more effective in discarding the inactive features than DPP. In this section, we give a more accurate estimation of the dual optimal solution than the ones in Theorems 10 and 16 by combining the aforementioned two approaches together. The resulting screening rule for Lasso is the so called enhanced DPP rule (EDPP). Again, (R1') implies that EDPP is more effective in discarding the inactive features than the screening rules in Improvements 1 and 2. We also present several experiments to demonstrate that EDPP is able to identify more inactive features than the screening rules in Improvements 1 and 2. Therefore, in the subsequent sections, we will focus on the generalizations and evaluations of EDPP.

To develop the EDPP rules, we still follow the three steps in Section 2.1. Indeed, by combining the two approaches proposed in Sections 2.3.1 and 2.3.2, we can further improve the estimation of the dual optimal solution in the following theorem.

Theorem 18 *For the Lasso problem, suppose that the dual optimal solution $\theta^*(\cdot)$ at $\lambda_0 \in (0, \lambda_{\max}]$ is known, and $\forall\, \lambda \in (0, \lambda_0]$, let $\mathbf{v}_2^{\perp}(\lambda, \lambda_0)$ be given by (17). Then, we have*

$$\left\|\theta^*(\lambda) - \left(\theta^*(\lambda_0) + \frac{1}{2}\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\right)\right\|_2 \leq \frac{1}{2}\|\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\|_2.$$

Proof Recall that $\theta_{\lambda_0}(t)$ is defined by (19) and (26). In view of (34), we have

$$\left\|P_F\left(\frac{\mathbf{y}}{\lambda}\right) - P_F\left(\theta_{\lambda_0}(t)\right)\right\|_2^2 + \left\|(\mathrm{Id} - P_F)\left(\frac{\mathbf{y}}{\lambda}\right) - (\mathrm{Id} - P_F)\left(\theta_{\lambda_0}(t)\right)\right\|_2^2 \leq \left\|\frac{\mathbf{y}}{\lambda} - \theta_{\lambda_0}(t)\right\|_2^2. \tag{37}$$

By expanding the second term on the left hand side of (37) and rearranging the terms, we obtain the following equivalent form:

$$\left\| P_F\left(\frac{\mathbf{y}}{\lambda}\right) - P_F\left(\theta_{\lambda_0}(t)\right)\right\|_2^2 \leq \left\langle \frac{\mathbf{y}}{\lambda} - \theta_{\lambda_0}(t), P_F\left(\frac{\mathbf{y}}{\lambda}\right) - P_F\left(\theta_{\lambda_0}(t)\right)\right\rangle. \tag{38}$$

In view of (6), (20) and (27), the inequality in (38) can be rewritten as

$$\begin{aligned} \|\theta^*(\lambda) - \theta^*(\lambda_0)\|_2^2 &\leq \left\langle \frac{\mathbf{y}}{\lambda} - \theta_{\lambda_0}(t), \theta^*(\lambda) - \theta^*(\lambda_0)\right\rangle \\ &= \left\langle \frac{\mathbf{y}}{\lambda} - \theta^*(\lambda_0) - t\mathbf{v}_1(\lambda_0), \theta^*(\lambda) - \theta^*(\lambda_0)\right\rangle \\ &= \langle \mathbf{v}_2(\lambda, \lambda_0) - t\mathbf{v}_1(\lambda_0), \theta^*(\lambda) - \theta^*(\lambda_0)\rangle, \ \ \forall t \geq 0. \end{aligned} \tag{39}$$

[Recall that $\mathbf{v}_1(\lambda_0)$ and $\mathbf{v}_2(\lambda, \lambda_0)$ are defined by (15) and (16) respectively.] Clearly, the inequality in (39) is equivalent to

$$\left\| \theta^*(\lambda) - \left(\theta^*(\lambda_0) + \frac{1}{2}(\mathbf{v}_2(\lambda, \lambda_0) - t\mathbf{v}_1(\lambda_0))\right)\right\|_2^2 \leq \frac{1}{4}\|\mathbf{v}_2(\lambda, \lambda_0) - t\mathbf{v}_1(\lambda_0)\|_2^2, \ \ \forall t \geq 0. \tag{40}$$

The statement follows easily by minimizing the right hand side of the inequality in (40), which has been done in the proof of Theorem 10. ■

Indeed, Theorem 18 is equivalent to bounding $\theta^*(\lambda)$ in a ball as follows:

$$\theta^*(\lambda) \in B\left(\theta^*(\lambda_0) + \frac{1}{2}\mathbf{v}_2^{\perp}(\lambda, \lambda_0), \frac{1}{2}\|\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\|_2\right). \tag{41}$$

Based on this estimation and (R1'), we immediately have the following result.

Theorem 19 *For the Lasso problem, assume that the dual optimal problem $\theta^*(\cdot)$ at $\lambda_0 \in (0, \lambda_{\max}]$ is known, and $\lambda \in (0, \lambda_0]$. Then $[\beta^*(\lambda)]_i = 0$ if the following holds:*

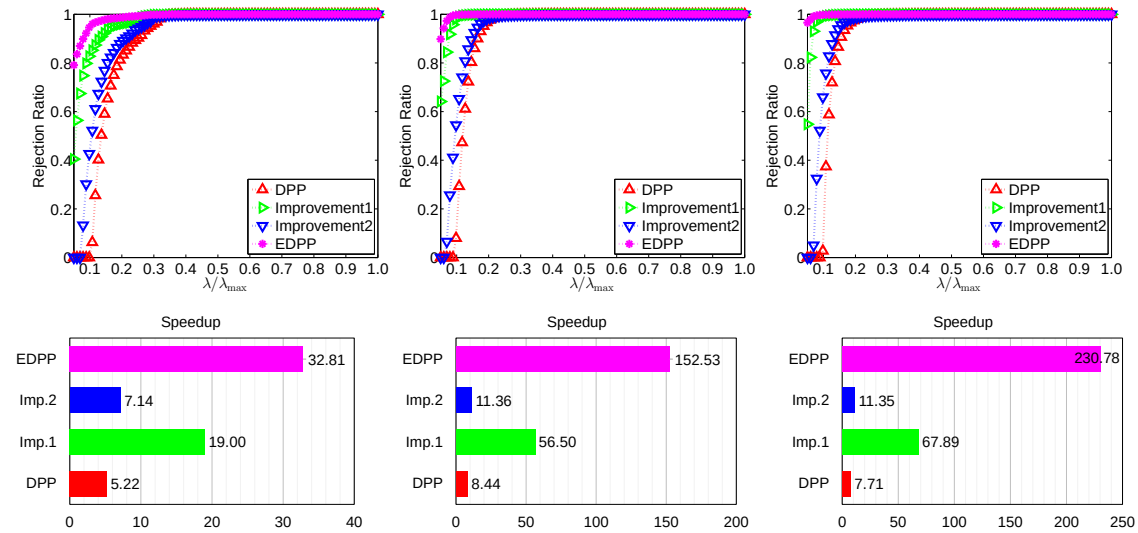
$$\left|\mathbf{x}_i^T\left(\theta^*(\lambda_0) + \frac{1}{2}\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\right)\right| < 1 - \frac{1}{2}\|\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\|_2\|\mathbf{x}_i\|_2.$$

We omit the proof of Theorem 19 since it is very similar to the one of Theorem 3. Based on Theorem 19, we can develop the EDPP rules as follows.

Corollary 20 EDPP: *For the Lasso problem, suppose that we are given a sequence of parameter values $\lambda_{\max} = \lambda_0 > \lambda_1 > \ldots > \lambda_{\mathcal{K}}$. Then for any integer $0 \leq k < \mathcal{K}$, we have $[\beta^*(\lambda_{k+1})]_i = 0$ if $\beta^*(\lambda_k)$ is known and the following holds:*

$$\left|\mathbf{x}_i^T\left(\frac{\mathbf{y} - \mathbf{X}\beta^*(\lambda_k)}{\lambda_k} + \frac{1}{2}\mathbf{v}_2^{\perp}(\lambda_{k+1}, \lambda_k)\right)\right| < 1 - \frac{1}{2}\|\mathbf{v}_2^{\perp}(\lambda_{k+1}, \lambda_k)\|_2\|\mathbf{x}_i\|_2. \tag{42}$$

It is easy to see that the ball in (41) has the smallest radius compared to the ones in Theorems 10 and 16, and thus it provides the most accurate estimation of the dual optimal solution. According to (R1'), EDPP is more effective in discarding the inactive features than DPP, Improvements 1 and 2.



(a) Prostate Cancer, $\mathbf{X} \in \mathbb{R}^{132\times 15154}$ (b) PIE, $\mathbf{X} \in \mathbb{R}^{1024\times 11553}$ (c) MNIST, $\mathbf{X} \in \mathbb{R}^{784\times 50000}$

Figure 1: Comparison of the family of DPP rules on three real data sets: Prostate Cancer digit data set (left), PIE data set (middle) and MNIST image data set (right). The first row shows the rejection ratios of DPP, Improvement 1, Improvement 2 and EDPP. The second row presents the speedup gained by these four methods.

We evaluate the performance of the family of DPP screening rules, i.e., DPP, Improvement 1, Improvement 2 and EDPP, on three real data sets: a) the Prostate Cancer (Petricoin et al., 2002); b) the PIE face image data set (Sim et al., 2003); c) the MNIST handwritten digit data set (Lecun et al., 1998). To measure the performance of the screening rules, we compute the following two quantities:

1. the *rejection ratio*, i.e., the ratio of the number of features discarded by screening rules to the actual number of zero features in the ground truth;

2. the *speedup*, i.e., the ratio of the running time of the solver with screening rules to the running time of the solver without screening.

For each data set, we run the solver with or without the screening rules to solve the Lasso problem along a sequence of 100 parameter values equally spaced on the $\lambda/\lambda_{\max}$ scale from 0.05 to 1.0. Figure 1 presents the rejection ratios and speedup by the family of DPP screening rules. Table 1 reports the running time of the solver with or without the screening rules for solving the 100 Lasso problems, as well as the time for running the screening rules.

The Prostate Cancer data set (Petricoin et al., 2002) is obtained by protein mass spectrometry. The features are indexed by time-of-flight values, which are related to the mass over charge ratios of the constituent proteins in the blood. The data set has 15154 measurements of 132 patients. 69 of the patients have prostate cancer and the rest are healthy. Therefore, the data matrix $\mathbf{X}$ is of size 132×15154, and the response vector $\mathbf{y} \in \{1, -1\}^{132}$ contains the binary labels of the patients.

