Safe and Efficient Screening For Sparse Support Vector Machine

Zheng Zhao, Jun Liu, James Cox
SAS Institute Inc. 600 Research Drive, Cary, NC 27513
{zheng.zhao, jun.liu, james.cox}@sas.com

ABSTRACT

Sparse support vector machine (SVM) is a robust predictive model that can effectively remove noise and preserve signals. Like Lasso, it can efficiently learn a solution path based on a set of predefined parameters and therefore provides strong support for model selection. Sparse SVM has been successfully applied in a variety of data mining applications including text mining, bioinformatics, and image processing. The emergence of big-data analysis poses new challenges for model selection with large-scale data that consist of tens of millions samples and features. In this paper, a novel screening technique is proposed to accelerate model selection for ℓ_1-regularized ℓ_2-SVM and effectively improve its scalability. This technique can precisely identify inactive features in the optimal solution of a ℓ_1-regularized ℓ_2-SVM model and remove them before training. The technique makes use of the variational inequality and provides a closed-form solution for screening inactive features in different situations. Every feature that is removed by the screening technique is guaranteed to be inactive in the optimal solution. Therefore, when ℓ_1-regularized ℓ_2-SVM uses the features selected by the technique, it achieves exactly the same result as when it uses the full feature set. Because the technique can remove a large number of inactive features, it can greatly increase the efficiency of model selection for ℓ_1-regularized ℓ_2-SVM. Experimental results on five high-dimensional benchmark data sets demonstrate the power of the proposed technique.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*data mining*; I.5.2 [**Pattern Recognition**]: Design Methodology—*feature evaluation and selection*

Keywords

Screening, sparse support vector machine, feature selection

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD'14, August 24–27, 2014, New York, NY, USA.
Copyright 2014 ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623686.

1. INTRODUCTION

Sparse predictive modeling algorithms provide powerful tools to analyze high-dimensional data and generate results that have high-degree of interpretability and robustness [5, 11]. In general, an ℓ_1-regularized sparse predictive modeling algorithm can be formulated as $\min_{\boldsymbol{w}} \text{loss}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1$. Here $\boldsymbol{w} \in \mathbb{R}^m$ contains the model coefficients, $\text{loss}(\boldsymbol{w})$ is a loss function, and $\lambda \geq 0$ is the regularization parameter that balances between the loss function and the regularizer. When the hinge loss or its square form is used as the loss function, the resulting sparse model is the ℓ_1-regularized support vector machine (SVM) [4, 18, 2, 6, 16]. An ℓ_1-regularized SVM model can simultaneously perform model fitting by margin maximization and remove noisy features by soft-thresholding. It has been successfully applied in a variety of data mining applications that include text mining, bioinformatics, and image processing. Compared to other variances of sparse SVM model [15, 8, 1], ℓ_1-regularized SVM enjoys two major advantages. First, it defines a convex problem; therefore, an optimal solution can always be achieved without any relaxation of the original problem. Second, its optimization is simple, and a well implemented ℓ_1-regularized SVM solver can readily handle large-scale problems with tens of millions samples and features [6].

The value of the regularization parameter λ is crucial to the performance of an ℓ_1-regularized SVM. To achieve good performance, model selection is often used to help choose an appropriate λ value. For example, given a series of regularization parameters, $\lambda_1 > \lambda_2 > \ldots > \lambda_k$, the corresponding solutions, $\boldsymbol{w}_1^*, \boldsymbol{w}_2^*, \ldots, \boldsymbol{w}_k^*$, can be obtained and the best solution can be chosen by using a prespecified criterion, such as the accuracy or the area under the curve (AUC) that is achieved by the resulting models on holdout samples.

Big-data analysis requires a higher standard of efficiency for predictive modeling. When data are huge, the computational cost of model selection can be prohibitive. An intuitive question is to ask whether the solution obtained in the kth step of model selection can be used in the $(k+1)$th step to speed up computation. For Lasso [11], the answer leads to the state-of-the-art screening techniques to accelerate model selection [17, 7, 14, 12, 10]. The key idea is that, given a solution $\boldsymbol{w}_1^*$ for $\lambda = \lambda_1$, one can identify many features that are guaranteed to have zero coefficients in $\boldsymbol{w}_2^*$ when $\lambda = \lambda_2$. By removing a large number of these inactive features, the cost for computing $\boldsymbol{w}_2^*$ can be greatly reduced.

Although screening algorithms have been designed for Lasso, very little research has been done for screening for ℓ_1-regularized SVMs except in [7], which proposes a safe screening tech-

nique for ℓ_1-regularized ℓ_1-SVM. This paper presents a novel screening technique that is designed to speed up model selection for ℓ_1-regularized ℓ_2-SVM. The technique makes use of the variational inequality [9] for constructing a tight convex set, which can be used to compute bounds for screening features. Features that are removed by this technique are guaranteed to be inactive in the optimal solution. Therefore, the screening is "safe." Experimental results on five high-dimensional benchmark data sets demonstrate that the proposed screening technique can dramatically speed up model selection for ℓ_1-regularized ℓ_2-SVM by efficiently removing a large number of inactive features.

2. ℓ_1-REGULARIZED ℓ_2-SVM

Assume that $\mathbf{X} \in \mathbb{R}^{m \times n}$ is a data set that contains n samples, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, and m features, $\mathbf{X} = \left(\mathbf{f}_1^\top, \ldots, \mathbf{f}_m^\top\right)^\top$. Assume also that $\mathbf{y} = (y_1, \ldots, y_n)$ contains n class labels, $y_i \in \{-1, +1\}$, $i = 1, \ldots, n$. Let $\mathbf{w} \in \mathbb{R}^m$ be the m-dimensional weight vector, let $\xi_i \geq 0, i = 1, \ldots, n$ be the n slack variables, and let $b \in \mathbb{R}$ and $\lambda \in \mathbb{R}^+$ be the bias and the regularization parameter, respectively. The primal form of the ℓ_1-regularized ℓ_2-SVM is defined as:

$$\min_{\boldsymbol{\xi}, \mathbf{w}} \frac{1}{2} \sum_{i=1}^{n} \xi_i^2 + \lambda ||\mathbf{w}||_1, \quad (1)$$
$$s.t. \;\; y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i, \;\; \xi_i \geq 0.$$

Eq. (1) specifies a convex problem that has a non-smooth ℓ_1 regularizer, which enforces that the solution is sparse. Let $\boldsymbol{w}^\star(\lambda)$ be the optimal solution of Eq. (1) for a given λ. All the features that have nonzero values in $\boldsymbol{w}^\star(\lambda)$ are called active features, and the other features are called inactive. Let $\boldsymbol{\alpha} \in \mathbb{R}^n$ be the n-dimensional dual variable. By applying the Lagrangian multiplier [3], the dual of the problem defined in Eq. (1) can be obtained as:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha} - \mathbf{1}\|_2^2, \quad (2)$$
$$s.t. \;\; \|\hat{\mathbf{f}}_j^\top \boldsymbol{\alpha}\| \leq \lambda, \;\; j = 1, \ldots, m,$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0, \; \boldsymbol{\alpha} \succcurlyeq \mathbf{0}.$$

Here, $\hat{\mathbf{f}} = \mathbf{Y}\mathbf{f}$, and $\mathbf{Y} = diag(\mathbf{y})$ is a diagonal matrix. By defining $\boldsymbol{\alpha} = \lambda \boldsymbol{\theta}$, Eq. (2) can be reformulated as:

$$\min_{\boldsymbol{\theta}} ||\boldsymbol{\theta} - \frac{\mathbf{1}}{\lambda}||_2^2, \quad (3)$$
$$s.t. \;\; \|\hat{\mathbf{f}}_j^\top \boldsymbol{\theta}\| \leq 1, \;\; j = 1, \ldots, m,$$
$$\sum_{i=1}^{n} \theta_i y_i = 0, \; \boldsymbol{\theta} \succcurlyeq \mathbf{0}.$$

In the primal formulation for the ℓ_1-regularized ℓ_2-SVM, the primal variables are b, $\mathbf{w}$, and $\boldsymbol{\xi}$. And in the dual formulation, the dual variables are $\boldsymbol{\alpha}$ or $\boldsymbol{\theta}$. When b and $\mathbf{w}$ are known, $\boldsymbol{\xi}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\theta}$ can be obtained as:

$$\xi_i = \alpha_i = \lambda \theta_i = \max\left(0, 1 - y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right)\right). \quad (4)$$

The relation between $\boldsymbol{\alpha}$ and $\mathbf{w}$ can be expressed as:

$$\boldsymbol{\alpha}^\top \hat{\mathbf{f}}_j = \begin{cases} \text{sign}\,(w_j)\,\lambda, & \text{if } w_j \neq 0 \\ [-\lambda, +\lambda], & \text{if } w_j = 0 \end{cases}, \quad j = 1, \ldots, m. \quad (5)$$

The relation between $\boldsymbol{\theta}$ and $\mathbf{w}$ can be expressed as:

$$\boldsymbol{\theta}^\top \hat{\mathbf{f}}_j = \begin{cases} \text{sign}\,(w_j), & \text{if } w_j \neq 0 \\ [-1, +1], & \text{if } w_j = 0 \end{cases}, \quad j = 1, \ldots, m. \quad (6)$$

$\lambda_{\max}$ is defined as the smallest λ value that leads to $\mathbf{w} = \mathbf{0}$ when it is used in Eq. (1). Given an input data set $(\mathbf{X}, \mathbf{y})$, $\lambda_{\max}$ can be obtained in a closed form as:

$$\lambda_{\max} = \left\| \sum_{i=1}^{n} \left(y_i - \frac{n_+ - n_-}{n}\right) \mathbf{x}_i \right\|_\infty, \quad (7)$$

where n_+ and n_- denote the number of positive and negative samples, respectively. And when $\lambda \geq \lambda_{max}$, the optimal solution of the problem defined in Eq. (1) can be written as:

$$\mathbf{w}^\star = \mathbf{0}, \;\; b^\star = \frac{(n_+ - n_-)}{n}. \quad (8)$$

Denote $\mathbf{m} = \sum_{i=1}^{n} \left(y_i - \frac{n_+ - n_-}{n}\right) \mathbf{x}_i$. The first feature to enter the model is the one that corresponds to the element that has the largest magnitude in $\mathbf{m}$.

