

Safe and Efficient Screening For Sparse Support Vector Machine

Zheng Zhao and Jun Liu

1 Sparse SVM in Primal Form

Assume that $\mathbf{X} \in \mathbb{R}^{m \times n}$ is a data set containing n samples, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, and m features, $\mathbf{X} = (\mathbf{f}_1^\top, \dots, \mathbf{f}_m^\top)^\top$, and $\mathbf{y} = (y_1, \dots, y_n)$ contains the class label of n samples, and $y_i \in \{-1, +1\}$, $i = 1, \dots, n$. The primal form of the L1-regularized L2-Loss support vector machine (SVM) is defined as:

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

Eq. (1) specifies a convex problem with a non-smooth L_1 regularizer, which enforce the solution to be sparse. Let $\mathbf{w}^*(\lambda)$ be the optimal solution of Eq. (1) for a given λ . All the features with nonzero values in $\mathbf{w}^*(\lambda)$ are called active features, and the other features are called inactive.

2 Sparse SVM in Dual

The Lagrangian multiplier [1] of the problem defined in Eq. (1) is:

$$\begin{aligned} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) &= \frac{1}{2} \sum_{i=1}^n \xi_i^2 + \lambda \|\mathbf{w}\|_1 \\ &- \sum_{i=1}^n \alpha_i (y_i (\mathbf{w}^\top \mathbf{x}_i + b) - 1 + \xi_i) \\ &- \sum_{i=1}^n \mu_i \xi_i. \end{aligned} \tag{2}$$

The corresponding Karush-Kuhn-Tucker (KKT) conditions [1] are:

$$\xi_i \geq 0 \quad (3)$$

$$\alpha_i \geq 0 \quad (4)$$

$$\mu_i \geq 0 \quad (5)$$

$$y_i (\mathbf{w}^\top \mathbf{x}_i + b) - 1 + \xi_i \geq 0 \quad (6)$$

$$\alpha_i (y_i (\mathbf{w}^\top \mathbf{x}_i + b) - 1 + \xi_i) = 0 \quad (7)$$

$$\xi_i \mu_i = 0 \quad (8)$$

By defining $L(\mathbf{w})$, $L(\xi_i)$, and $L(b)$ as:

$$L(\mathbf{w}) = \lambda \|\mathbf{w}\|_1 - \sum_{i=1}^n \alpha_i y_i \mathbf{w}^\top \mathbf{x}_i, \quad (9)$$

$$L(\xi_i) = \frac{1}{2} \xi_i^2 - \alpha_i \xi_i - \mu_i \xi_i, \quad (10)$$

$$L(b) = \sum_{i=1}^n \alpha_i y_i b, \quad (11)$$

The Eq. (2) can be reformulated as:

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = L(\mathbf{w}) + \sum_{i=1}^n L(\xi_i) + L(b) + \sum_{i=1}^n \alpha_i. \quad (12)$$

The minimum of $L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu})$ can be expressed as:

$$\inf_{\mathbf{w}, b, \boldsymbol{\xi}} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \inf_{\mathbf{w}} L(\mathbf{w}) + \sum_{i=1}^n \inf_{\xi_i} L(\xi_i) + \inf_b L(b) + \sum_{i=1}^n \alpha_i. \quad (13)$$

Since the problem defined in Eq. (1) is convex and the optimal value of the objective function is achievable, the strong duality condition holds. Therefore, $\inf_{\mathbf{w}} L(\mathbf{w}) < -\infty$, $\inf_{\xi_i} L(\xi_i) < -\infty$, $\inf_b L(b) < -\infty$. By applying standard optimization technique, we can obtain their minimum.

The minimum of $L(\mathbf{w})$

The minimum of $L(\mathbf{w})$ is given by the following equation:

$$\inf_{\mathbf{w}} L(\mathbf{w}) = 0, \quad \text{when } \|\hat{\mathbf{f}}_j^\top \boldsymbol{\alpha}\| \leq \lambda, \quad j = 1, \dots, m. \quad (14)$$