Data	solver	DPP+solver	Imp.1+solver	Imp.2+solver	EDPP+solver	DPP	Imp.1	Imp.2	EDPP
Prostate Cancer	121.41	23.36	6.39	17.00	3.70	0.30	0.27	0.28	0.23
PIE	629.94	74.66	11.15	55.45	4.13	1.63	1.34	1.54	1.33
MNIST	2566.26	332.87	37.80	226.02	11.12	5.28	4.36	4.94	4.19

Table 1: Running time (in seconds) for solving the Lasso problems along a sequence of 100 tuning parameter values equally spaced on the scale of $\lambda/\lambda_{\max}$ from 0.05 to 1 by (a): the solver (Liu et al., 2009) (reported in the second column) without screening; (b): the solver combined with different screening methods (reported in the 3rd to the 6th columns). The last four columns report the total running time (in seconds) for the screening methods.

The PIE face image data set used in this experiment[1] (Cai et al., 2007) contains 11554 gray face images of 68 people, taken under different poses, illumination conditions and expressions. Each of the images has 32×32 pixels. Therefore, in each trial, we first randomly pick an image as the response $\mathbf{y} \in \mathbb{R}^{1024}$, and then use the remaining images to form the data matrix $\mathbf{X} \in \mathbb{R}^{1024 \times 11553}$. We run 100 trials and report the average performance of the screening rules.

The MNIST data set contains gray images of scanned handwritten digits, including $60,000$ for training and $10,000$ for testing. The dimension of each image is 28×28. We first randomly select 5000 images for each digit from the training set (and in total we have 50000 images) and get a data matrix $\mathbf{X} \in \mathbb{R}^{784 \times 50000}$. Then in each trial, we randomly select an image from the testing set as the response $\mathbf{y} \in \mathbb{R}^{784}$. We run 100 trials and report the average performance of the screening rules.

From Figure 1, we can see that both Improvements 1 and 2 are able to discard more inactive features than DPP, and thus lead to a higher speedup. Compared to Improvement 2, we can also observe that Improvement 1 is more effective in discarding the inactive features. For the three data sets, the second row of Figure 1 shows that Improvement 1 leads to about 20, 60, 70 times speedup respectively, which are much higher than the ones gained by Improvement 1 (roughly 10 times for all the three cases).

Moreover, the EDPP rule, which combines the ideas of both Improvements 1 and 2, is even more effective in discarding the inactive features than Improvement 1. We can see that, for all of the three data sets and most of the 100 parameter values, the rejection ratios of EDPP are very close to 100%. In other words, EDPP is able to discard almost all of the inactive features. Thus, the resulting speedup of EDPP is significantly better than the ones gained by the other three DPP rules. For the PIE and MNIST data sets, we can see that the speedup gained EDPP is about 150 and 230 times, which are two orders of magnitude. In view of Table 1, for the MNIST data set, the solver without screening needs about 2566.26 seconds to solve the 100 Lasso problems. In contrast, the solver with EDPP only needs 11.12 seconds, leading to substantial savings in the computational cost. Moreover, from

1. `http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html`

the last four columns of Table 1, we can also observe that the computational cost of the family of DPP rules are very low. Compared to that of the solver without screening, the computational cost of the family of DPP rules is negligible.

In Section 4, we will only compare the performance of EDPP against several other state-of-the-art screening rules.

3. Extensions to Group Lasso

To demonstrate the flexibility of the family of DPP rules, we extend the idea of EDPP to the group Lasso problem (Yuan and Lin, 2006) in this section. Although the Lasso and group Lasso problems are very different from each other, we will see that their dual problems share a lot of similarities. For example, both of the dual problems can be formulated as looking for projections onto nonempty closed convex subsets of a Hilbert space. Recall that, the EDPP rule for the Lasso problem is entirely based on the properties of the projection operators. Therefore, the framework of the EDPP screening rule we developed for Lasso is also applicable for the group Lasso problem. In Section 3.1, we briefly review some basics of the group Lasso problem and explore the geometric properties of its dual problem. In Section 3.2, we develop the EDPP rule for the group Lasso problem.

3.1 Basics

With the group information available, the group Lasso problem takes the form of:

$$\inf_{\beta\in\mathbb{R}^p} \frac{1}{2}\left\|\mathbf{y}-\sum\nolimits_{g=1}^{G}\mathbf{X}_g\beta_g\right\|_2^2+\lambda\sum\nolimits_{g=1}^{G}\sqrt{n_g}\|\beta_g\|_2, \tag{43}$$

where $\mathbf{X}_g\in\mathbb{R}^{N\times n_g}$ is the data matrix for the gth group and $p=\sum_{g=1}^{G}n_g$. The dual problem of (43) is (see detailed derivation in the appendix):

$$\sup_{\theta}\quad\left\{\frac{1}{2}\|\mathbf{y}\|_2^2-\frac{\lambda^2}{2}\left\|\theta-\frac{\mathbf{y}}{\lambda}\right\|_2^2:\ \|\mathbf{X}_g^T\theta\|_2\le\sqrt{n_g},\ g=1,2,\ldots,G\right\} \tag{44}$$

The KKT conditions are given by

$$\mathbf{y}=\sum\nolimits_{g=1}^{G}\mathbf{X}_g\beta_g^*(\lambda)+\lambda\theta^*(\lambda), \tag{45}$$

$$(\theta^*(\lambda))^T\mathbf{X}_g\in\begin{cases}\sqrt{n_g}\frac{\beta_g^*(\lambda)}{\|\beta_g^*(\lambda)\|_2}, & \text{if}\beta_g^*(\lambda)\neq 0,\\ \sqrt{n_g}\mathbf{u},\ \|\mathbf{u}\|_2\le 1, & \text{if}\beta_g^*(\lambda)=0.\end{cases} \tag{46}$$

for $g=1,2,\ldots,G$. Clearly, in view of (46), we can see that

$$\|(\theta^*(\lambda))^T\mathbf{X}_g\|_2<\sqrt{n_g}\Rightarrow\beta_g^*(\lambda)=0 \tag{R2}$$

However, since $\theta^*(\lambda)$ is generally unknown, (R2) is not applicable to identify the *inactive groups*, i.e., the groups which have 0 coefficients in the solution vector, for the group Lasso problem. Therefore, similar to the Lasso problem, we can first find a region $\overline{\mathbf{\Theta}}$ which contains $\theta^*(\lambda)$, and then (R2) can be relaxed as follows:

$$\sup_{\theta\in\overline{\mathbf{\Theta}}}\|(\theta)^T\mathbf{X}_g\|_2<\sqrt{n_g}\Rightarrow\beta_g^*(\lambda)=0. \tag{R2$'$}$$

Therefore, to develop screening rules for the group Lasso problem, we only need to estimate the region $\overline{\mathbf{\Theta}}$ which contains $\theta^*(\lambda)$, solve the maximization problem in (R2′), and plug it into (R2′). In other words, the three steps proposed in Section 2.1 can also be applied to develop screening rules for the group Lasso problem. Moreover, (R2′) also implies that the smaller the region $\overline{\mathbf{\Theta}}$ is, the more accurate the estimation of the dual optimal solution is. As a result, the more effective the resulting screening rule is in discarding the inactive features.

The dual problem of group Lasso has similar geometric interpretations to the one of Lasso. For notational convenience, let $\overline{F}$ be the feasible set of problem (44). Similar to the case of Lasso, problem (44)implies that the dual optimal $\theta^*(\lambda)$ is the projection of $\mathbf{y}/\lambda$ onto the feasible set $\overline{F}$, i.e.,

$$\theta^*(\lambda) = P_{\overline{F}}\left(\frac{\mathbf{y}}{\lambda}\right), \ \ \forall\, \lambda > 0. \tag{47}$$

Compared to (6), the only difference in (47) is that the feasible set $\overline{F}$ is the intersection of a set of ellipsoids, and thus not a polytope. However, similar to F, $\overline{F}$ is also a nonempty closed and convex (notice that 0 is a feasible point). Therefore, we can make use of all the aforementioned properties of the projection operators, e.g., Lemmas 9 and 13, Theorems 12 and 15, to develop screening rules for the group Lasso problem.

Moreover, similar to the case of Lasso, we also have a specific parameter value (Tibshirani et al., 2012) for the group Lasso problem, i.e.,

$$\overline{\lambda}_{\max} = \max_g \frac{\|\mathbf{X}_g^T\mathbf{y}\|_2}{\sqrt{n_g}}. \tag{48}$$

Indeed, $\overline{\lambda}_{\max}$ is the smallest parameter value such that the optimal solution of problem (43) is 0. More specifically, we have:

$$\beta^*(\lambda) = 0, \ \ \forall\, \lambda \in [\overline{\lambda}_{\max}, \infty). \tag{49}$$

Combining the result in (49) and (45), we immediately have

$$\theta^*(\lambda) = \frac{\mathbf{y}}{\lambda}, \ \ \forall\, \lambda \in [\overline{\lambda}_{\max}, \infty). \tag{50}$$

Therefore, all through the subsequent sections, we will focus on the cases with $\lambda \in (0, \overline{\lambda}_{\max})$.

3.2 Enhanced DPP rule for Group Lasso

In view of (R2′), we can see that the estimation of the dual optimal solution is the key step to develop a screening rule for the group Lasso problem. Because $\theta^*(\lambda)$ is the projection of $\mathbf{y}/\lambda$ onto the nonempty closed convex set $\overline{F}$ [please refer to (47)], we can make use of all the properties of projection operators, e.g., Lemmas 9 and 13, Theorems 12 and 15, to estimate the dual optimal solution. First, let us develop a useful technical result as follows.