3. EFFICIENT AND SAFE SCREENING FOR ℓ_1-REGULARIZED ℓ_2-SVM

Eq. (6) shows that the necessary condition for a feature $\mathbf{f}$ to be active in an optimal solution is $|\boldsymbol{\theta}^\top \hat{\mathbf{f}}| = 1$. On the other hand, for any feature $\mathbf{f}$, if $|\boldsymbol{\theta}^\top \hat{\mathbf{f}}| < 1$, it must be inactive in the optimal solution. Given a λ value, this condition can be used to develop a screening rule for removing inactive features to speed up training for the ℓ_1-regularized ℓ_2-SVM. The key is to compute the upper bound of $|\boldsymbol{\theta}^\top \hat{\mathbf{f}}|$ for features. A feature can be safely removed if its upper bound value is less than 1. The cost of computing the upper bounds can be much lower than training ℓ_1-regularized ℓ_2-SVM. Therefore, screening can greatly lower the computational cost by removing many inactive features before training. To bound the value of $|\boldsymbol{\theta}^\top \hat{\mathbf{f}}|$, it is necessary to construct a closed convex set $\mathbf{K}$ that contains $\boldsymbol{\theta}$. The upper bound value can be then computed by maximizing $|\boldsymbol{\theta}^\top \hat{\mathbf{f}}|$ over $\mathbf{K}$.

3.1 Constructing the Convex Set K

In the following, Eq. (3) and the variational inequality [9] are used to construct a closed convex set $\mathbf{K}$ to bound $|\boldsymbol{\theta}^\top \hat{\mathbf{f}}|$. Proposition 3.1 introduces the variational inequality for a convex optimization problem.

PROPOSITION 3.1. *Let $\boldsymbol{\theta}^\star$ be an optimal solution of the following convex optimization problem:*

$$\min g(\boldsymbol{\theta}), \quad s.t. \;\; \boldsymbol{\theta} \in \mathbf{K},$$

where g is continuously differentiable and $\mathbf{K}$ is closed and convex. Then the following variational inequality holds:

$$\nabla g\,(\boldsymbol{\theta}^\star)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \geq 0, \quad \forall \boldsymbol{\theta} \in \mathbf{K}.$$

The proof of this proposition can be found in [9].

Given $\lambda_2 < \lambda_{max}$, assume that there is a λ_1, such that $\lambda_{max} \geq \lambda_1 > \lambda_2$ and its corresponding solution $\boldsymbol{\theta}_1$ is known[1]. The reason to introduce λ_1 is that when λ_1 is close to λ_2 and $\boldsymbol{\theta}_1$ is known, $\boldsymbol{\theta}_1$ can be used to construct a tighter convex set that contains $\boldsymbol{\theta}_2$ to bound the value of $|\boldsymbol{\theta}_2^\top \hat{\mathbf{f}}|$.

[1]When $\lambda_1 = \lambda_{max}$, $\boldsymbol{\theta}_1$ is given in Eq. (4).

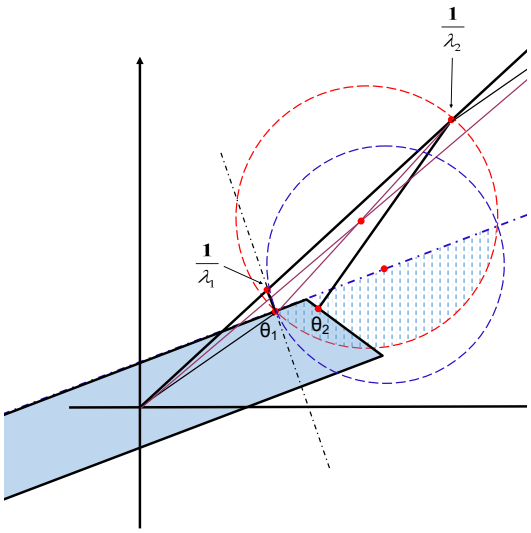

Figure 1: The K in a two-dimensional (2D) space when different t values are used. The red circle corresponds to $t = 0$, and the blue circle corresponds to $t = 1 + \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right) \frac{\mathbf{a}^\top \mathbf{1}}{\left\| \frac{\mathbf{1}}{\lambda_1} - \boldsymbol{\theta}_1 \right\|_2}$.

Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be the optimal solutions of the problem defined in Eq. (3) for λ_1 and λ_2, respectively. Assume that $\lambda_1 > \lambda_2$ and that $\boldsymbol{\theta}_1$ is known. The following results can be obtained by applying Proposition 3.1 to the convex problem defined in Eq. (3) for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively:

$$\left(\boldsymbol{\theta}_1 - \frac{\mathbf{1}}{\lambda_1}\right)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_1) \geq 0, \tag{9}$$

$$\left(\boldsymbol{\theta}_2 - \frac{\mathbf{1}}{\lambda_2}\right)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_2) \geq 0. \tag{10}$$

By substituting $\boldsymbol{\theta} = \boldsymbol{\theta}_2$ into Eq. (9), and $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ into Eq. (10), the following equations can be obtained:

$$\left(\boldsymbol{\theta}_1 - \frac{\mathbf{1}}{\lambda_1}\right)^\top (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \geq 0, \tag{11}$$

$$\left(\boldsymbol{\theta}_2 - \frac{\mathbf{1}}{\lambda_2}\right)^\top (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \leq 0. \tag{12}$$

In the preceding equations, $\boldsymbol{\theta}_1$, λ_1, and λ_2 are known. Therefore, Eq. (11) defines an n-dimensional halfspace and Eq. (12) defines an n-dimensional hyperball. Because $\boldsymbol{\theta}_2$ needs to satisfy both equations, it must reside in the intersection of the halfspace and the hyperball. Obviously, this region is a closed convex set, and it can be used as $\mathbf{K}$ to bound $|\boldsymbol{\theta}_2^\top \hat{\mathbf{f}}|$. Fig. 1 shows an example of the $\mathbf{K}$ in a two-dimensional space. In the figure, Eq. (11) defines the area below the blue line, Eq. (12) defines the area in the red circle, and $\mathbf{K}$ is indicated by the shaded area.

Besides the n-dimensional hyperball defined in Eq. (12), it is possible to derive a series of hyperballs by combining Eq. (11) and Eq. (12). Assume that $\boldsymbol{\theta}^\star$ is the optimal solution of Eq. (3) and $t \geq 0$. It is easy to verify that $\boldsymbol{\theta}^\star$ is also the optimal solution of the following problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \left\| \boldsymbol{\theta} - \left(t\frac{\mathbf{1}}{\lambda} + (1-t)\,\boldsymbol{\theta}^\star \right) \right\|_2^2, \\ s.t. \quad \|\hat{\mathbf{f}}_j^\top \boldsymbol{\theta}\| \leq 1, \quad j = 1, \ldots, m, \\ \sum_{i=1}^{n} \theta_i y_i = 0, \ \boldsymbol{\theta} \succcurlyeq \mathbf{0}. \end{aligned} \tag{13}$$

By applying Proposition 3.1 to the problem defined in Eq. (13) for $\boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$, the following results can be obtained:

$$\left(\boldsymbol{\theta}_1 - \left(t_1\frac{\mathbf{1}}{\lambda_1} + (1-t_1)\,\boldsymbol{\theta}_1\right)\right)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_1) \geq 0, \tag{14}$$

$$\left(\boldsymbol{\theta}_2 - \left(t_2\frac{\mathbf{1}}{\lambda_2} + (1-t_2)\,\boldsymbol{\theta}_2\right)\right)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_2) \geq 0. \tag{15}$$

Let $t = \frac{t_1}{t_2} \geq 0$. By substituting $\boldsymbol{\theta} = \boldsymbol{\theta}_2$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ into Eq. (14) and Eq. (15), respectively, and then combining the two inequalities, the following equations can be obtained:

$$\mathbf{B}_t = \left\{ \boldsymbol{\theta}_2 : (\boldsymbol{\theta}_2 - \mathbf{c})^\top (\boldsymbol{\theta}_2 - \mathbf{c}) \leq l^2 \right\}, \tag{16}$$

$$\mathbf{c} = \frac{1}{2}\left(t\boldsymbol{\theta}_1 - t\frac{\mathbf{1}}{\lambda_1} + \frac{\mathbf{1}}{\lambda_2} + \boldsymbol{\theta}_1\right), l = \frac{1}{2}\left\| t\boldsymbol{\theta}_1 - t\frac{\mathbf{1}}{\lambda_1} + \frac{\mathbf{1}}{\lambda_2} - \boldsymbol{\theta}_1 \right\|_2.$$

As the value of t changes from 0 to ∞, Eq. (16) generates a series of hyperballs. When $t = 0$, $\mathbf{c} = \frac{1}{2}(\frac{\mathbf{1}}{\lambda_2} + \boldsymbol{\theta}_1)$ and $l = \frac{1}{2}\|\frac{\mathbf{1}}{\lambda_2} - \boldsymbol{\theta}_1\|_2$. This corresponds to the hyperball defined by Eq. (12). The following theorems provide some insights about the properties of the hyperballs generated by Eq. (16):

THEOREM 3.2. *Let* $\mathbf{a} = \frac{\frac{\mathbf{1}}{\lambda_1} - \boldsymbol{\theta}_1}{\left\| \frac{\mathbf{1}}{\lambda_1} - \boldsymbol{\theta}_1 \right\|_2}$. *The radius of the hyperball generated by Eq. (16) reaches it minimum when*

$$t = 1 + \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right) \frac{\mathbf{a}^\top \mathbf{1}}{\left\| \frac{\mathbf{1}}{\lambda_1} - \boldsymbol{\theta}_1 \right\|_2}.$$

Let $\hat{\mathbf{c}}$ *be the center of the ball and* l *be the radius. When the minimum is reached, they can be computed as:*

$$\hat{\mathbf{c}} = \frac{1}{2}\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right) P_{\mathbf{a}}(\mathbf{1}) + \boldsymbol{\theta}_1, \ l = \frac{1}{2}\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right) \|P_{\mathbf{a}}(\mathbf{1})\|.$$

Here, $P_{\mathbf{u}}(\mathbf{v}) = \mathbf{v} - \frac{\mathbf{v}^\top \mathbf{u}}{\|\mathbf{u}\|_2^2}\mathbf{u}$ *is an operator that projects* $\mathbf{v}$ *to the null-space of* $\mathbf{u}$. *Since* $\|\mathbf{a}\|_2 = 1$, $P_{\mathbf{a}}(\mathbf{1}) = \mathbf{1} - (\mathbf{a}^\top \mathbf{1})\,\mathbf{a}$.