In the preceding equation $\hat{\mathbf{f}}_j = \mathbf{Y} \mathbf{f}_j$, and \mathbf{Y} is a diagonal matrix and $Y_{i,i} = y_i$, $i = 1, \dots, n$. Also, the following equation holds when minimum is achieved:

$$\boldsymbol{\alpha}^\top \hat{\mathbf{f}}_j = \begin{cases} \text{sign}(w_j) \lambda, & \text{if } w_j \neq 0 \\ [-\lambda, +\lambda], & \text{if } w_j = 0 \end{cases} \quad j = 1, \dots, m \quad (15)$$

The minimum of $L(\xi_i)$

The minimum of $L(\xi_i)$ is given by the following equation:

$$\inf_{\xi_i} L(\xi_i) = -\frac{1}{2}\alpha_i^2, \quad \text{when } \xi_i = \alpha_i, \mu_i = 0 \quad i = 1, \dots, n \quad (16)$$

The minimum of $L(b)$

The minimum of $L(b)$ is given by the following equation:

$$\inf_b L(b) = 0, \quad \text{when } \sum_{i=1}^n \alpha_i y_i = 0 \quad (17)$$

The Dual

By substituting Equations (14), (16), and (17) into Eq. (13), the dual of the L1-regularized L2-Loss SVM can be expressed as the following equation:

$$\begin{aligned} & \min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha} - \mathbf{1}\|_2^2 \\ & s.t. \quad \|\hat{\mathbf{f}}_j^\top \boldsymbol{\alpha}\| \leq \lambda, \quad j = 1, \dots, m \\ & \quad \sum_{i=1}^n \alpha_i y_i = 0 \\ & \quad \boldsymbol{\alpha} \succcurlyeq \mathbf{0} \end{aligned} \quad (18)$$

By defining $\boldsymbol{\alpha} = \lambda \boldsymbol{\theta}$, the preceding equation can be reformulated as:

$$\begin{aligned} & \min_{\boldsymbol{\theta}} \|\boldsymbol{\theta} - \frac{\mathbf{1}}{\lambda}\|_2^2 \\ & s.t. \quad \|\hat{\mathbf{f}}_j^\top \boldsymbol{\theta}\| \leq 1, \quad j = 1, \dots, m \\ & \quad \sum_{i=1}^n \theta_i y_i = 0 \\ & \quad \boldsymbol{\theta} \succcurlyeq \mathbf{0} \end{aligned} \quad (19)$$

3 The Relationship between Primal and Dual Variables

In the primal formulation for the L1-regularized L2-loss SVM, the primal variables are b , \mathbf{w} , and $\boldsymbol{\xi}$. And in the dual formulation, the dual variables are $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$. When b and \mathbf{w} is known $\boldsymbol{\xi}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\theta}$ can be obtained as:

$$\mu_i = 0, \quad \xi_i = \alpha_i = \lambda \theta_i = \max(0, 1 - y_i (\mathbf{w}^\top \mathbf{x}_i + b)), \quad i = 1, \dots, n. \quad (20)$$

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

$$\boldsymbol{\alpha}^\top \hat{\mathbf{f}}_j = \begin{cases} \text{sign}(w_j) \lambda, & \text{if } w_j \neq 0 \\ [-\lambda, +\lambda], & \text{if } w_j = 0 \end{cases}, \quad j = 1, \dots, m \quad (21)$$

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

$$\boldsymbol{\theta}^\top \hat{\mathbf{f}}_j = \begin{cases} \text{sign}(w_j), & \text{if } w_j \neq 0 \\ [-1, +1], & \text{if } w_j = 0 \end{cases}, \quad j = 1, \dots, m \quad (22)$$

4 Computing λ_{\max}

λ_{\max} is defined as the smallest value of λ that results $\mathbf{w} = \mathbf{0}$ when it is used in Eq. (1). When the input is given, it can be obtained in a closed form.

The L1-regularized L2-Loss SVM in Eq. (1) can be rewritten in an unconstrained form as:

$$\min h(\mathbf{w}, b) + \lambda \|\mathbf{w}\|_1, \quad (23)$$

where $h(\mathbf{w}, b) = \frac{1}{2} \sum_{i=1}^n \max(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b), 0)^2$. The derivative of $h(\mathbf{w}, b)$ with regard to \mathbf{w} and b can be computed as:

$$h'_{\mathbf{w}}(\mathbf{w}, b) = - \sum_{i=1}^n \max(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b), 0) y_i \mathbf{x}_i \quad (24)$$

$$h'_b(\mathbf{w}, b) = - \sum_{i=1}^n \max(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b), 0) y_i \quad (25)$$