Lemma 21 *For the group Lasso problem, let $\overline{\lambda}_{\max}$ be given by (48) and*

$$\mathbf{X}_* := \mathrm{argmax}_{\mathbf{X}_g} \frac{\|\mathbf{X}_g^T\mathbf{y}\|_2}{\sqrt{n_g}}. \tag{51}$$

Suppose that the dual optimal solution $\theta^*(\cdot)$ is known at $\lambda_0 \in (0, \overline{\lambda}_{\max}]$, let us define

$$\overline{\mathbf{v}}_1(\lambda_0) = \begin{cases} \dfrac{\mathbf{y}}{\lambda_0} - \theta^*(\lambda_0), & \text{if } \lambda_0 \in (0, \overline{\lambda}_{\max}), \\ \mathbf{X}_* \mathbf{X}_*^T \mathbf{y}, & \text{if } \lambda_0 = \overline{\lambda}_{\max}. \end{cases} \tag{52}$$
$$\overline{\theta}_{\lambda_0}(t) = \theta^*(\lambda_0) + t\overline{\mathbf{v}}_1(\lambda_0), \ \ t \geq 0.$$

Then, we have the following result holds

$$P_{\overline{F}}(\overline{\theta}_{\lambda_0}(t)) = \theta^*(\lambda_0), \ \forall \ t \geq 0. \tag{53}$$

Proof Let us first consider the cases with $\lambda_0 \in (0, \overline{\lambda}_{\max})$. In view of the definition of $\overline{\lambda}_{\max}$, it is easy to see that $\mathbf{y}/\lambda_0 \notin \overline{F}$. Therefore, in view of (47) and Lemma 9, the statement in (53) follows immediately.

We next consider the case with $\lambda_0 = \overline{\lambda}_{\max}$. By Theorem 12, we only need to check if

$$\overline{\mathbf{v}}_1(\overline{\lambda}_{\max}) \in N_{\overline{F}}(\theta^*(\overline{\lambda}_{\max})) \Leftrightarrow \left\langle \overline{\mathbf{v}}_1(\overline{\lambda}_{\max}), \theta - \theta^*(\overline{\lambda}_{\max}) \right\rangle \leq 0, \ \ \forall \, \theta \in \overline{F}. \tag{54}$$

Indeed, in view of (48) and (50), we can see that

$$\langle \overline{\mathbf{v}}_1(\overline{\lambda}_{\max}), \theta^*(\overline{\lambda}_{\max}) \rangle = \left\langle \mathbf{X}_* \mathbf{X}_*^T \mathbf{y}, \frac{\mathbf{y}}{\overline{\lambda}_{\max}} \right\rangle = \frac{\|\mathbf{X}_*^T \mathbf{y}\|_2^2}{\overline{\lambda}_{\max}}. \tag{55}$$

On the other hand, by (48) and (51), we can see that

$$\|\mathbf{X}_*^T \mathbf{y}\|_2 = \overline{\lambda}_{\max} \sqrt{n_*}, \tag{56}$$

where n_* is the number of columns of $\mathbf{X}_*$. By plugging (56) into (55), we have

$$\langle \overline{\mathbf{v}}_1(\overline{\lambda}_{\max}), \theta^*(\overline{\lambda}_{\max}) \rangle = \overline{\lambda}_{\max} \cdot n_*.$$

Moreover, for any feasible point $\theta \in \overline{F}$, we can see that

$$\|\mathbf{X}_*^T \theta\|_2 \leq \sqrt{n_*}. \tag{57}$$

In view of the result in (57) and (56), it is easy to see that

$$\left\langle \overline{\mathbf{v}}_1(\overline{\lambda}_{\max}), \theta \right\rangle = \left\langle \mathbf{X}_* \mathbf{X}_*^T \mathbf{y}, \theta \right\rangle = \left\langle \mathbf{X}_*^T \mathbf{y}, \mathbf{X}_*^T \theta \right\rangle \leq \|\mathbf{X}_*^T \mathbf{y}\|_2 \|\mathbf{X}_*^T \theta\|_2 = \overline{\lambda}_{\max} \cdot n_*. \tag{58}$$

Combining the result in (55) and (58), it is easy to see that the inequality in (54) holds for all $\theta \in \overline{F}$, which completes the proof. ■

By Lemma 21, we can accurately estimate the dual optimal solution of the group Lasso problem in the following theorem. It is easy to see that the result in Theorem 22 is very similar to the one in Theorem 18 for the Lasso problem.

Theorem 22 *For the group Lasso problem, suppose that the dual optimal solution $\theta^*(\cdot)$ at $\theta_0 \in (0, \overline{\lambda}_{\max}]$ is known, and $\overline{\mathbf{v}}_1(\lambda_0)$ is given by (52). For any $\lambda \in (0, \lambda_0]$, let us define*

$$\overline{\mathbf{v}}_2(\lambda, \lambda_0) = \frac{\mathbf{y}}{\lambda} - \theta^*(\lambda_0),$$

$$\overline{\mathbf{v}}_2^{\perp}(\lambda, \lambda_0) = \overline{\mathbf{v}}_2(\lambda, \lambda_0) - \frac{\langle \overline{\mathbf{v}}_1(\lambda_0), \overline{\mathbf{v}}_2(\lambda, \lambda_0)\rangle}{\|\overline{\mathbf{v}}_1(\lambda_0)\|_2^2}\overline{\mathbf{v}}_1(\lambda_0).$$

Then, the dual optimal solution $\theta^*(\lambda)$ can be estimated as follows:

$$\left\|\theta^*(\lambda) - \left(\theta^*(\lambda_0) + \frac{1}{2}\overline{\mathbf{v}}_2^{\perp}(\lambda, \lambda_0)\right)\right\|_2 \leq \frac{1}{2}\|\overline{\mathbf{v}}_2^{\perp}(\lambda, \lambda_0)\|_2.$$

We omit the proof of Theorem 22 since it is exactly the same as the one of Theorem 18. Indeed, Theorem 22 is equivalent to estimating $\theta^*(\lambda)$ in a ball as follows:

$$\theta^*(\lambda) \in B\left(\theta^*(\lambda_0) + \frac{1}{2}\overline{\mathbf{v}}_2^{\perp}(\lambda, \lambda_0), \frac{1}{2}\|\overline{\mathbf{v}}_2^{\perp}(\lambda, \lambda_0)\|_2\right). \tag{59}$$

Based on this estimation and (R2′), we immediately have the following result.

Theorem 23 *For the group Lasso problem, assume the dual optimal solution $\theta^*(\cdot)$ is known at $\lambda_0 \in (0, \overline{\lambda}_{\max}]$, and $\lambda \in (0, \lambda_0]$. Then $\beta_g^*(\lambda) = 0$ if the following holds*

$$\left\|\mathbf{X}_g^T\left(\theta^*(\lambda_0) + \frac{1}{2}\overline{\mathbf{v}}_2^{\perp}(\lambda, \lambda_0)\right)\right\|_2 < \sqrt{n_g} - \frac{1}{2}\|\overline{\mathbf{v}}_2^{\perp}(\lambda, \lambda_0)\|_2\|\mathbf{X}_g\|_2. \tag{60}$$

Proof In view of (R2′), we only need to check if

$$\left\|\mathbf{X}_g^T\theta^*(\lambda)\right\|_2 < \sqrt{n_g}.$$

To simplify notations, let

$$\mathbf{o} = \theta^*(\lambda_0) + \frac{1}{2}\overline{\mathbf{v}}_2^{\perp}(\lambda, \lambda_0), \;\; r = \frac{1}{2}\|\overline{\mathbf{v}}_2^{\perp}(\lambda, \lambda_0)\|_2.$$

It is easy to see that

$$\begin{aligned}\left\|\mathbf{X}_g^T\theta^*(\lambda)\right\|_2 &\leq \|\mathbf{X}_g^T(\theta^*(\lambda) - \mathbf{o})\|_2 + \|\mathbf{X}_g^T\mathbf{o}\|_2 \\ &< \|\mathbf{X}_g\|_2\|\theta^*(\lambda) - \mathbf{o}\|_2 + \sqrt{n_g} - r\|\mathbf{X}_g\|_2 \\ &\leq r\|\mathbf{X}_g\|_2 + \sqrt{n_g} - r\|\mathbf{X}_g\|_2 = \sqrt{n_g},\end{aligned} \tag{61}$$

which completes the proof. The second and third inequalities in (61) are due to (60) and Theorem 22, respectively. ■

In view of (45) and Theorem 23, we can derive the EDPP rule to discard the inactive groups for the group Lasso problem as follows.

Corollary 24 EDPP: *For the group Lasso problem (43), suppose we are given a sequence of parameter values $\overline{\lambda}_{\max} = \lambda_0 > \lambda_1 > \ldots > \lambda_{\mathcal{K}}$. For any integer $0 \leq k < \mathcal{K}$, we have $\beta_g^*(\lambda_{k+1}) = 0$ if $\beta^*(\lambda_k)$ is known and the following holds:*

$$\left\|\mathbf{X}_g^T\left(\frac{\mathbf{y} - \sum_{g=1}^G \mathbf{X}_g\beta_g^*(\lambda_k)}{\lambda_k} + \frac{1}{2}\overline{\mathbf{v}}_2^{\perp}(\lambda_{k+1}, \lambda_k)\right)\right\|_2 < \sqrt{n_g} - \frac{1}{2}\|\overline{\mathbf{v}}_2^{\perp}(\lambda_{k+1}, \lambda_k)\|_2\|\mathbf{X}_g\|_2.$$

4. Experiments

In this section, we evaluate the proposed EDPP rules for Lasso and group Lasso on both synthetic and real data sets. To measure the performance of our screening rules, we compute the *rejection ratio* and *speedup* (please refer to Section 2.3.3 for details). Because the EDPP rule is safe, i.e., no active features/groups will be mistakenly discarded, the rejection ratio will be less than one.

In Section 4.1, we conduct two sets of experiments to compare the performance of EDPP against several state-of-the-art screening methods. We first compare the performance of the basic versions of EDPP, DOME, SAFE, and strong rule. Then, we focus on the sequential versions of EDPP, SAFE, and strong rule. Notice that, SAFE and EDPP are safe. However, strong rule may mistakenly discard features with nonzero coefficients in the solution. Although DOME is also safe for the Lasso problem, it is unclear if there exists a sequential version of DOME. Recall that, real applications usually favor the sequential screening rules because we need to solve a sequence of of Lasso problems to determine an appropriate parameter value (Tibshirani et al., 2012). Moreover, DOME assumes special structure on the data, i.e., each feature and the response vector should be normalized to have unit length.

In Section 4.2, we compare EDPP with strong rule for the group Lasso problem on synthetic data sets. We are not aware of any *safe* screening rules for the group Lasso problem at this point. For SAFE and Dome, it is not straightforward to extend them to the group Lasso problem.

An efficient MATLAB implementation of the EDPP screening rules—combined with the solvers from SLEP package (Liu et al., 2009)—for both Lasso and group Lasso is available at `http://dpc-screening.github.io/`.