PROOF. The theorem can be proved by minimizing the r defined in Eq. (16). □

THEOREM 3.3. *Let the intersection of the hyperplane* $\left(\boldsymbol{\theta}_1 - \frac{\mathbf{1}}{\lambda_1}\right)^\top (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) = 0$ *and the hyperball defined by Eq. (16) be* $\mathbf{P}_t$. *The following equation holds:*

$$\mathbf{P}_{t_1} = \mathbf{P}_{t_2}, \ for \ \ \forall t_1, t_2 \geq 0, t_1 \neq t_2.$$

PROOF. The theorem can be proved by showing that $\mathbf{P}_t$ is independent of t. □

This theorem shows that the intersection between the hyperball $\mathbf{B}_t$ and the hyperplane $(\boldsymbol{\theta}_1 - \frac{\mathbf{1}}{\lambda_1})^\top (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) = 0$ is the same for different t values.

THEOREM 3.4. *Let the intersection of the halfspace* $(\boldsymbol{\theta}_1-\frac{\mathbf{1}}{\lambda_1})^\top(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1)\geq 0$ *and the hyperball defined by Eq. (16) be* $\mathbf{Q}_t$*. The following inequality holds:*

$$\mathbf{Q}_{t_1}\subseteq\mathbf{Q}_{t_2},\ \textit{for}\ \ \forall t_1,t_2\geq 0, t_1\leq t_2.$$

PROOF. The theorem can be proved by showing that $\forall t_1$, $t_2\geq 0$ and $t_1\leq t_2$, if $\boldsymbol{\theta}_2\in\mathbf{Q}_{t_1}$, then $\boldsymbol{\theta}_2\in\mathbf{Q}_{t_2}$ also holds. □

This theorem shows that the volume of $\mathbf{Q}_t$ becomes larger when t becomes larger. And $\mathbf{Q}_{t_1}\subseteq\mathbf{Q}_{t_2}$ if $t_1\leq t_2$.

Fig. 1 shows two circles in a 2D space. The red circle corresponds to the one obtained by setting $t_1=0$ in Eq. (16). The blue circle corresponds to the one obtained by setting

$$t_2=1+\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\frac{\mathbf{a}^\top\mathbf{1}}{\left\|\frac{\mathbf{1}}{\lambda_1}-\boldsymbol{\theta}_1\right\|_2}$$

in Eq. (16). It can be observed that the intersections of the two circles pass the line $(\boldsymbol{\theta}_1-\frac{\mathbf{1}}{\lambda_1})^\top(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1)=0$. This is consistent with Theorem 3.3. Also since $t_1\leq t_2$, $\mathbf{Q}_{t_1}\subseteq\mathbf{Q}_{t_2}$, which is consistent with Theorem 3.4.

Thereom 3.4 suggests that $\mathbf{Q}_{t=0}$ should be used to construct $\mathbf{K}$, because when $t=0$, the volume of $\mathbf{Q}_t$ is minimized. The equality $\boldsymbol{\theta}^\top\mathbf{y}=0$ in Eq. (3) of the dual formulation can also be used to further reduce the volume of $\mathbf{K}$.

$$\mathbf{K}=\left\{\boldsymbol{\theta}_2:(\boldsymbol{\theta}_2-\mathbf{c})^\top(\boldsymbol{\theta}_2-\mathbf{c})\leq l^2,\right.$$
$$\left.\left(\boldsymbol{\theta}_1-\frac{\mathbf{1}}{\lambda_1}\right)^\top(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1)\geq 0,\ \boldsymbol{\theta}_2^\top\mathbf{y}=0\right\}.$$

Here $\mathbf{c}=\frac{1}{2}\left(\frac{\mathbf{1}}{\lambda_2}+\boldsymbol{\theta}_1\right)$ and $l=\frac{1}{2}\left\|\frac{\mathbf{1}}{\lambda_2}-\boldsymbol{\theta}_1\right\|_2$. Let $\boldsymbol{\theta}_2=\mathbf{c}+\mathbf{r}$, $\mathbf{a}=\frac{\frac{\mathbf{1}}{\lambda_1}-\boldsymbol{\theta}_1}{\left\|\frac{\mathbf{1}}{\lambda_1}-\boldsymbol{\theta}_1\right\|_2}$, and $\mathbf{b}=\frac{1}{2}\left(\frac{\mathbf{1}}{\lambda_2}-\boldsymbol{\theta}_1\right)$. $\mathbf{K}$ can be written as:

$$\mathbf{K}=\left\{\boldsymbol{\theta}_2:\boldsymbol{\theta}_2=\mathbf{c}+\mathbf{r},\ \|\mathbf{r}\|^2\leq\|\mathbf{b}\|^2,\right.$$
$$\left.\mathbf{a}^\top(\mathbf{b}+\mathbf{r})\leq 0,\ (\mathbf{c}+\mathbf{r})^\top\mathbf{y}=0\right\}. \quad (17)$$

Theorem 3.3 shows that when t varies, the intersection of the hyperball $\mathbf{B}_t$ and the hyperplane $(\boldsymbol{\theta}_1-\frac{\mathbf{1}}{\lambda_1})^\top(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1)=0$ remains unchanged. This suggests that if the maximum value of $|\boldsymbol{\theta}^\top\hat{\mathbf{f}}|$ is achieved with a $\boldsymbol{\theta}$ that is in this area, no matter which $\mathbf{B}_t$ is used, the maximum value will be the same. This property can be used to simplify the computation. Section 3.2.4 will show that when the maximum value of $|\boldsymbol{\theta}^\top\hat{\mathbf{f}}|$ is achieved with a $\boldsymbol{\theta}$ on the intersection of the hyperball $\mathbf{B}_{t=0}$ and the hyperplane $(\boldsymbol{\theta}_1-\frac{\mathbf{1}}{\lambda_1})^\top(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1)=0$, the computation of the maximum value can be simplified by switching to $\mathbf{B}_t$ with $t=1+\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\frac{\mathbf{a}^\top\mathbf{1}}{\left\|\frac{\mathbf{1}}{\lambda_1}-\boldsymbol{\theta}_1\right\|_2}$.

3.2 Computing the Upper Bound

Given the convex set $\mathbf{K}$ defined in Eq. (17), the maximum value of $\left|\boldsymbol{\theta}_2^\top\hat{\mathbf{f}}\right|$ can be computed by solving the problem:

$$\max\left|(\mathbf{c}+\mathbf{r})^\top\hat{\mathbf{f}}\right| \quad (18)$$
$$s.t.\quad \mathbf{a}^\top(\mathbf{b}+\mathbf{r})\leq 0\ ,\|\mathbf{r}\|^2-\|\mathbf{b}\|^2\leq 0,\ \ (\mathbf{c}+\mathbf{r})^\top\mathbf{y}=0.$$

In Eq. (18), $\boldsymbol{\theta}_2=\mathbf{c}+\mathbf{r}$. Also, $\mathbf{r}$ is unknown, and $\hat{\mathbf{f}}$, $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, and $\mathbf{y}$ are known. Since the following equation holds:

$$\begin{aligned}\max|x| &= \max\{-\min(x),\max(x)\}\\ &= \max\{-\min(x),-\min(-x)\},\end{aligned}$$

$\max\left|(\mathbf{c}+\mathbf{r})^\top\hat{\mathbf{f}}\right|$ can be decomposed to two subproblems:

$$m_1=-\min\boldsymbol{\theta}_2^\top\hat{\mathbf{f}}=-\min\mathbf{r}^\top\hat{\mathbf{f}}-\mathbf{c}^\top\hat{\mathbf{f}} \quad (19)$$
$$s.t.\quad \mathbf{a}^\top(\mathbf{b}+\mathbf{r})\leq 0\ ,\|\mathbf{r}\|^2-\|\mathbf{b}\|^2\leq 0,\ \ (\mathbf{c}+\mathbf{r})^\top\mathbf{y}=0,$$

$$m_2=\max\boldsymbol{\theta}_2^\top\hat{\mathbf{f}}=-\min\mathbf{r}^\top\left(-\hat{\mathbf{f}}\right)+\mathbf{c}^\top\hat{\mathbf{f}} \quad (20)$$
$$s.t.\quad \mathbf{a}^\top(\mathbf{b}+\mathbf{r})\leq 0\ ,\|\mathbf{r}\|^2-\|\mathbf{b}\|^2\leq 0,\ \ (\mathbf{c}+\mathbf{r})^\top\mathbf{y}=0,$$

and

$$\max\left|\boldsymbol{\theta}_2^\top\hat{\mathbf{f}}\right|=\max\left|(\mathbf{c}+\mathbf{r})^\top\hat{\mathbf{f}}\right|=\max(m_1,m_2).$$

Eq. (19) and Eq. (20) suggest that the key to bound $\left|\boldsymbol{\theta}_2^\top\hat{\mathbf{f}}\right|$ is to solve the following problem:

$$\min\mathbf{r}^\top\hat{\mathbf{f}} \quad (21)$$
$$s.t.\quad \mathbf{a}^\top(\mathbf{b}+\mathbf{r})\leq 0,\|\mathbf{r}\|^2-\|\mathbf{b}\|^2\leq 0,(\mathbf{c}+\mathbf{r})^\top\mathbf{y}=0.$$

Its Lagrangian $L(\mathbf{r},\alpha,\beta,\rho)$ can be written as:

$$\mathbf{r}^\top\hat{\mathbf{f}}+\alpha\mathbf{a}^\top(\mathbf{b}+\mathbf{r})+\frac{1}{2}\beta\left(\|\mathbf{r}\|_2^2-\|\mathbf{b}\|_2^2\right)+\rho(\mathbf{c}+\mathbf{r})^\top\mathbf{y}. \quad (22)$$

And the Karush-Kuhn-Tucker (KKT) conditions are:

$$\begin{aligned}
&\text{(dual feasibility)} && \alpha\geq 0,\beta\geq 0, && (23)\\
&\text{(primal feasibility)} && \|\mathbf{r}\|_2^2-\|\mathbf{b}\|_2^2\leq 0, && (24)\\
& && \mathbf{a}^\top(\mathbf{b}+\mathbf{r})\leq 0, && (25)\\
& && (\mathbf{c}+\mathbf{r})^\top\mathbf{y}=0, && (26)\\
&\text{(complementary slackness)} && \alpha\mathbf{a}^\top(\mathbf{b}+\mathbf{r})=0, && (27)\\
& && \beta\left(\|\mathbf{r}\|_2^2-\|\mathbf{b}\|_2^2\right)=0, && (28)\\
&\text{(stationarity)} && \nabla_{\mathbf{r}}L(\mathbf{r},\alpha,\beta,\rho)=0. && (29)
\end{aligned}$$

Since $\|\mathbf{r}\|_2^2\leq\|\mathbf{b}\|_2^2$, the problem specified in Eq. (21) is bounded from below by $-\|\mathbf{b}\|_2\|\mathbf{f}\|_2$. Thus, $\min_{\mathbf{r}}L(\mathbf{r},\alpha,\beta,\rho)$ is also bounded from below.