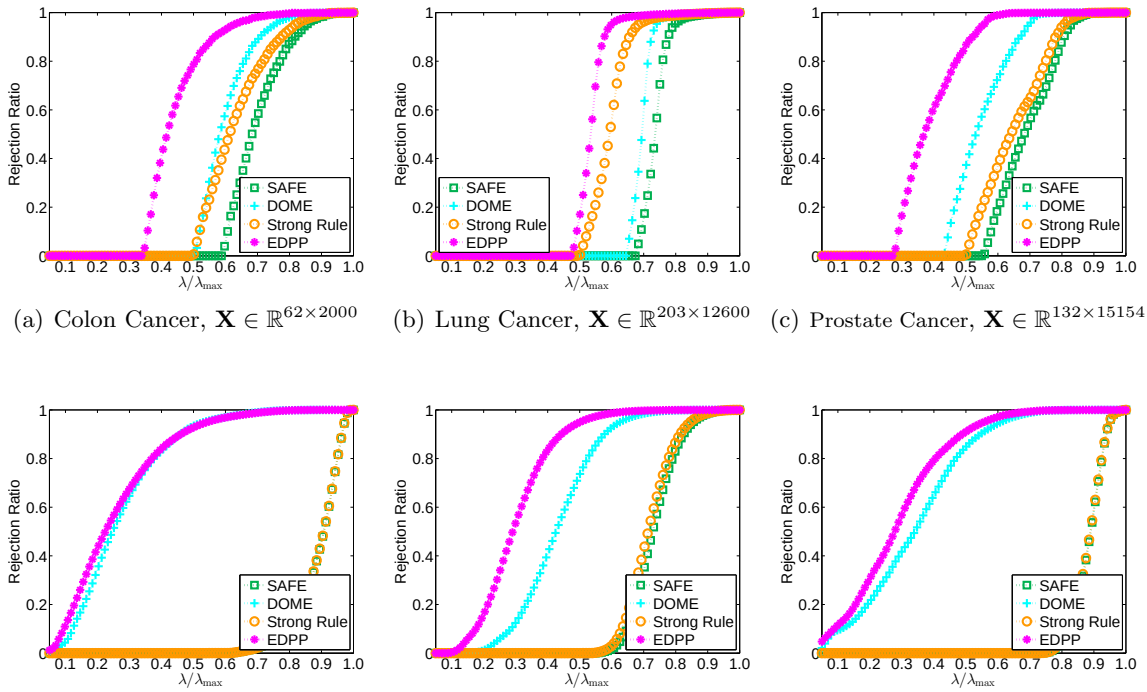
4.1 EDPP for the Lasso Problem

For the Lasso problem, we first compare the performance of the basic versions of EDPP, DOME, SAFE and strong rule in Section 4.1.1. Then, we compare the performance of the sequential versions of EDPP, SAFE and strong rule in Section 4.1.2.

4.1.1 Evaluation of the Basic EDPP Rule

In this section, we perform experiments on six real data sets to compare the performance of the basic versions of SAFE, DOME, strong rule and EDPP. Briefly speaking, suppose that we are given a parameter value λ. Basic versions of the aforementioned screening rules always make use of $\beta^*(\lambda_{\max})$ to identify the zero components of $\beta^*(\lambda)$. Take EDPP for example. The basic version of EDPP can be obtained by replacing $\beta^*(\lambda_k)$ and $\mathbf{v}_2^{\perp}(\lambda_{k+1}, \lambda_k)$ with $\beta^*(\lambda_0)$ and $\mathbf{v}_2^{\perp}(\lambda_k, \lambda_0)$, respectively, in (42) for all $k = 1, \ldots, \mathcal{K}$.

In this experiment, we report the rejection ratios of the basic SAFE, DOME, strong rule and EDPP along a sequence of 100 parameter values equally spaced on the $\lambda/\lambda_{\max}$ scale from 0.05 to 1.0. We note that DOME requires that all features of the data sets have unit length. Therefore, to compare the performance of DOME with SAFE, strong rule and EDPP, we normalize the features of all the data sets used in this section. However,



(a) Colon Cancer, $\mathbf{X} \in \mathbb{R}^{62\times 2000}$ (b) Lung Cancer, $\mathbf{X} \in \mathbb{R}^{203\times 12600}$ (c) Prostate Cancer, $\mathbf{X} \in \mathbb{R}^{132\times 15154}$

(d) PIE, $\mathbf{X} \in \mathbb{R}^{1024\times 11553}$ (e) MNIST, $\mathbf{X} \in \mathbb{R}^{784\times 50000}$ (f) COIL-100, $\mathbf{X} \in \mathbb{R}^{1024\times 7199}$

Figure 2: Comparison of basic versions of SAFE, DOME, Strong Rule and EDPP on six real data sets.

it is worthwhile to mention that SAFE, strong rule and EDPP do not assume any specific structures on the data set. The data sets used in this section are listed as follows:

a) Colon Cancer data set (Alon et al., 1999);

b) Lung Cancer data set (Bhattacharjee et al., 2001);

c) Prostate Cancer data set (Petricoin et al., 2002);

d) PIE face image data set (Sim et al., 2003; Cai et al., 2007);

e) MNIST handwritten digit data set (Lecun et al., 1998);

f) COIL-100 image data set (Nene et al., 1996; Cai et al., 2011).

The Colon Cancer data set contains gene expression information of 22 normal tissues and 40 colon cancer tissues, and each has 2000 gene expression values.

The Lung Cancer data set contains gene expression information of 186 lung tumors and 17 normal lung specimens. Each specimen has 12600 expression values.

The COIL-100 image data set consists of images of 100 objects. The images of each object are taken every 5 degree by rotating the object, yielding 72 images per object. The

dimension of each image is 32×32. In each trial, we randomly select one image as the response vector and use the remaining ones as the data matrix. We run 100 trials and report the average performance of the screening rules.

The description and the experimental settings for the Prostate Cancer data set, the PIE face image data set and the MNIST handwritten digit data set are given in Section 2.3.3.

Figure 2 reports the rejection ratios of the basic versions of SAFE, DOME, strong rule and EDPP. We can see that EDPP significantly outperforms the other three screening methods on five of the six data sets, i.e., the Colon Cancer, Lung Cancer, Prostate Cancer, MNIST, and COIL-100 data sets. On the PIE face image data set, EDPP and DOME provide similar performance and both significantly outperform SAFE and strong rule.

However, as pointed out by Tibshirani et al. (2012), the real strength of screening methods stems from their sequential versions. The reason is because the optimal parameter value is unknown in real applications. Typical approaches for model selection usually involve solving the Lasso problems many times along a sequence of parameter values. Thus, the sequential screening methods are more suitable in facilitating the aforementioned scenario and more useful than their basic-version counterparts in practice (Tibshirani et al., 2012).

4.1.2 Evaluation of the Sequential EDPP Rule

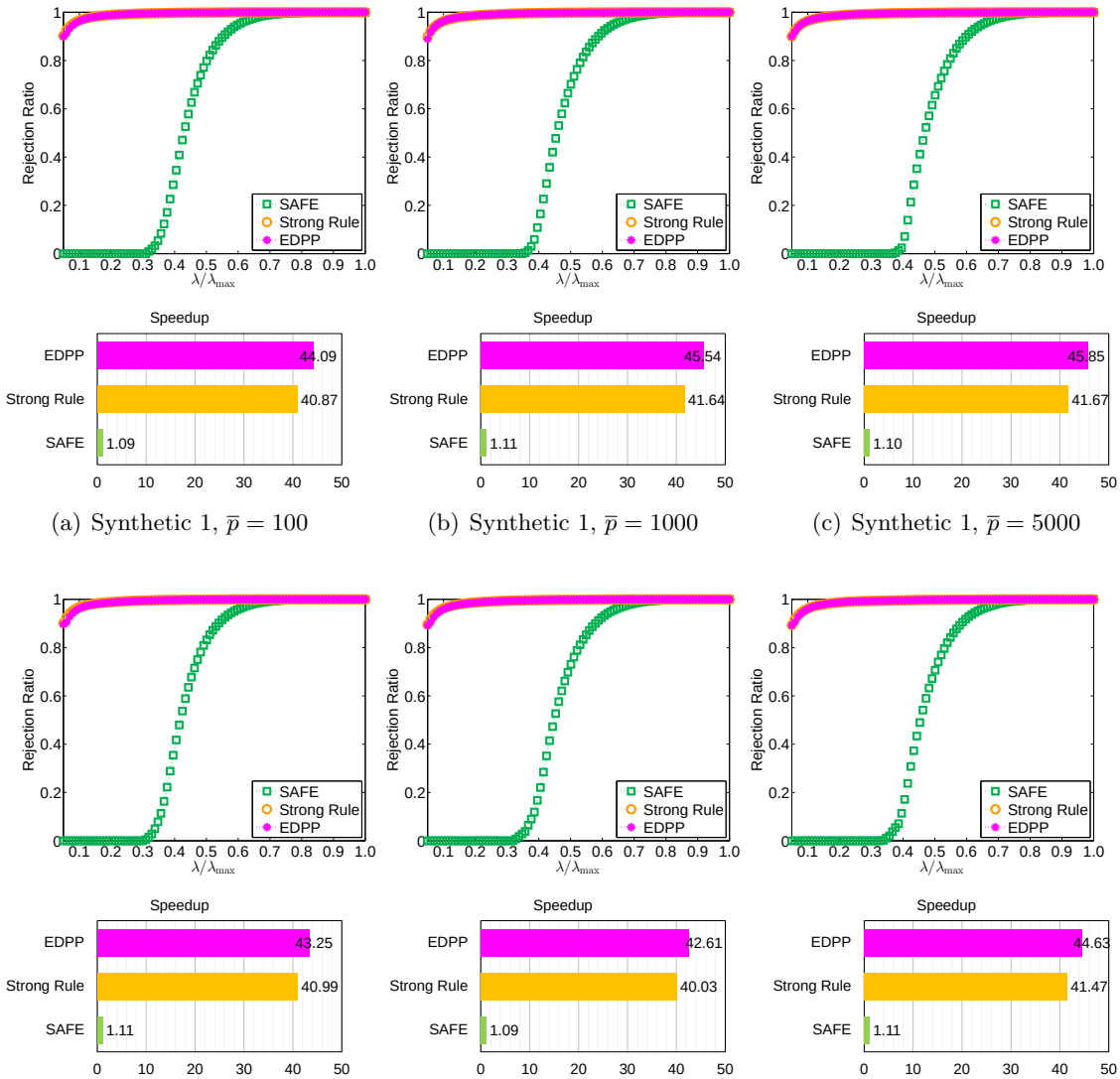
In this section, we compare the performance of the sequential versions of SAFE, strong rule and EDPP by the rejection ratio and speedup. We first perform experiments on two synthetic data sets. We then apply the three screening rules to six real data sets.