According to whether the inequality constraints are active, the problem can have different minimum values. This requires a discussion of the following four different cases: **(1)**, $\beta=0$, $\hat{\mathbf{f}}+\alpha\mathbf{a}+\rho\mathbf{y}\neq 0$; **(2)**, $\beta=0$, $\hat{\mathbf{f}}+\alpha\mathbf{a}+\rho\mathbf{y}=0$; **(3)**, $\beta>0$, $\alpha=0$; **(4)**, $\beta>0$, $\alpha>0$. The following sections study these cases in detail.

3.2.1 *The Case* $\beta=0,\ \hat{\mathbf{f}}+\alpha\mathbf{a}+\rho\mathbf{y}\neq\mathbf{0}$

In this case, set $\mathbf{r}=t(\mathbf{f}+\alpha\mathbf{a}+\rho\mathbf{y})$, and let $t\to-\infty$. Then $L(\mathbf{r},\alpha,0,\rho)\to-\infty$. This contradicts the observation that $\min_{\mathbf{r}}L(\mathbf{r},\alpha,\beta,\rho)$ must be bounded from below. So when $\hat{\mathbf{f}}+\alpha\mathbf{a}+\rho\mathbf{y}\neq\mathbf{0}$, β must be positive.

3.2.2 *The Case* $\beta=0,\ \hat{\mathbf{f}}+\alpha\mathbf{a}+\rho\mathbf{y}=\mathbf{0}$

Let $P_{\mathbf{u}}(\mathbf{v})=\mathbf{v}-\frac{\mathbf{v}^\top\mathbf{u}}{\|\mathbf{u}\|_2^2}\mathbf{u}$ be the projection of $\mathbf{v}$ onto the null-space of $\mathbf{u}$. Given $\hat{\mathbf{f}}+\alpha\mathbf{a}+\rho\mathbf{y}=\mathbf{0}$, it is easy to verify that $\alpha P_{\mathbf{y}}(\mathbf{a})=-P_{\mathbf{y}}(\hat{\mathbf{f}})$. This suggests that $\alpha P_{\mathbf{y}}(\mathbf{a})$ and $P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)$ are colinear. Also since $\alpha\geq 0$, the following must

hold:

$$\frac{P_{\mathbf{y}}(\mathbf{a})^{\top} P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)}{\|P_{\mathbf{y}}(\mathbf{a})\| \|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|} = -1.$$

Given $\alpha P_{\mathbf{y}}(\mathbf{a}) = -P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)$, α can be computed by:

$$\alpha = -\frac{P_{\mathbf{y}}(\mathbf{a})^{\top} P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)}{\|P_{\mathbf{y}}(\mathbf{a})\|_2^2} = \frac{\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2}.$$

Similarly, the value of ρ can be computed by:

$$\rho = -\frac{\hat{\mathbf{f}}^{\top}\mathbf{y}}{\|\mathbf{y}\|_2^2} - \alpha\frac{\mathbf{a}^{\top}\mathbf{y}}{\|\mathbf{y}\|_2^2} = -\frac{\hat{\mathbf{f}}^{\top}\mathbf{y}}{\|\mathbf{y}\|_2^2} - \frac{\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2}\frac{\mathbf{a}^{\top}\mathbf{y}}{\|\mathbf{y}\|_2^2}$$

By plugging $\beta = 0$ and the obtained value of α and ρ into Eq. (22), $L(\mathbf{r}, \alpha, 0, \rho)$ can be written as:

$$L(\mathbf{r}, \alpha, 0, \rho) = -\frac{\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2}\mathbf{a}^{\top}\boldsymbol{\theta_1} - \mathbf{c}^{\top}\hat{\mathbf{f}} \tag{30}$$

It can be verified that in this case, all KKT conditions specified in Eq. (23)–Eq. (29) are satisfied. Since the problem defined in Eq. (19) is convex and its domain is also convex, Eq. (30) defines the minimum value of the problem. The following theorem summarizes the result when $\beta = 0$.

THEOREM 3.5. *When* $\frac{P_{\mathbf{y}}(\mathbf{a})^{\top} P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)}{\|P_{\mathbf{y}}(\mathbf{a})\| \|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|} = -1$, $\mathbf{r}^{\top}\hat{\mathbf{f}}$ *achieves its minimum value at* $\beta = 0$*, which can be computed as:*

$$\min_{\mathbf{r}} \mathbf{r}^{\top}\hat{\mathbf{f}} = -\frac{\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2}\mathbf{a}^{\top}\boldsymbol{\theta_1} - \mathbf{c}^{\top}\hat{\mathbf{f}}.$$

And in this case, the value of the dual variables are:

$$\alpha = \frac{\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2},\ \beta = 0,\ \rho = -\frac{\hat{\mathbf{f}}^{\top}\mathbf{y}}{\|\mathbf{y}\|_2^2} - \frac{\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2}\frac{\mathbf{a}^{\top}\mathbf{y}}{\|\mathbf{y}\|_2^2}.$$

Since $\alpha = \frac{\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2} > 0$, the minimum value is achieved on the hyperplane defined by $\mathbf{a}^{\top}(\mathbf{b} + \mathbf{r}) = 0$.

COROLLARY 3.6. *When* $\frac{\left|P_{\mathbf{y}}(\mathbf{a})^{\top} P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\right|}{\|P_{\mathbf{y}}(\mathbf{a})\| \|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|} = 1$, $\mathbf{r}^{\top}\hat{\mathbf{f}}$ *achieves its maximum value at* $\beta = 0$*. In this case* $-\min \boldsymbol{\theta}^{\top}\hat{\mathbf{f}}$ *can be computed as:*

$$-\min \boldsymbol{\theta}_2^{\top}\hat{\mathbf{f}} = -\min \mathbf{r}^{\top}\hat{\mathbf{f}} - \mathbf{c}^{\top}\hat{\mathbf{f}} = \frac{\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2}\mathbf{a}^{\top}\boldsymbol{\theta_1}. \tag{31}$$

$\max \boldsymbol{\theta}_2^{\top}\hat{\mathbf{f}}$ can be computed by replacing $\hat{\mathbf{f}}$ with $-\hat{\mathbf{f}}$ in Eq. (31).

In the computation, $\|\mathbf{P_y}\left(\hat{\mathbf{f}}\right)\|_2$ are independent to λ_1, λ_2, and $\boldsymbol{\theta_1}$. Therefore, it can be precomputed. $\|\mathbf{P_y}(\mathbf{a})\|_2$ and $\mathbf{a}^{\top}\boldsymbol{\theta_1}$ are shared by all features. These properties can be used to accelerate the computation.

3.2.3 The Case: $\beta > 0,\ \alpha = 0$

In this case, since $\beta > 0$ and $\alpha = 0$, the minimum value of $\mathbf{r}^{\top}\hat{\mathbf{f}}$ is achieved on the boundary of the hyperball. In Figure 1, this corresponds to the arc of the red circle under the blue line. Plugging $\alpha = 0$ in Eq. (22) results in:

$$L(\mathbf{r}, 0, \beta, \rho) = \mathbf{r}^{\top}\hat{\mathbf{f}} + \frac{1}{2}\beta\left(\|\mathbf{r}\|_2^2 - \|\mathbf{b}\|_2^2\right) + \rho(\mathbf{c} + \mathbf{r})^{\top}\mathbf{y} \tag{32}$$

The dual function $g(0, \beta, \rho)$ can be obtained by setting

$$\nabla_{\mathbf{r}} L(\mathbf{r}, 0, \beta, \rho) = \hat{\mathbf{f}} + \beta\mathbf{r} + \rho\mathbf{y} = 0 \Rightarrow \mathbf{r} = -\frac{1}{\beta}\left(\hat{\mathbf{f}} + \rho\mathbf{y}\right).$$

Since $\beta > 0$, $\|\mathbf{b}\|_2 = \|\mathbf{r}\|_2$. Therefore β can be written as:

$$\beta = \frac{\|\hat{\mathbf{f}} + \rho\mathbf{y}\|_2}{\|\mathbf{b}\|_2}$$

Plugging the obtained $\mathbf{r}$ and β into $L(\mathbf{r}, 0, \beta, \rho)$ leads to:

$$g(\rho) = \min_{\mathbf{r}} L(\mathbf{r}, 0, \beta, \rho) = -\|\mathbf{b}\|_2\|\hat{\mathbf{f}} + \rho\mathbf{y}\|_2 + \rho\mathbf{c}^{\top}\mathbf{y}. \tag{33}$$

The dual function can be maximized by setting $\frac{\partial g(\rho)}{\partial \rho} = 0$. Also since $\mathbf{b}^{\top}\mathbf{y} = \mathbf{c}^{\top}\mathbf{y}$, the following equation holds:

$$-\|\mathbf{b}\|_2\frac{\rho\mathbf{y}^{\top}\mathbf{y} + \hat{\mathbf{f}}^{\top}\mathbf{y}}{\|\hat{\mathbf{f}} + \rho\mathbf{y}\|_2} + \mathbf{b}^{\top}\mathbf{y} = 0. \tag{34}$$