We first perform experiments on several synthetic problems, which have been commonly used in the sparse learning literature (Bondell and Reich, 2008; Zou and Hastie, 2005; Tibshirani, 1996). We simulate data from the true model

$$\mathbf{y} = \mathbf{X}\beta^* + \sigma\epsilon, \;\; \epsilon \sim N(0,1). \tag{62}$$

We generate two data sets with 250×10000 entries: Synthetic 1 and Synthetic 2. For Synthetic 1, the entries of the data matrix $\mathbf{X}$ are i.i.d. standard Gaussian with pairwise correlation zero, i.e., $\text{corr}(\mathbf{x}_i, \mathbf{x}_i) = 0$. For Synthetic 2, the entries of the data matrix $\mathbf{X}$ are drawn from i.i.d. standard Gaussian with pairwise correlation $0.5^{|i-j|}$, i.e., $\text{corr}(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}$. To generate the response vector $\mathbf{y} \in \mathbb{R}^{250}$ by the model in (62), we need to set the parameter σ and construct the ground truth $\beta^* \in \mathbb{R}^{10000}$. Throughout this section, σ is set to be 0.1. To construct β^*, we randomly select $\overline{p}$ components which are populated from a uniform $[-1, 1]$ distribution, and set the remaining ones as 0. After we generate the data matrix $\mathbf{X}$ and the response vector $\mathbf{y}$, we run the solver with or without screening rules to solve the Lasso problems along a sequence of 100 parameter values equally spaced on the $\lambda/\lambda_{\max}$ scale from 0.05 to 1.0. We then run 100 trials and report the average performance.

We first apply the screening rules, i.e., SAFE, strong rule and EDPP to Synthetic 1 with $\overline{p} = 100, 1000, 5000$ respectively. Figure 3(a), Figure 3(b) and Figure 3(c) present the corresponding rejection ratios and speedup of SAFE, strong rule and EDPP. We can see that the rejection ratios of strong rule and EDPP are comparable to each other, and both of them are more effective in discarding inactive features than SAFE. In terms of the speedup, EDPP provides better performance than strong rule. The reason is because strong rule is a heuristic screening method, i.e., it may mistakenly discard active features which have



(a) Synthetic 1, $\overline{p} = 100$ (b) Synthetic 1, $\overline{p} = 1000$ (c) Synthetic 1, $\overline{p} = 5000$

(d) Synthetic 2, $\overline{p} = 100$ (e) Synthetic 2, $\overline{p} = 1000$ (f) Synthetic 2, $\overline{p} = 5000$

Figure 3: Comparison of SAFE, Strong Rule and EDPP on two synthetic data sets with different numbers of nonzero components of the ground truth.

nonzero components in the solution. Thus, strong rule needs to check the KKT conditions to ensure the correctness of the screening result. In contrast, the EDPP rule does not need to check the KKT conditions since the discarded features are guaranteed to be absent from the resulting sparse representation. From the last two columns of Table 2, we can observe that the running time of strong rule is about twice of that of EDPP.

Figure 3(d), Figure 3(e) and Figure 3(f) present the rejection ratios and speedup of SAFE, strong rule and EDPP on Synthetic 2 with $\overline{p} = 100, 1000, 5000$ respectively. We can

Data	$\overline{p}$	solver	SAFE+solver	Strong Rule+solver	EDPP+solver	SAFE	Strong Rule	EDPP
Synthetic 1	100	109.01	100.09	2.67	2.47	4.60	0.65	0.36
	1000	123.60	111.32	2.97	2.71	4.59	0.66	0.37
	5000	124.92	113.09	3.00	2.72	4.57	0.65	0.36
Synthetic 2	100	107.50	96.94	2.62	2.49	4.61	0.67	0.37
	1000	113.59	104.29	2.84	2.67	4.57	0.63	0.35
	5000	125.25	113.35	3.02	2.81	4.62	0.65	0.36

Table 2: Running time (in seconds) for solving the Lasso problems along a sequence of 100 tuning parameter values equally spaced on the scale of $\lambda/\lambda_{\max}$ from 0.05 to 1 by (a): the solver (Liu et al., 2009) (reported in the third column) without screening; (b): the solver combined with different screening methods (reported in the 4th to the 6th columns). The last four columns report the total running time (in seconds) for the screening methods.

observe patterns similar to Synthetic 1. Clearly, our method, EDPP, is very robust to the variations of the intrinsic structures of the data sets and the sparsity of the ground truth.

We next compare the performance of the EDPP rule with SAFE and strong rule on six real data sets along a sequence of 100 parameter values equally spaced on the $\lambda/\lambda_{\max}$ scale from 0.05 to 1.0. The data sets are listed as follows:

a) Breast Cancer data set (West et al., 2001; Shevade and Keerthi, 2003);

b) Leukemia data set (Armstrong et al., 2002);

c) Prostate Cancer data set (Petricoin et al., 2002);

d) PIE face image data set (Sim et al., 2003; Cai et al., 2007);

e) MNIST handwritten digit data set (Lecun et al., 1998);

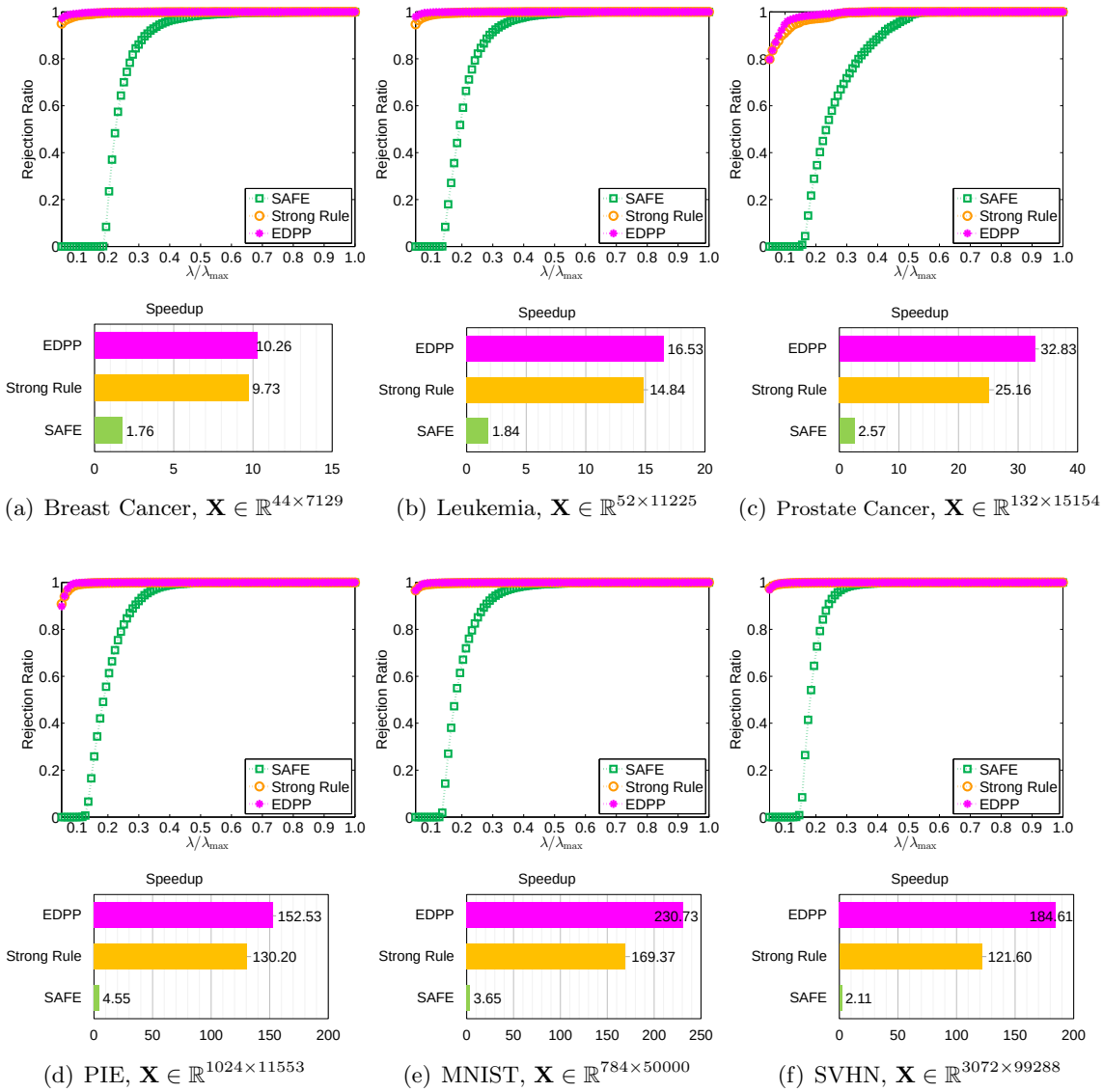
f) Street View House Number (SVHN) data set (Netzer et al., 2001).

We present the rejection ratios and speedup of EDPP, SAFE and strong rule in Figure 4. Table 3 reports the running time of the solver with or without screening for solving the 100 Lasso problems, and that of the screening rules.

The Breast Cancer data set contains 44 tumor samples, each of which is represented by 7129 genes. Therefore, the data matrix $\mathbf{X}$ is of 44×7129. The response vector $\mathbf{y} \in \{1, -1\}^{44}$ contains the binary label of each sample.

The Leukemia data set is a DNA microarray data set, containing 52 samples and 11225 genes. Therefore, the data matrix $\mathbf{X}$ is of 52×11225. The response vector $\mathbf{y} \in \{1, -1\}^{52}$ contains the binary label of each sample.

The SVHN data set contains color images of street view house numbers, including 73257 images for training and 26032 for testing. The dimension of each image is 32×32.



(a) Breast Cancer, $\mathbf{X} \in \mathbb{R}^{44\times 7129}$ (b) Leukemia, $\mathbf{X} \in \mathbb{R}^{52\times 11225}$ (c) Prostate Cancer, $\mathbf{X} \in \mathbb{R}^{132\times 15154}$

(d) PIE, $\mathbf{X} \in \mathbb{R}^{1024\times 11553}$ (e) MNIST, $\mathbf{X} \in \mathbb{R}^{784\times 50000}$ (f) SVHN, $\mathbf{X} \in \mathbb{R}^{3072\times 99288}$

Figure 4: Comparison of SAFE, Strong Rule, and EDPP on six real data sets.

In each trial, we first randomly select an image as the response $\mathbf{y} \in \mathbb{R}^{3072}$, and then use the remaining ones to form the data matrix $\mathbf{X} \in \mathbb{R}^{3072\times 99288}$. We run 100 trials and report the average performance.

The description and the experiment settings for the Prostate Cancer data set, the PIE face image data set and the MNIST handwritten digit data set are given in Section 2.3.3.

From Figure 4, we can see that the rejection ratios of strong rule and EDPP are comparable to each other. Compared to SAFE, both of strong rule and EDPP are able to identify far more inactive features, leading to a much higher speedup. However, because strong rule needs to check the KKT conditions to ensure the correctness of the screening

Data	solver	SAFE+solver	Strong Rule+solver	EDPP+solver	SAFE	Strong Rule	EDPP
Breast Cancer	12.70	7.20	1.31	1.24	0.44	0.06	0.05
Leukemia	16.99	9.22	1.15	1.03	0.91	0.09	0.07
Prostate Cancer	121.41	47.17	4.83	3.70	3.60	0.46	0.23
PIE	629.94	138.33	4.84	4.13	19.93	2.54	1.33
MNIST	2566.26	702.21	15.15	11.12	64.81	8.14	4.19
SVHN	11023.30	5220.88	90.65	59.71	583.12	61.02	31.64

Table 3: Running time (in seconds) for solving the Lasso problems along a sequence of 100 tuning parameter values equally spaced on the scale of $\lambda/\lambda_{\max}$ from 0.05 to 1 by (a): the solver (Liu et al., 2009) (reported in the second column) without screening; (b): the solver combined with different screening methods (reported in the 3rd to the 5th columns). The last three columns report the total running time (in seconds) for the screening methods.

results, the speedup gained by EDPP is higher than that by strong rule. When the size of the data matrix is not very large, e.g., the Breast Cancer and Leukemia data sets, the speedup gained by EDPP are slightly higher than that by strong rule. However, when the size of the data matrix is large, e.g., the MNIST and SVHN data sets, the speedup gained by EDPP are significantly higher than that by strong rule. Moreover, we can also observe from Figure 4 that, the larger the data matrix is, the higher the speedup can be gained by EDPP. More specifically, for the small data sets, e.g., the Breast Cancer, Leukemia and Prostate Cancer data sets, the speedup gained by EDPP is about 10, 17 and 30 times. In contrast, for the large data sets, e.g., the PIE, MNIST and SVHN data sets, the speedup gained by EDPP is two orders of magnitude. Take the SVHN data set for example. The solver without screening needs about 3 *hours* to solve the 100 Lasso problems. Combined with the EDPP rule, the solver only needs less than 1 *minute* to complete the task.

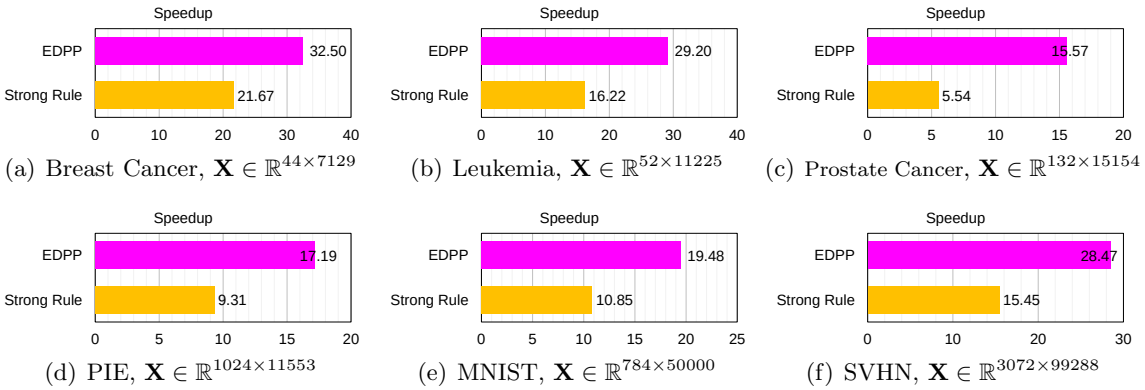
Clearly, the proposed EDPP screening rule is very effective in accelerating the computation of Lasso especially for large-scale problems, and outperforms the state-of-the-art approaches like SAFE and strong rule. Notice that, the EDPP method is safe in the sense that the discarded features are guaranteed to have zero coefficients in the solution.

4.1.3 EDPP with Least-Angle Regression (LARS)

As we mentioned in the introduction, we can combine EDPP with any existing solver. In this experiment, we integrate EDPP and strong rule with another state-of-the-art solver for Lasso, i.e., Least-Angle Regression (LARS) (Efron et al., 2004). We perform experiments on the same real data sets used in the last section with the same experiment settings. Because the rejection ratios of screening methods are irrelevant to the solvers, we only report the speedup. Table 4 reports the running time of LARS with or without screening for solving the 100 Lasso problems, and that of the screening methods. Figure 5 shows the speedup of these two methods. We can still observe a substantial speedup gained by EDPP. The

Data	LARS	Strong Rule+LARS	EDPP+LARS	Strong Rule	EDPP
Breast Cancer	1.30	0.06	0.04	0.04	0.03
Leukemia	1.46	0.09	0.05	0.07	0.04
Prostate Cancer	5.76	1.04	0.37	0.42	0.24
PIE	22.52	2.42	1.31	2.30	1.21
MNIST	92.53	8.53	4.75	8.36	4.34
SVHN	1017.20	65.83	35.73	62.53	32.00

Table 4: Running time (in seconds) for solving the Lasso problems along a sequence of 100 tuning parameter values equally spaced on the scale of $\lambda/\lambda_{\max}$ from 0.05 to 1 by (a): the solver (Efron et al., 2004; Mairal et al., 2010) (reported in the second column) without screening; (b): the solver combined with different screening methods (reported in the 3rd and 4th columns). The last two columns report the total running time (in seconds) for the screening methods.



(a) Breast Cancer, $\mathbf{X} \in \mathbb{R}^{44\times 7129}$ (b) Leukemia, $\mathbf{X} \in \mathbb{R}^{52\times 11225}$ (c) Prostate Cancer, $\mathbf{X} \in \mathbb{R}^{132\times 15154}$

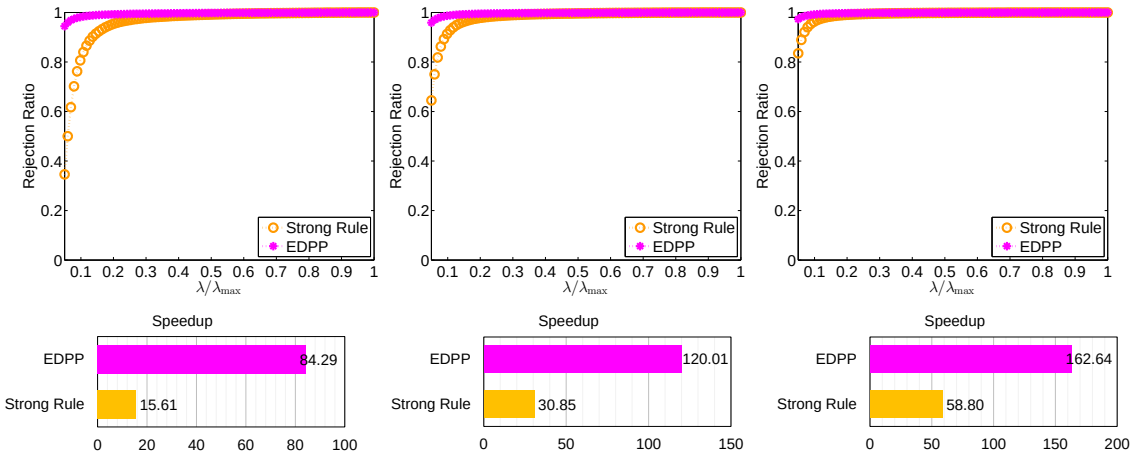
(d) PIE, $\mathbf{X} \in \mathbb{R}^{1024\times 11553}$ (e) MNIST, $\mathbf{X} \in \mathbb{R}^{784\times 50000}$ (f) SVHN, $\mathbf{X} \in \mathbb{R}^{3072\times 99288}$

Figure 5: The speedup gained by Strong Rule and EDPP combined with LARS on six real data sets.

reason is that EDPP has a very low computational cost (see Table 4) and it is very effective in discarding inactive features (see Figure 4).

4.2 EDPP for the Group Lasso Problem

In this experiment, we evaluate the performance of EDPP and strong rule with different numbers of groups. The data matrix $\mathbf{X}$ is fixed to be 250×200000. The entries of the response vector $\mathbf{y}$ and the data matrix $\mathbf{X}$ are generated i.i.d. from a standard Gaussian distribution. For each experiment, we repeat the computation 20 times and report the average results. Moreover, let n_g denote the number of groups and s_g be the average group size. For example, if n_g is 10000, then $s_g = p/n_g = 20$.



(a) $n_g = 10000$ (b) $n_g = 20000$ (c) $n_g = 40000$

Figure 6: Comparison of EDPP and strong rules with different numbers of groups.

n_g	solver	Strong Rule+solver	EDPP+solver	Strong Rule	EDPP
10000	4535.54	296.60	53.81	13.99	8.32
20000	5536.18	179.48	46.13	14.16	8.61
40000	6144.48	104.50	37.78	13.13	8.37

Table 5: Running time (in seconds) for solving the group Lasso problems along a sequence of 100 tuning parameter values equally spaced on the scale of $\lambda/\lambda_{\max}$ from 0.05 to 1.0 by (a): the solver from SLEP (reported in the second column) without screening; (b): the solver combined with different screening methods (reported in the 3rd and 4th columns). The last two columns report the total running time (in seconds) for the screening methods. The data matrix $\mathbf{X}$ is of size 250×200000.