Squaring both sides of the equation and solving the obtained equation leads to the result:

$$\rho = -\frac{\hat{\mathbf{f}}^{\top}\mathbf{y}}{\mathbf{y}^{\top}\mathbf{y}} \pm \frac{\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|_2}{\|P_{\mathbf{y}}(\mathbf{b})\|_2}\frac{\mathbf{b}^{\top}\mathbf{y}}{\mathbf{y}^{\top}\mathbf{y}}.$$

To obtain this equation, the following facts are used:

$$\mathbf{b}^{\top}\mathbf{b} - \frac{\left(\mathbf{b}^{\top}\mathbf{y}\right)^2}{\mathbf{y}^{\top}\mathbf{y}} = \left\|\mathbf{b} - \frac{\mathbf{b}^{\top}\mathbf{y}}{\mathbf{y}^{\top}\mathbf{y}}\mathbf{y}\right\|_2^2 = \|P_{\mathbf{y}}(\mathbf{b})\|_2^2,$$

$$\hat{\mathbf{f}}^{\top}\hat{\mathbf{f}} - \frac{\left(\hat{\mathbf{f}}^{\top}\mathbf{y}\right)^2}{\mathbf{y}^{\top}\mathbf{y}} = \left\|\hat{\mathbf{f}} - \frac{\hat{\mathbf{f}}^{\top}\mathbf{y}}{\mathbf{y}^{\top}\mathbf{y}}\mathbf{y}\right\|_2^2 = \left\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\right\|_2^2.$$

Since $(\mathbf{c} + \mathbf{r})^{\top}\mathbf{y} = 0$ and $\mathbf{r} = -\frac{1}{\beta}\left(\hat{\mathbf{f}} + \rho\mathbf{y}\right)$, it can be verified that $\beta = \frac{\hat{\mathbf{f}}^{\top}\mathbf{y} + \rho\mathbf{y}^{\top}\mathbf{y}}{\mathbf{c}^{\top}\mathbf{y}}$. To ensure that β is positive, the following equation must hold:

$$\rho = -\frac{\hat{\mathbf{f}}^{\top}\mathbf{y}}{\mathbf{y}^{\top}\mathbf{y}} + \frac{\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|_2}{\|P_{\mathbf{y}}(\mathbf{b})\|_2}\frac{\mathbf{b}^{\top}\mathbf{y}}{\mathbf{y}^{\top}\mathbf{y}}. \tag{35}$$

And in this case, β can be written in the form:

$$\beta = \frac{\|\hat{\mathbf{f}} + \rho\mathbf{y}\|_2}{\|\mathbf{b}\|_2} = \frac{\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|_2}{\|P_{\mathbf{y}}(\mathbf{b})\|_2} \tag{36}$$

To compute $\max_{\rho} g(\rho)$, first, Eq. (34) can be rewritten as:

$$\|\mathbf{b}\|_2\|\hat{\mathbf{f}} + \rho\mathbf{y}\|_2 = \|\mathbf{b}\|_2^2\frac{\rho\mathbf{y}^{\top}\mathbf{y} + \hat{\mathbf{f}}^{\top}\mathbf{y}}{\mathbf{b}^{\top}\mathbf{y}}. \tag{37}$$

Plugging Eq. (35) and Eq. (37) into Eq. (33) leads to:

$$\begin{aligned}\max_{\rho} g(\rho) &= -\|\mathbf{b}\|_2^2\frac{\rho\mathbf{y}^{\top}\mathbf{y} + \hat{\mathbf{f}}^{\top}\mathbf{y}}{\mathbf{b}^{\top}\mathbf{y}} + \rho\mathbf{b}^{\top}\mathbf{y} \\ &= -\|P_{\mathbf{y}}(\mathbf{b})\|_2\left\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\right\|_2 - \frac{\hat{\mathbf{f}}^{\top}\mathbf{y}\mathbf{b}^{\top}\mathbf{y}}{\mathbf{y}^{\top}\mathbf{y}}\end{aligned}$$

Since $\frac{\hat{\mathbf{f}}^{\top}\mathbf{y}\mathbf{b}^{\top}\mathbf{y}}{\mathbf{y}^{\top}\mathbf{y}} = \hat{\mathbf{f}}^{\top}\mathbf{b} - P_{\mathbf{y}}^{\top}(\mathbf{b})P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)$, $\max_{\rho} g(\rho)$ can also be written in the following form:

$$\max_{\rho} g(\rho) = -\|P_{\mathbf{y}}(\mathbf{b})\|_2\left\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\right\|_2 + P_{\mathbf{y}}(\mathbf{b})^{\top}P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right) - \hat{\mathbf{f}}^{\top}\mathbf{b}.$$

It can be verified that in this case, all the KKT conditions specified in Eq. (23)–Eq. (24) and Eq. (26)–Eq.(29) are satisfied. The additional condition for Eq. (25) to be satisfied

can be derived as follows. First setting the derivative of Eq. (22) to be zero leads to the following equation:

$$\mathbf{r} = -\frac{1}{\beta}\left(\hat{\mathbf{f}} + \alpha\mathbf{a} + \rho\mathbf{y}\right)$$

Plugging this equation to $\mathbf{a}^\top(\mathbf{b}+\mathbf{r}) \leq 0$ results in:

$$\alpha \geq \beta\mathbf{a}^\top\mathbf{b} - \mathbf{a}^\top\mathbf{f} - \rho\mathbf{a}^\top\mathbf{y}.$$

If $\beta\mathbf{a}^\top\mathbf{b} - \mathbf{a}^\top\mathbf{f} - \rho\mathbf{a}^\top\mathbf{y} > 0$, $\alpha > 0$ must hold. According to the complementary slackness condition, $\mathbf{a}^\top(\mathbf{b}+\mathbf{r}) = 0$. Therefore, $\alpha = \beta\mathbf{a}^\top\mathbf{b} - \mathbf{a}^\top\mathbf{f} - \rho\mathbf{a}^\top\mathbf{y} > 0$. However, this contradicts the requirement that $\alpha = 0$. On the other hand, if $\beta\mathbf{a}^\top\mathbf{b} - \mathbf{a}^\top\mathbf{f} - \rho\mathbf{a}^\top\mathbf{y} \leq 0$, $\alpha = 0$ must hold. Otherwise, $\alpha > 0$ and $\alpha = \beta\mathbf{a}^\top\mathbf{b} - \mathbf{a}^\top\mathbf{f} - \rho\mathbf{a}^\top\mathbf{y} \leq 0$ form a contradiction. Therefore, to satisfy Eq. (25), the condition $\beta\mathbf{a}^\top\mathbf{b} - \mathbf{a}^\top\mathbf{f} - \rho\mathbf{a}^\top\mathbf{y} \leq 0$ must be true. Plugging Eq. (35) and Eq. (36) in $\beta\mathbf{a}^\top\mathbf{b} - \mathbf{a}^\top\mathbf{f} - \rho\mathbf{a}^\top\mathbf{y} \leq 0$ leads to:

$$\left\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\right\|_2 P_{\mathbf{y}}(\mathbf{a})^\top \left(\frac{P_{\mathbf{y}}(\mathbf{b})}{\|P_{\mathbf{y}}(\mathbf{b})\|_2} - \frac{P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)}{\left\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\right\|_2}\right) \leq 0$$

Under this condition, the KKT condition $\mathbf{a}^\top(\mathbf{b}+\mathbf{r}) \geq 0$ must be satisfied. The following theorem summarizes the result for the case $\beta > 0$ and $\alpha = 0$.

THEOREM 3.7. *When* $P_{\mathbf{y}}(\mathbf{a})^\top \left(\frac{P_{\mathbf{y}}(\mathbf{b})}{\|P_{\mathbf{y}}(\mathbf{b})\|_2} - \frac{P_{\mathbf{y}}(\hat{\mathbf{f}})}{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}\right) \leq 0$, $\mathbf{r}^\top\hat{\mathbf{f}}$ *achieves its minimum value at* $\beta > 0$ *and* $\alpha = 0$*:*

$$\min_{\mathbf{r}} \mathbf{r}^\top\hat{\mathbf{f}} = -\|P_{\mathbf{y}}(\mathbf{b})\|_2 \left\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\right\|_2 + P_{\mathbf{y}}(\mathbf{b})^\top P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right) - \hat{\mathbf{f}}^\top\mathbf{b} \tag{38}$$

In this case, the values of the dual variables are:

$$\alpha = 0, \beta = \frac{\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|_2}{\|P_{\mathbf{y}}(\mathbf{b})\|_2}, \rho = -\frac{\hat{\mathbf{f}}^\top\mathbf{y}}{\mathbf{y}^\top\mathbf{y}} - \frac{\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\|_2}{\|P_{\mathbf{y}}(\mathbf{b})\|_2}\frac{\mathbf{b}^\top\mathbf{y}}{\mathbf{y}^\top\mathbf{y}}. \tag{39}$$

COROLLARY 3.8. *When* $P_{\mathbf{y}}(\mathbf{a})^\top \left(\frac{P_{\mathbf{y}}(\mathbf{b})}{\|P_{\mathbf{y}}(\mathbf{b})\|_2} - \frac{P_{\mathbf{y}}(\hat{\mathbf{f}})}{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}\right) \leq 0$, $\mathbf{r}^\top\hat{\mathbf{f}}$ *achieves its minimum value at* $\beta > 0$ *and* $\alpha = 0$*. And in this case,* $-\min\boldsymbol{\theta}^\top\hat{\mathbf{f}}$ *can be computed as:*

$$\begin{aligned} &-\min\boldsymbol{\theta}_2^\top\hat{\mathbf{f}} = -\min\mathbf{r}^\top\hat{\mathbf{f}} - \mathbf{c}^\top\hat{\mathbf{f}} \\ &= \|P_{\mathbf{y}}(\mathbf{b})\|_2 \left\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\right\|_2 - P_{\mathbf{y}}(\mathbf{b})^\top P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right) - \hat{\mathbf{f}}^\top\boldsymbol{\theta}_1 \end{aligned} \tag{40}$$

$\max\boldsymbol{\theta}_2^\top\hat{\mathbf{f}}$ can be computed by replacing $\hat{\mathbf{f}}$ with $-\hat{\mathbf{f}}$ in Eq. (40).

In the computation, $\left\|P_{\mathbf{y}}\left(\hat{\mathbf{f}}\right)\right\|_2$ does not rely on λ_1, λ_2 and $\boldsymbol{\theta}_1$. Therefore, it can be precomputed. $P_{\mathbf{y}}(\mathbf{a})^\top P_{\mathbf{y}}(\mathbf{b})$ and $\|P_{\mathbf{y}}(\mathbf{b})\|_2$, although rely on λ_1, λ_2 or $\boldsymbol{\theta}_1$, are shared by all features and only need to be computed once. These properties can be used to accelerate computation.

3.2.4 The Case: $\beta > 0,\ \alpha > 0$

In this case, the minimum value of $\mathbf{r}^\top\hat{\mathbf{f}}$ is achieved on the intersection of the boundary of the hyperball and the hyperplane. In Figure 1, this corresponds to the two points on the intersection of the red circle and the blue line. It turns out that when $\beta > 0$ and $\alpha > 0$, deriving a closed form solution for the problem specified in Eq. (19) is difficult. Theorem 3.3 suggests that when the minimum value is achieved on the intersection of the hyperball and the hyperplane, one could switch the hyperball used in Eq. (19) to simplify the computation. It turns out that a closed form solution can be obtained by using the hyperball $\mathbf{B}_t$ with $t = 1 + \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right)\frac{\mathbf{a}^\top\mathbf{1}}{\left\|\frac{\mathbf{1}}{\lambda_1} - \boldsymbol{\theta}_1\right\|_2}$. This corresponds to the hyperball defined in Theorem 3.2. As proved in Theorem 3.3, the intersections of different $\mathbf{B}_t$ and $(\frac{\mathbf{1}}{\lambda_1} - \boldsymbol{\theta}_1)^\top(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) = 0$ are identical. Therefore, switching the hyperball $\mathbf{B}_t$ does not change the maximum value of $|\boldsymbol{\theta}^\top\hat{\mathbf{f}}|$.

When $\mathbf{B}_t$ with $t = 1 + \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right)\frac{\mathbf{a}^\top\mathbf{1}}{\left\|\frac{\mathbf{1}}{\lambda_1} - \boldsymbol{\theta}_1\right\|_2}$ is used and assuming that the minimum is achieved on the intersection of the hyperball and the hyperplane, the problem specified in Eq. (19) can be rewritten as:

$$\begin{aligned} &\arg_{\mathbf{r}}\min\mathbf{r}^\top\hat{\mathbf{f}} \\ &s.t.\quad \mathbf{a}^\top\mathbf{r} = 0,\ \|\mathbf{r}\|_2^2 - l^2 \leq 0,\ (\hat{\mathbf{c}}+\mathbf{r})^\top\mathbf{y} = 0. \end{aligned} \tag{41}$$

And its Lagrangian can be written as:

$$L(\mathbf{r}, \alpha, \beta, \rho) = \mathbf{r}^\top\hat{\mathbf{f}} + \alpha\mathbf{a}^\top\mathbf{r} + \frac{1}{2}\beta\left(\|\mathbf{r}\|_2^2 - l^2\right) + \rho(\hat{\mathbf{c}}+\mathbf{r})^\top\mathbf{y}.$$

In the preceding equation, $\mathbf{c}$ is the center of the hyperfall, and l is the radius of the hyperfall. They are defined as:

$$\hat{\mathbf{c}} = \frac{1}{2}\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right)P_{\mathbf{a}}(\mathbf{1}) + \boldsymbol{\theta}_1,\ \ l = \frac{1}{2}\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right)\|P_{\mathbf{a}}(\mathbf{1})\|.$$

The dual function $g(\alpha, \beta, \rho) = \min_{\mathbf{r}} L(\mathbf{r}, \alpha, \beta, \rho)$ can be obtained by setting $\nabla_{\mathbf{r}} L(\mathbf{r}, \alpha, \beta, \rho) = 0$, which leads to:

$$\mathbf{r} = -\frac{1}{\beta}\left(\hat{\mathbf{f}} + \alpha\mathbf{a} + \rho\mathbf{y}\right).$$

Since $\beta \neq 0$, $\|\mathbf{r}\|_2 = l$. Therefore, β can be written as:

$$\beta = \frac{\|\hat{\mathbf{f}} + \alpha\mathbf{a} + \rho\mathbf{y}\|_2}{l}$$

Since $\alpha \neq 0$, $\mathbf{a}^\top\mathbf{r} = 0$. Therefore, α can be written as:

$$\alpha = -\mathbf{a}^\top\left(\hat{\mathbf{f}} + \rho\mathbf{y}\right)$$

Plugging the obtained $\mathbf{r}$, α, and β into $L(\mathbf{r}, \alpha, \beta, \rho)$ leads to:

$$\begin{aligned} g(\rho) &= \min_{\mathbf{r}} L(\mathbf{r}, \alpha, \beta, \rho) \\ &= -l\|\hat{\mathbf{f}} + \alpha\mathbf{a} + \rho\mathbf{y}\|_2 + \rho\hat{\mathbf{c}}^\top\mathbf{y} \\ &= -l\|\hat{\mathbf{f}} - \mathbf{a}^\top\hat{\mathbf{f}}\mathbf{a} + \rho\mathbf{y} - \mathbf{a}^\top\mathbf{y}\mathbf{a}\|_2 + \rho\hat{\mathbf{c}}^\top\mathbf{y} \\ &= -l\|P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right) + \rho P_{\mathbf{a}}(\mathbf{y})\|_2 + \rho\hat{\mathbf{c}}^\top\mathbf{y}. \end{aligned} \tag{42}$$

$g(\rho)$ can be maximized by setting $\frac{\partial g(\rho)}{\partial\rho} = 0$, which leads to:

$$l\frac{\rho P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y}) + P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)^\top P_{\mathbf{a}}(\mathbf{y})}{\|P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right) + \rho P_{\mathbf{a}}(\mathbf{y})\|_2} = \hat{\mathbf{c}}^\top\mathbf{y}. \tag{43}$$

Squaring both sides of the equation and solving the resulting problem yields a closed-form solution for ρ:

$$\rho = -\frac{P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)^\top P_{\mathbf{a}}(\mathbf{y})}{P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})} \pm \frac{\|P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)\right)\|_2}{\|P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\mathbf{1}))\|_2}\frac{P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{y})}{P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})}.$$

To obtain this equation, the following facts are used:

$$\begin{aligned} \hat{\mathbf{c}}^\top\mathbf{y} &= \frac{1}{2}\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right)P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{1}), \\ l^2 &= \frac{1}{4}\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right)^2 P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{1}). \end{aligned}$$

Since $(\hat{\mathbf{c}}+\mathbf{r})^\top \mathbf{y}=0$, $\left(P_{\mathbf{a}}(\hat{\mathbf{c}})+P_{\mathbf{a}}(\mathbf{r})\right)^\top P_{\mathbf{a}}(\mathbf{y})=0$. It can be verified that $P_{\mathbf{a}}(\mathbf{r})=-\frac{1}{\beta}\left(P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)+\rho P_{\mathbf{a}}(\mathbf{y})\right)$. Therefore, $\beta=\frac{P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)^\top P_{\mathbf{a}}(\mathbf{y})+\rho P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})}{P_{\mathbf{a}}(\hat{\mathbf{c}})^\top P_{\mathbf{a}}(\mathbf{y})}$. To ensure that β is positive, the following equation holds:

$$\rho=-\frac{P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)^\top P_{\mathbf{a}}(\mathbf{y})}{P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})}-\frac{\|P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)\right)\|_2}{\|P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\mathbf{1}))\|_2}\frac{P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{y})}{P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})}$$

In this case, β can be written in the form:

$$\beta=2\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)^{-1}\frac{\|P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)\right)\|_2}{\|P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\mathbf{1}))\|_2}$$

To compute $\max_\rho g(\rho)$, first, Eq. (43) can be rewritten as:

$$l\|P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)+\rho P_{\mathbf{a}}(\mathbf{y})\|_2=l^2\frac{\rho P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})+P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)^\top P_{\mathbf{a}}(\mathbf{y})}{\hat{\mathbf{c}}^\top\mathbf{y}}.$$

By plugging this equation and ρ into Eq. (42), $\max_\rho g(\rho)$ can be written in the following form:

$$\frac{1}{2}\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\left(-\left\|P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)\right)\right\|_2\|P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\mathbf{1}))\|_2-\frac{P_{\mathbf{a}}^\top\left(\hat{\mathbf{f}}\right)^\top P_{\mathbf{a}}(\mathbf{y})P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{y})}{P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})}\right).$$

$$\text{Since }\ P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{f})-\frac{P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)^\top P_{\mathbf{a}}(\mathbf{y})P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{y})}{P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})}=P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\mathbf{1}))^\top P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)\right).$$

$\max_\rho g(\rho)$ can also be written in the form:

$$\frac{1}{2}\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\left(-\left\|P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)\right)\right\|_2\left\|P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}(\mathbf{1})\right)\right\|_2+P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}(\mathbf{1})\right)^\top P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)\right)-P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{f})\right).$$

Theorem 3.9 summarizes the result when $\beta>0$ and $\alpha>0$.

THEOREM 3.9. *When $\mathbf{r}^\top\hat{\mathbf{f}}$ achieves its minimum value at $\beta>0$ and $\alpha>0$, it can be computed as:*

$$\min_{\mathbf{r}}\mathbf{r}^\top\hat{\mathbf{f}}=$$
$$\frac{1}{2}\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\left(-\left\|P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)\right)\right\|_2\left\|P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}(\mathbf{1})\right)\right\|_2+P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}(\mathbf{1})\right)^\top P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)\right)-P_{\mathbf{a}}^\top(\mathbf{1})P_{\mathbf{a}}(\mathbf{f})\right).$$

COROLLARY 3.10. *When $\boldsymbol{\theta}^\top\hat{\mathbf{f}}$ achieves its minimum value at $\beta>0$ and $\alpha>0$, it can be computed as:*

$$-\min\boldsymbol{\theta}_2^\top\hat{\mathbf{f}}=-\min\mathbf{r}^\top\hat{\mathbf{f}}-\hat{\mathbf{c}}^\top\hat{\mathbf{f}}= \tag{44}$$
$$\frac{1}{2}\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\left(\left\|P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)\right)\right\|_2\left\|P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}(\mathbf{1})\right)\right\|_2-P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}(\mathbf{1})\right)^\top P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}\left(\hat{\mathbf{f}}\right)\right)\right)-\hat{\mathbf{f}}^\top\boldsymbol{\theta}_1.$$

$\max\boldsymbol{\theta}_2^\top\hat{\mathbf{f}}$ can be computed by replacing $\hat{\mathbf{f}}$ with $-\hat{\mathbf{f}}$ in Eq. (44).