From Figure 6, we can see that EDPP and strong rule are able to discard more inactive groups when the number of groups n_g increases. The intuition behind this observation is that the estimation of the dual optimal solution is more accurate with a smaller group size. Notice that, a large n_g implies a small average group size. Figure 6 also implies that compared to strong rule, EDPP is able to discard more inactive groups and is more robust with respect to different values of n_g.

Table 5 further demonstrates the effectiveness of EDPP in improving the efficiency of the solver. When $n_g = 10000$, the efficiency of the solver is improved by about 80 times. When $n_g = 20000$ and 40000, the efficiency of the solver is boosted by about 120 and 160 times with EDPP respectively.

5. Conclusion

In this paper, we develop new screening rules for the Lasso problem by making use of the properties of the projection operators with respect to a closed convex set. Our proposed methods, i.e., DPP screening rules, are able to effectively identify inactive predictors of the Lasso problem, thus greatly reducing the size of the optimization problem. Moreover, we further improve DPP rule and propose the enhanced DPP rule, which is more effective in discarding inactive features than DPP rule. The idea of the family of DPP rules can be easily generalized to identify the inactive groups of the group Lasso problem. Extensive numerical experiments on both synthetic and real data demonstrate the effectiveness of the proposed rules. It is worthwhile to mention that the family of DPP rules can be combined with any Lasso solver as a speedup tool. In the future, we plan to generalize our ideas to other sparse formulations consisting of more general structured sparse penalties, e.g., tree/graph Lasso, fused Lasso.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (IIS-0953662, IIS-1421057, and IIS-1421100) and the National Institutes of Health (R01 LM010730 and U54 EB020403).

Appendix A. The Dual Problem of Lasso

In this appendix, we give the detailed derivation of the dual problem of Lasso.

A.1 Dual Formulation

Assuming the data matrix is $\mathbf{X} \in \mathbb{R}^{N\times p}$, the standard Lasso problem is given by:

$$\inf_{\beta\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1. \tag{63}$$

For completeness, we give a detailed deviation of the dual formulation of (63) in this section. Note that problem (63) has no constraints. Therefore the dual problem is trivial and useless. A common trick (Boyd and Vandenberghe, 2004) is to introduce a new set of variables $\mathbf{z} = \mathbf{y} - \mathbf{X}\beta$ such that problem (63) becomes:

$$\begin{aligned} \inf_{\beta} \quad & \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda\|\beta\|_1, \\ \text{subject to} \quad & \mathbf{z} = \mathbf{y} - \mathbf{X}\beta. \end{aligned} \tag{64}$$

By introducing the dual variables $\eta \in \mathbb{R}^N$, we get the Lagrangian of problem (64):

$$L(\beta, \mathbf{z}, \eta) = \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda\|\beta\|_1 + \eta^T \cdot (\mathbf{y} - \mathbf{X}\beta - \mathbf{z}).$$

For the Lagrangian, the primal variables are β and $\mathbf{z}$. And the dual function $g(\eta)$ is:

$$g(\eta) = \inf_{\beta,\mathbf{z}} L(\beta, \mathbf{z}, \eta) = \eta^T\mathbf{y} + \inf_{\beta}(-\eta^T\mathbf{X}\beta + \lambda\|\beta\|_1) + \inf_{\mathbf{z}}\big(\frac{1}{2}\|\mathbf{z}\|_2^2 - \eta^T\mathbf{z}\big).$$

In order to get $g(\eta)$, we need to solve the following two optimization problems.

$$\inf_{\beta} -\eta^T \mathbf{X}\beta + \lambda\|\beta\|_1, \tag{65}$$

and

$$\inf_{\mathbf{z}} \frac{1}{2}\|\mathbf{z}\|_2^2 - \eta^T \mathbf{z}. \tag{66}$$

Let us first consider problem (65). Denote the objective function of problem (65) as

$$f_1(\beta) = -\eta^T \mathbf{X}\beta + \lambda\|\beta\|_1. \tag{67}$$

$f_1(\beta)$ is convex but not smooth. Therefore let us consider its subgradient

$$\partial f_1(\beta) = -\mathbf{X}^T\eta + \lambda\mathbf{v},$$

in which $\|\mathbf{v}\|_\infty \leq 1$ and $\mathbf{v}^\mathbf{T}\beta = \|\beta\|_1$, i.e., $\mathbf{v}$ is the subgradient of $\|\beta\|_1$.

The necessary condition for f_1 to attain an optimum is

$$\exists\, \beta', \text{ such that } 0 \in \partial f_1(\beta') = \{-\mathbf{X}^T\eta + \lambda\mathbf{v}'\},$$

where $\mathbf{v}' \in \partial\|\beta'\|_1$. In other words, $\beta', \mathbf{v}'$ should satisfy

$$\mathbf{v}' = \frac{\mathbf{X}^T\eta}{\lambda}, \|\mathbf{v}'\|_\infty \leq 1, \mathbf{v}'^T\beta' = \|\beta'\|_1,$$

which is equivalent to

$$|\mathbf{x}_i^T\eta| \leq \lambda, i = 1, 2, \ldots, p.$$

Then we plug $\mathbf{v}' = \frac{\mathbf{X}^T\eta}{\lambda}$ and $\mathbf{v}'^T\beta' = \|\beta'\|_1$ into (67):

$$f_1(\beta') = \inf_{\beta} f_1(\beta) = -\eta^T\mathbf{X}\beta' + \lambda\big(\frac{\mathbf{X}^T\eta}{\lambda}\big)^T\beta' = 0.$$

Therefore, the optimum value of problem (65) is 0.

Next, let us consider problem (66). Denote the objective function of problem (66) as $f_2(\mathbf{z})$. Let us rewrite $f_2(\mathbf{z})$ as:

$$f_2(\mathbf{z}) = \frac{1}{2}(\|\mathbf{z} - \eta\|_2^2 - \|\eta\|_2^2).$$

Clearly,

$$\mathbf{z}' = \operatorname*{argmin}_{\mathbf{z}} f_2(\mathbf{z}) = \eta,$$

and

$$\inf_{\mathbf{z}} f_2(\mathbf{z}) = -\frac{1}{2}\|\eta\|_2^2.$$

Combining everything above, we get the dual problem:

$$\begin{aligned} \sup_{\eta} \quad & g(\eta) = \eta^T\mathbf{y} - \frac{1}{2}\|\eta\|_2^2, \\ \text{subject to} \quad & |\mathbf{x}_i^T\eta| \leq \lambda,\ i = 1, 2, \ldots, p. \end{aligned}$$

which is equivalent to

$$\begin{aligned} \sup_{\eta} \quad & g(\eta) = \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\|\eta - \mathbf{y}\|_2^2, \\ \text{subject to} \quad & |\mathbf{x}_i^T \eta| \leq \lambda,\ i = 1, 2, \ldots, p. \end{aligned} \tag{68}$$

By a simple re-scaling of the dual variables η, i.e., let $\theta = \frac{\eta}{\lambda}$, problem (68) transforms to:

$$\begin{aligned} \sup_{\theta} \quad & g(\theta) = \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{\lambda^2}{2}\|\theta - \frac{\mathbf{y}}{\lambda}\|_2^2, \\ \text{subject to} \quad & |\mathbf{x}_i^T \theta| \leq 1,\ i = 1, 2, \ldots, p. \end{aligned}$$

A.2 The KKT Conditions

Problem (64) is clearly convex and its constraints are all affine. By Slater's condition, as long as problem (64) is feasible we will have strong duality. Denote β^*, $\mathbf{z}^*$ and θ^* as optimal primal and dual variables. The Lagrangian is

$$L(\beta, \mathbf{z}, \theta) = \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda\|\beta\|_1 + \lambda\theta^T \cdot (\mathbf{y} - \mathbf{X}\beta - \mathbf{z}).$$

From the KKT condition, we have

$$0 \in \partial_\beta L(\beta^*, \mathbf{z}^*, \theta^*) = -\lambda \mathbf{X}^T \theta^* + \lambda \mathbf{v}, \text{ in which } \|\mathbf{v}\|_\infty \leq 1 \text{ and } \mathbf{v}^T \beta^* = \|\beta^*\|_1, \tag{69}$$

$$\nabla_{\mathbf{z}} L(\beta^*, \mathbf{z}^*, \theta^*) = \mathbf{z}^* - \lambda\theta^* = 0, \tag{70}$$

$$\nabla_{\theta} L(\beta^*, \mathbf{z}^*, \theta^*) = \lambda(\mathbf{y} - \mathbf{X}\beta^* - \mathbf{z}^*) = 0. \tag{71}$$

From (70) and (71), we have:

$$\mathbf{y} = \mathbf{X}\beta^* + \lambda\theta^*.$$

From (69), we know there exists $\mathbf{v}^* \in \partial\|\beta^*\|_1$ such that

$$\mathbf{X}^T\theta^* = \mathbf{v}^*,\ \|\mathbf{v}^*\|_\infty \leq 1 \text{ and } (\mathbf{v}^*)^T\beta^* = \|\beta^*\|_1,$$

which is equivalent to

$$|\mathbf{x}_i^T\theta^*| \leq 1, i = 1, 2, \ldots, p, \text{ and } (\theta^*)^T\mathbf{X}\beta^* = \|\beta^*\|_1. \tag{72}$$

From (72), it is easy to conclude:

$$(\theta^*)^T\mathbf{x}_i \in \begin{cases} \text{sign}(\beta_i^*) & \text{if } \beta_i^* \neq 0, \\ [-1, 1] & \text{if } \beta_i^* = 0. \end{cases}$$

Appendix B. The Dual Problem of Group Lasso

In this appendix, we present the detailed derivation of the dual problem of group Lasso.