In the computation, $\left\|P_{P_{\mathbf{a}}(\mathbf{y})}\left(P_{\mathbf{a}}(\mathbf{1})\right)\right\|_2$ is shared by all features and needs to be computed only once. This property can be used to accelerate computation.

3.3 The Screening Algorithm

Algorithm 1 shows the procedure of screening features for ℓ_1-regularized ℓ_2-SVM. Given λ_1, λ_2, and $\boldsymbol{\theta}_1$, the algorithm returns a list $\mathbb{L}$, which contains the indices of the features that are potentially active in the optimal solution that corresponds to λ_2. The algorithm first weights all features using $\mathbf{Y}$ in Line 3. It then computes $\max|\hat{\mathbf{f}}^\top\boldsymbol{\theta}|$ for features in Line 4 and Line 5. If the value is larger than 1, it adds the index of the feature to $\mathbb{L}$ in Line 7. The function $\mathsf{neg\_min}(\hat{\mathbf{f}})$ computes $-\min\boldsymbol{\theta}_2^\top\hat{\mathbf{f}}$. Since $P_{\mathbf{u}}(-\mathbf{v})=-P_{\mathbf{u}}(\mathbf{v})$, the intermediate results computed for $\mathsf{neg\_min}(\hat{\mathbf{f}})$ can be used by $\mathsf{neg\_min}(-\hat{\mathbf{f}})$ to accelerate its computation.

The algorithm needs to be implemented carefully to ensure efficiency. First, each step of the computation needs to be decomposed to many small substeps, so that the intermediate results obtained in the preceding substeps can be used by the following substeps to accelerate computation. Second, the substeps need to be organized and ordered properly so that no redundant computation is performed. It turns out the procedure listed in Algorithm 1 is surprisingly fast. First, $\mathbf{y}^\top\mathbf{1}$, $\mathbf{f}^\top\mathbf{1}$, $\mathbf{f}^\top\mathbf{y}$, and $\mathbf{f}^\top\mathbf{f}$ are independent of $\boldsymbol{\theta}_1$, λ_1, and λ_2. Therefore, they can been precomputed before training, and the cost is $O(mn)$. $\boldsymbol{\theta}_1^\top\mathbf{y}$, $\boldsymbol{\theta}_1^\top\mathbf{1}$, and $\boldsymbol{\theta}_1^\top\boldsymbol{\theta}_1$ are shared by all the features. So they can be computed at the begining of screening, and the cost is $O(n)$. Given these intermediate results, most substeps for computing $\max|\hat{\mathbf{f}}^\top\boldsymbol{\theta}|$ can be obtained in $O(1)$. The only expensive substep is to compute $\boldsymbol{\theta}_1^\top\mathbf{f}$, and its cost is $O(mn)$ for m features. However, when a solver fits a ℓ_1-regularized ℓ_2-SVM model, it might have already computed $\hat{\mathbf{f}}^\top\boldsymbol{\theta}_1$ as an intermediate result for all the features. In this case, $\hat{\mathbf{f}}^\top\boldsymbol{\theta}_1$ can be obtained from the solver for screening features at no cost.

In summary, in the worst case of the proposed procedure, the total computational cost for screening a data set that has m features and n samples is $O(mn)$. And if $\hat{\mathbf{f}}^\top\boldsymbol{\theta}_1$, $\mathbf{f}^\top\mathbf{1}$, $\mathbf{f}^\top\mathbf{y}$, and $\mathbf{f}^\top\mathbf{f}$ can be obtained from the intermediate results generated by the ℓ_1-regularized ℓ_2-SVM solver, the total cost can decrease to just $O(m)$.

4. EMPIRICAL STUDY

The proposed screening method was implemented in the C language and compiled as a library that can be conveniently accessed in a high-level programming language, such as the Python or SAS®. This section evaluates its power for accelerating model selection for ℓ_1-regularized ℓ_2-SVM. Experiments are performed on a Windows Server 2008 R2 with two Intel Xeon® L5530 CPUs and 72GB memory.

Input: $\mathbf{X} \in \mathbb{R}^{n\times m}$, $\mathbf{y} \in \mathbb{R}^n$, λ_1, λ_2, $\boldsymbol{\theta}_1 \in \mathbb{R}^n$.
Output: $\mathbb{L}$, the retained feature list.

$\mathbb{L} = \emptyset$, $i = 1$, $\mathbf{Y} = diag(\mathbf{y})$;
for $i \leq m$ **do**
  $\hat{\mathbf{f}} = \mathbf{Y}\mathbf{f}_i$;
  m_1=neg_min($\hat{\mathbf{f}}$), m_2=neg_min($-\hat{\mathbf{f}}$);
  $m = \max\{m_1, m_2\}$;
  if $m \geq 1$ **then**
    $\mathbb{L} = \mathbb{L} \cup \{i\}$;
  end
  $i = i + 1$;
end
return $\mathbb{L}$;

Function neg_min($\hat{\mathbf{f}}$)
  if $\frac{P_{\mathbf{y}}(\mathbf{a})^\top P_{\mathbf{y}}(\hat{\mathbf{f}})}{\|P_{\mathbf{y}}(\mathbf{a})\|\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|} = -1$ **then**
    compute m using Eq. (31);
    return m;
  end
  if $P_{\mathbf{y}}(\mathbf{a})^\top \left(\frac{P_{\mathbf{y}}(\mathbf{b})}{\|P_{\mathbf{y}}(\mathbf{b})\|_2} - \frac{P_{\mathbf{y}}(\hat{\mathbf{f}})}{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}\right) \leq 0$ **then**
    compute m using Eq. (40);
    return m;
  end
  compute m using Eq. (44);
  return m;
end

Algorithm 1: Screening for ℓ_1-regularized ℓ_2-SVM.

4.1 Experiment Setup

Five benchmark data sets are used in the experiment. One is a microarray data set: gli_85. Three are text data sets: rcv1.binary(rcv1b), real-sim, and news20.binary (news20b). And one is a educational data mining data set: kdd2010 bridge-to-algebra (kddb). The gli_85 data set is downloaded from Gene Expression Omnibus,[2] and the other four data sets are downloaded from the LIBSVM data repository.[3] According to the feature-to-sample ratio (m/n), the five data sets fall into three groups: (1) the $m \gg n$ group, including the gli_85 and news20b data sets; (2) the $m \approx n$ group, including the rcv1b and kddb data sets; and (3) the $m \ll n$ group, including the real-sim data set. Table 1 shows detailed information about the five benchmark data sets.

Table 1: Summary of the benchmark data sets

Data Set	sample (n)	feature (m)	m/n
gli_85	85	22283	262.15
rcv1b	20242	47236	2.33
real-sim	72309	20958	0.29
news20b	19996	1355191	67.77
kddb	19264097	29890095	1.55

[2]www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4412
[3]www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

A solver based on the coordinate gradient descent (cgd) algorithm [13] is implemented in the C language for training the ℓ_1-regularized ℓ_2-SVM model. This solver improves the one that is implemented in the liblinear package [6]. In liblinear, the bias term b is also penalized by the ℓ_1 regularizer and is inactive in most cases. In contrast, the improved one solves the problem specified in Eq. (1) exactly. Therefore, the bias term is not penalized and is alway active. Since $-\hat{\mathbf{f}}^\top\boldsymbol{\theta}_1$ is the gradient on a coordinate, $\hat{\mathbf{f}}^\top\boldsymbol{\theta}_1$ is computed in the solver as an intermediate result. Therefore, in screening, $\hat{\mathbf{f}}^\top\boldsymbol{\theta}_1$ values can be obtained from the solver at no cost.

For each given benchmark data set, the cgd solver is used to fit the ℓ_1-regularized ℓ_2-SVM model along a sequence of 20 λ values: $\left\{\lambda_k = \frac{1}{k}\lambda_{max} - \epsilon, k = 1, \ldots, 20, \epsilon = 10^{-8}\right\}$. When $\lambda = \lambda_{max} - \epsilon$, only one feature is active. Denote n_+ and n_- as the number of positive and negative samples, respectively. And let $\mathbf{m} = \sum_{i=1}^{n}\left(y_i - \frac{n_+ - n_-}{n}\right)\mathbf{x}_i$. This feature corresponds to the largest element in $\mathbf{m}$.

For each given benchmark data set, the cgd solver runs in seven different configurations: (1) In **org**, the solver runs without any accelerating technique. (2) In **warm**, the solver runs with warm-start. In the kth iteration, the $\mathbf{w}_{k-1}$ obtained in the $(k-1)$th iteration is used as the initial $\mathbf{w}_k$ for fitting the model. When λ_k and λ_{k-1} are close, warm-start can effectively speed up training by reducing the number of iterations for the solver to converge. (3) In **shr**, the solver runs with the shrinking strategy. During each iteration of the cgd solver run, if a feature's current weight is 0 and its gradient is very small, the feature is set to be inactive [6]. (4) In **warm_shr**, the solver runs with both warm-start and the shrinking strategy. (5) In **scr**, the solver runs with the screening technique. (6) In **warm_scr**, the solver runs with both warm-start and the screening technique. (7) In **scr_shr**, the solver runs with both the shrinking strategy and the screening technique.

Warm-start and screening are designed to speed up model selection, and shrinking is designed to speed up training. These techniques can be combined for further performance improvement. The main purpose of running the ℓ_1-regularized ℓ_2-SVM solver with different configurations is not only to compare different accelerating techniques, but also to provide a sensitivity study for exploring how these techniques can be combined to achieve the best performance.