B.1 Dual Formulation

Assuming the data matrix is $\mathbf{X}_g \in \mathbb{R}^{N\times n_g}$ and $p = \sum_{g=1}^{G} n_g$, the group Lasso problem is given by:

$$\inf_{\beta\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g\beta_g\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{n_g}\|\beta_g\|_2. \tag{73}$$

Let $\mathbf{z} = \mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g\beta_g$ and problem (73) becomes:

$$\begin{aligned} \inf_{\beta} \quad & \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{n_g}\|\beta_g\|_2, \\ \text{subject to} \quad & \mathbf{z} = \mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g\beta_g. \end{aligned} \tag{74}$$

By introducing the dual variables $\eta \in \mathbb{R}^N$, the Lagrangian of problem (74) is:

$$L(\beta, \mathbf{z}, \eta) = \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{n_g}\|\beta_g\|_2 + \eta^T \cdot (\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g\beta_g - \mathbf{z}).$$

and the dual function $g(\eta)$ is:

$$\begin{aligned} g(\eta) &= \inf_{\beta,\mathbf{z}} L(\beta, \mathbf{z}, \eta) \\ &= \eta^T\mathbf{y} + \inf_{\beta} \Big(- \eta^T \sum_{g=1}^{G} \mathbf{X}_g\beta_g + \lambda \sum_{g=1}^{G} \sqrt{n_g}\|\beta_g\|_2 \Big) + \inf_{\mathbf{z}} \big(\frac{1}{2}\|\mathbf{z}\|_2^2 - \eta^T\mathbf{z}\big). \end{aligned}$$

In order to get $g(\eta)$, let us solve the following two optimization problems.

$$\inf_{\beta} -\eta^T \sum_{g=1}^{G} \mathbf{X}_g\beta_g + \lambda \sum_{g=1}^{G} \sqrt{n_g}\|\beta_g\|_2, \tag{75}$$

and

$$\inf_{\mathbf{z}} \frac{1}{2}\|\mathbf{z}\|_2^2 - \eta^T\mathbf{z}. \tag{76}$$

Let us first consider problem (75). Denote the objective function of problem (75) as

$$\hat{f}(\beta) = -\eta^T \sum_{g=1}^{G} \mathbf{X}_g\beta_g + \lambda \sum_{g=1}^{G} \sqrt{n_g}\|\beta_g\|_2,$$

Let

$$\hat{f}_g(\beta_g) = -\eta^T\mathbf{X}_g\beta_g + \lambda\sqrt{n_g}\|\beta_g\|_2, \qquad g = 1, 2, \ldots, G.$$

then we can split problem (75) into a set of subproblems. Clearly $\hat{f}_g(\beta_g)$ is convex but not smooth because it has a singular point at 0. Consider the subgradient of $\hat{f}_g$,

$$\partial\hat{f}_g(\beta_g) = -\mathbf{X}_g^T\eta + \lambda\sqrt{n_g}\mathbf{v}_g, \qquad g = 1, 2, \ldots, G,$$

where $\mathbf{v}_g$ is the subgradient of $\|\beta_g\|_2$:

$$\mathbf{v}_g \in \begin{cases} \frac{\beta_g}{\|\beta_g\|_2} & \text{if } \beta_g \neq 0, \\ \mathbf{u},\ \|\mathbf{u}\|_2 \leq 1 & \text{if } \beta_g = 0. \end{cases} \tag{77}$$

Let β_g' be the optimal solution of $\hat{f}_g$, then β_g' satisfy

$$\exists \mathbf{v}_g' \in \partial \|\beta_g'\|_2, \quad -\mathbf{X}_g^T \eta + \lambda \sqrt{n_g} \mathbf{v}_g' = 0.$$

If $\beta_g' = 0$, clearly, $\hat{f}_g(\beta_g') = 0$. Otherwise, since $\lambda \sqrt{n_g} \mathbf{v}_g' = \mathbf{X}_g^T \eta$ and $\mathbf{v}_g' = \frac{\beta_g'}{\|\beta_g'\|_2}$, we have

$$\hat{f}_g(\beta_g') = -\lambda \sqrt{n_g} \frac{(\beta_g')^T}{\|\beta_g'\|_2} \beta_g' + \lambda \sqrt{n_g} \|\beta_g'\|_2 = 0.$$

All together, we can conclude the

$$\inf_{\beta_g} \hat{f}_g(\beta_g) = 0, \quad g = 1, 2, \ldots, G$$

and thus

$$\inf_{\beta} \hat{f}(\beta) = \inf_{\beta} \sum_{g=1}^{G} \hat{f}_g(\beta_g) = \sum_{g=1}^{G} \inf_{\beta_g} \hat{f}_g(\beta_g) = 0.$$

The second equality is due to the fact that β_g's are independent.

Note, from (77), it is easy to see $\|\mathbf{v}_g\|_2 \leq 1$. Since $\lambda \sqrt{n_g} \mathbf{v}_g' = \mathbf{X}_g^T \eta$, we get a constraint on η, i.e., η should satisfy:

$$\|\mathbf{X}_g^T \eta\|_2 \leq \lambda \sqrt{n_g}, \qquad g = 1, 2, \ldots, G.$$

Next, let us consider problem (76). Since problem (76) is exactly the same as problem (66), we conclude:

$$\mathbf{z}' = \operatorname*{argmin}_{\mathbf{z}} \frac{1}{2} \|\mathbf{z}\|_2^2 - \eta^T \mathbf{z} = \eta,$$

and

$$\inf_{\mathbf{z}} \frac{1}{2} \|\mathbf{z}\|_2^2 - \eta^T \mathbf{z} = -\frac{1}{2} \|\eta\|_2^2.$$

Therefore the dual function $g(\eta)$ is:

$$g(\eta) = \eta^T \mathbf{y} - \frac{1}{2} \|\eta\|_2^2.$$

Combining everything above, we get the dual formulation of the group Lasso:

$$\begin{aligned} \sup_{\eta} \quad & g(\eta) = \eta^T \mathbf{y} - \frac{1}{2} \|\eta\|_2^2, \\ \text{subject to} \quad & \|\mathbf{X}_g^T \eta\|_2 \leq \lambda \sqrt{n_g},\ g = 1, 2, \ldots, G. \end{aligned}$$

which is equivalent to

$$\begin{aligned} \sup_{\eta} \quad & g(\eta) = \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\|\eta - \mathbf{y}\|_2^2, \\ \text{subject to} \quad & \|\mathbf{X}_g^T \eta\|_2 \leq \lambda\sqrt{n_g},\ g = 1, 2, \ldots, G. \end{aligned} \tag{78}$$

By a simple re-scaling of the dual variables η, i.e., let $\theta = \frac{\eta}{\lambda}$, problem (78) transforms to:

$$\begin{aligned} \sup_{\theta} \quad & g(\theta) = \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{\lambda^2}{2}\|\theta - \frac{\mathbf{y}}{\lambda}\|_2^2, \\ \text{subject to} \quad & \|\mathbf{X}_g^T \theta\|_2 \leq \sqrt{n_g},\ g = 1, 2, \ldots, G. \end{aligned}$$

B.2 The KKT Conditions

Clearly, problem (74) is convex and its constraints are all affine. By Slater's condition, as long as problem (74) is feasible we will have strong duality. Denote β^*, $\mathbf{z}^*$ and θ^* as optimal primal and dual variables. The Lagrangian is

$$L(\beta, \mathbf{z}, \theta) = \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{n_g}\|\beta_g\|_2 + \lambda\theta^T \cdot (\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g \beta_g - \mathbf{z}).$$

From the KKT condition, we have

$$0 \in \partial_{\beta_g} L(\beta^*, \mathbf{z}^*, \theta^*) = -\lambda \mathbf{X}_g^T \theta^* + \lambda\sqrt{n_g}\mathbf{v}_g, \text{ in which } \mathbf{v}_g \in \partial\|\beta_g^*\|_2, \quad g = 1, 2, \ldots, G, \tag{79}$$

$$\nabla_{\mathbf{z}} L(\beta^*, \mathbf{z}^*, \theta^*) = \mathbf{z}^* - \lambda\theta^* = 0, \tag{80}$$

$$\nabla_{\theta} L(\beta^*, \mathbf{z}^*, \theta^*) = \lambda \cdot (\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g \beta_g^* - \mathbf{z}^*) = 0. \tag{81}$$

From (80) and (81), we have:

$$\mathbf{y} = \sum_{g=1}^{G} \mathbf{X}_g \beta_g^* + \lambda\theta^*.$$

From (79), we know there exists $\mathbf{v}_g' \in \partial\|\beta_g^*\|_2$ such that

$$\mathbf{X}_g^T \theta^* = \sqrt{n_g}\mathbf{v}_g'$$

and

$$\mathbf{v}_g' \in \begin{cases} \frac{\beta_g^*}{\|\beta_g^*\|_2} & \text{if } \beta_g^* \neq 0, \\ \mathbf{u},\ \|\mathbf{u}\|_2 \leq 1 & \text{if } \beta_g^* = 0, \end{cases}$$

Then the following holds:

$$\mathbf{X}_g^T \theta^* \in \begin{cases} \sqrt{n_g}\frac{\beta_g^*}{\|\beta_g^*\|_2} & \text{if } \beta_g^* \neq 0, \\ \sqrt{n_g}\mathbf{u},\ \|\mathbf{u}\|_2 \leq 1 & \text{if } \beta_g^* = 0, \end{cases}$$

for $g = 1, 2, \ldots, G$. Clearly, if $\|\mathbf{X}_g^T \theta^*\|_2 < \sqrt{n_g}$, we can conclude $\beta_g^* = 0$.

References

U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Cell Biology*, 96:6745–6750, 1999.

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

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

S. R. Becker, E. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. Technical report, Standford University, 2010.

D. P. Bertsekas. *Convex Analysis and Optimization*. Athena Scientific, 2003.

A. Bhattacharjee, W. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. Mark, E. Lander, W. Wong, B. Johnson, T. Golub, D. Sugarbaker, and M. Meyerson. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98:13790–13795, 2001.

H. Bondell and B. Reich. Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics*, 64:115–123, 2008.

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

A. Bruckstein, D. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51:34–81, 2009.

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

D. Cai, X. He, J. Han, and T. Huang. Graph regularized non-negative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1548–1560, 2011.

E. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematics*, 2006.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43:129–159, 2001.

D. L. Donoho and Y. Tsaig. Fast solution of l-1 norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54:4789–4812, 2008.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

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

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature spaces. *Journal of the Royal Statistical Society Series B*, 70:849–911, 2008.

J. Friedman, T. Hastie, H. Hëfling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1:302–332, 2007.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.

S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large scale l1-regularized least squares. *IEEE Journal on Selected Topics in Signal Processing*, 1:606–617, 2007.

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

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

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

S. Nene, S. Nayar, and H. Murase. Columbia object image library (coil-100). Technical report, CUCS-006-96, Columbia University, 1996.

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

M. Y. Park and T. Hastie. L1-regularized path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B*, 69:659–677, 2007.

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

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

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

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

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

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

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

J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. In *Proceedings of IEEE*, 2010.

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

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

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

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