Both screening and shrinking reduce computational cost by removing inactive features. Their major differences include the following: (1) Shrinking is performed in each iteration of training to reduce the search space of the solver, whereas screening is performed only once before training. (2) Shrinking is a heuristic method for removing inactive features. Sometimes it might also remove active features; when this happens, recovering the true result leads to extra cost. In contrast, the proposed screening technique is safe, because all the removed features are guaranteed to be inactive the optimal solution. (3) The introduced shrinking technique works only for the cgd solver. In contrast, the proposed screening technique can be applied with any ℓ_1-regularized ℓ_2-SVM solver to speed up model selection. Therefore, the proposed screening technique is more general.

Table 2: Total run time of the ℓ_1-regularized ℓ_2-SVM solver when different combinations of accelerating techniques are used to speed up model selection.

Tech.	gli_85	rcv1b	real-sim	news20b	kddb
org	328.7	17.92	20.81	943.67	9209.06
warm	376.6	10.30	13.48	682.08	7752.80
shr	2.78	4.49	7.25	62.44	3374.53
warm_shr	3.10	2.31	4.32	32.62	2395.45
scr	0.78	3.35	6.67	25.53	1126.05
warm_scr	**0.74**	**1.78**	**4.30**	**17.84**	**831.87**
scr_shr	1.45	4.00	7.14	48.29	2603.94

Table 3: Total number of iterations for the ℓ_1-regularized ℓ_2-SVM solver to converge when different combinations of accelerating techniques are used.

Tech.	gli_85	rcv1b	real-sim	news20b	kddb
org	15535	1062	568	2579	755
warm	14610	615	**373**	**1898**	628
shr	16046	1737	815	4995	2008
warm_shr	14888	713	431	2157	1046
scr	15376	1059	596	2862	843
warm_scr	**14599**	**590**	390	1999	**569**
scr_shr	16150	1695	942	4908	1901

4.2 Results

Table 2 and Table 3 show the results of the total run time and the total number of iterations for the ℓ_1-regularized ℓ_2-SVM solver to converge when different combinations of accelerating techniques are used. The total run time and total number of iterations are obtained by aggregating the time and number of iterations used by the ℓ_1-regularized ℓ_2-SVM solver when it fits models using different regularization parameters. In terms of total running time, screening with warm-start (**warm_scr**) provides the best performance. Compared to **org**, for the $m \gg n$ group, the speed-up ratio is about 445 for the gli_85 data and 53 for the news20b data. For the $m \approx n$ group, the speed-up ratio is about 10 for the rcv1b data and 11 for the kddb data. And for the $m \ll n$ group, the speed-up ratio is about 5 for the real-sim data. The result shows that **warm_scr** is more effective when the number of features is larger than the number of samples. A similar trend is observed on **scr** and **scr_shr**. In terms of the total iteration number, the best performance is achieved by **warm** and **warm_scr**. This suggests that warm-start can effectively speed up convergence by providing a good start point for optimization. A similar trend is observed when **shr** is compared to **warm_shr**.

When **org** is compared to **scr**, the result suggests that the proposed screening technique can significantly improve the performance of the ℓ_1-regularized ℓ_2-SVM solver. This justifies that screening can effectively reduce the computational cost of training by removing most inactive features. When **shr** is compared to **scr**, the result suggests that screening performs faster. This is because shrinking is a heuristic method for removing inactive features. Sometimes, it might remove active features during training. When this happens, violations can be detected by using the KKT conditions for the optimal solution. However, recovering the optimal solution leads to extra cost. This is supported by the observation that with **shr** the solver usually takes more iterations to converge than with **org** and **src**. When **warm** is compared to **scr** and **shr**, the results suggest that removing inactive features for training is more effective than providing a good starting point for optimization.

The results presented in Table 2 and Table 3 suggest that the performance of screening and shrinking can be further improved by combining them with warm-start. This is because warm-start can effectively speed up convergence by providing a good starting point for optimization. However, combining screening with shrinking does not improve the performance of screening because that screening has already removed many inactive features before training is performed. When shrinking is used in training, its benefit for removing inactive features becomes insubstantial and is overwhelmed by the cost of recovering the optimal solution when it accidentally removes active features.

Figure 2 shows detailed information about how different combinations of accelerating techniques perform on the news20b data set when λ decreases from λ_{max} to $\frac{1}{20}\lambda_{max}$. The result shows that screening with warm-start is effective for accelerating and its performance is stable. It also shows that when λ decreases, the proposed screening technique can stably select a small set of features for reducing computational cost. Let k be the number of active features. On the news20b data set, when λ decreases from λ_{max} to $\frac{1}{20}\lambda_{max}$, the proposed screening technique retains about k+430 features for training the ℓ_1-regularized ℓ_2-SVM model. This number is much smaller than the dimensionality of the news20b data set, which is about 1.3 million. Similar trends are also observed on other data sets and are not presented in the paper because of the space limit.

Table 4: Comparison of screening to training time

Tech.	gli_85	rcv1b	real-sim	news20b	kddb
scr					
scr	0.03	0.06	0.04	1.91	41.65
tr	0.75	3.29	6.63	23.63	1084.40
ratio	0.04	0.02	0.01	0.08	0.04
warm_scr					
scr	0.03	0.07	0.04	1.91	41.71
tr	0.70	1.72	4.26	15.93	790.15
ratio	0.05	0.04	0.01	0.12	0.05

Table 4 compares the time used by screening to the time used by training. Notice that for training, the solver uses only the features that are selected by screening. Compared to training, the time used by screening is marginal.

The results presented in this section indicate that the proposed screening technqiue is effective for removing inactive features to improve training efficiency. And with warm-start they form the most powerful combination for accelerating model selection for the ℓ_1-regularized ℓ_2-SVM.

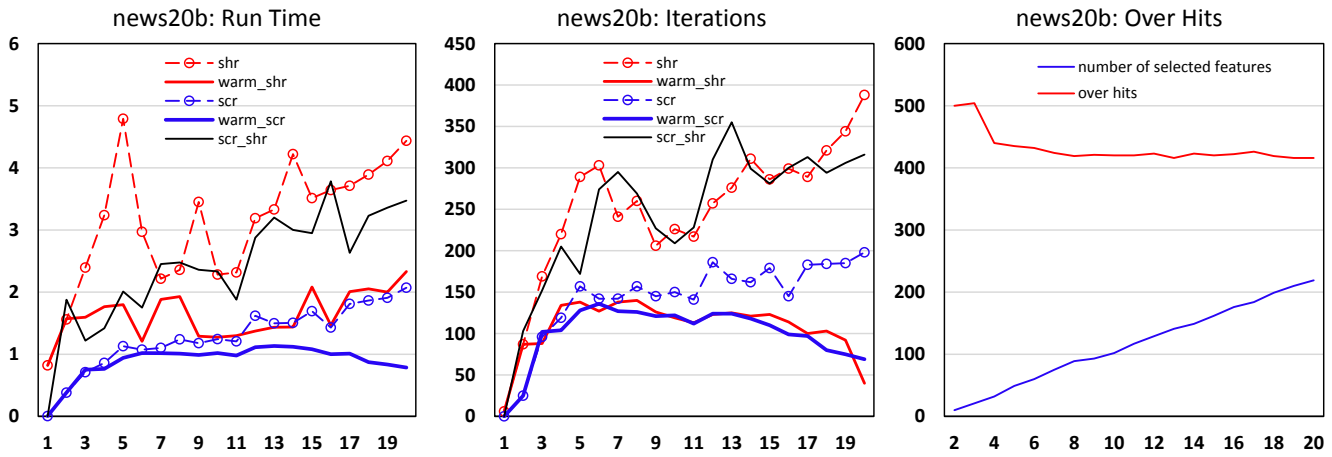

Figure 2: Detailed information about how different combinations of accelerating techniques perform when λ decreases from λ_{max} to $\frac{1}{20}\lambda_{max}$. Results are reported for the news20b data set. "Run time" is the time that is used for training. For scr, warm_scr, and scr_shr, run time includes screening time. "Iterations" is the number of iterations for the solver to converge. "Over hits" is the number of inactive features that are not removed by screening. The results show that the proposed screening technique improves efficiency significantly. And when λ decreases, the number of leftover inactive features is small and stable.

5. CONCLUSION

Screening is an effective technique for improving model selection efficiency by eliminating inactive features. This paper proposes a novel technique to screen features for ℓ_1-regularized ℓ_2-SVM. The key contribution of this paper is the usage of the variational inequality for deriving closed-form criteria to screen features for the ℓ_1-regularized ℓ_2-SVM model in different situations. Empirical study shows that the proposed technique can greatly improve model selection efficiency by stably eliminating a large number of inactive. Our ongoing work will extend the technique to screen features for the ℓ_1-regularized ℓ_1-SVM model.

6. ACKNOWLEDGMENTS

The authors would like to thank Anne Baxter, Russell Albright, and the anonymous reviewers for their valuable suggestions to improve this paper.

7. REFERENCES

[1] M. T. abd Li Wang and I. W. Tsang. Learning sparse svm for feature selection on very high dimensional datasets. In *ICML*, 2010.

[2] J. Bi, M. Embrechts, C. M. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *JMLR*, 3:1229–1243, 2003.

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

[4] P. S. Bradley and L. O. Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, 1998.

[5] E. Candes and M. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25:21–30, 2008.

[6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

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

[8] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[9] J. L. Lions and G. Stampacchia. Variational inequalities. *Communications on Pure and Applied Mathematics*, 20, (3):493–519, 1967.

[10] J. Liu, Z. Zhao, J. Wang, and J. Ye. Safe screening with variational inequalities and its application to lasso. In *ICML*, 2014.

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

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

[13] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117:387–423, 2009.

[14] J. Wang and *et al.* Lasso screening rules via dual polytope projection. In *NIPS*, 2013.

[15] J. Weston, A. Elisseff, B. Schoelkopf, and M. Tipping. Use of the zero norm with linear models and kernel methods. *JMLR*, 3:1439–1461, 2003.

[16] G.-X. Yuan and K.-L. Ma. Scalable training of sparse linear svms. In *ICDM*, 2012.

[17] J. X. Zhen, X. Hao, and J. R. Peter. Learning sparse representations of high dimensional data on large scale dictionaries. In *NIPS*, 2011.

[18] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. In *NIPS*, 2003.