By the definition of λ_{\max} , when λ is larger than λ_{\max} , $\mathbf{w}^* = \mathbf{0}$, therefore,

$$h'_b(\mathbf{0}, b^*) = - \sum_{i=1}^n \max(1 - y_i b^*, 0) y_i = 0,$$

and

$$\|h'_{\mathbf{w}}(\mathbf{0}, b^*)\|_\infty = \left\| \sum_{i=1}^n \max(1 - y_i b^*, 0) y_i \mathbf{x}_i \right\|_\infty \leq \lambda,$$

This leads to the result:

$$b^* = \frac{(n_+ - n_-)}{n},$$

where n_+ and n_- denote the number of positive and negative samples, respectively. Since $\lambda_{\max} = \left\| \sum_{i=1}^n \max(1 - y_i b^*, 0) y_i \mathbf{x}_i \right\|_\infty$. It is easy to verify that $b^* \in [-1, 1]$, thus $\max(1 - y_i b^*, 0) = 1 - y_i b^*$. Therefore,

$$\lambda_{\max} = \left\| \sum_{i=1}^n \left(y_i - \frac{n_+ - n_-}{n} \right) \mathbf{x}_i \right\|_\infty. \quad (26)$$

5 The First Feature(s) to Enter Into the Model

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 corresponding to the element with the largest magnitude in \mathbf{m} .

6 Screening Rule Based on Dual Variable θ

Eq. (22) shows that the necessary condition for a feature \mathbf{f} to be active in the optimal solution is $|\theta^\top \hat{\mathbf{f}}| = 1$, where $\hat{\mathbf{f}} = \mathbf{Y}\mathbf{f}$ and \mathbf{Y} is a diagonal matrix and $Y_{i,i} = y_i$, $i = 1, \dots, n$. This condition can be used to develop a screening rule for the L1-regularized L2-Loss SVM to speedup its training. More specifically, given λ , we can compute the upper bound of the value of $|\theta^\top \hat{\mathbf{f}}|$, and remove all the features with its upper bound values being less than 1, which are guaranteed to be inactive for the given λ . If the cost of computing this upper bound is low, we can use it to speedup the training process by removing many features. To bound value of $|\theta^\top \hat{\mathbf{f}}|$, we need to first construct a closed convex set \mathbf{K} that contains θ . Then we can obtain the upper bound value by maximizing $|\theta^\top \hat{\mathbf{f}}|$ over \mathbf{K} . We first study how to construct the convex set \mathbf{K} .

6.1 Constructing The Convex Set \mathbf{K}

In the following, we construct a closed convex set \mathbf{K} based on Eq. (19) and the variational inequality [2]. We first introduce the variational inequality for convex optimization.

Proposition 6.1. *Let θ be a solution to the optimization problem:*

$$\min g(\theta), \quad \text{s.t. } \theta \in \mathbf{K} \quad (27)$$

where g is continuously differentiable and \mathbf{K} is closed and convex. Then θ^ is a solution of the variational inequality problem:*

$$\nabla g(\theta^*)^\top (\theta - \theta^*) \geq 0, \quad \forall \theta \in \mathbf{K}. \quad (28)$$

The proof of this proposition can be found in [2].

Given $\lambda_2 < \lambda_{max}$, we assume that there is a λ_1 , such that $\lambda_{max} \geq \lambda_1 > \lambda_2$ and its corresponding solution θ_1 is known¹. The reason to introduce λ_1 is that when λ_1 is close to λ_2 and θ_1 is known, this can help us to construct a tighter convex set that contains θ_2 to bound the value of $|\theta_2^\top \hat{\mathbf{f}}|$ in a better way.

Let θ_1 and θ_2 be the optimal solutions of the problem defined in Eq. (19) for λ_1 and λ_2 , respectively. Assume that $\lambda_1 > \lambda_2$, and θ_1 is known. The following results can be obtained by applying Proposition 6.1 to the objective function defined in Eq. (19) for θ_1 and θ_2 , respectively.

$$\left(\theta_1 - \frac{1}{\lambda_1}\right)^\top (\theta - \theta_1) \geq 0 \quad (29)$$

$$\left(\theta_2 - \frac{1}{\lambda_2}\right)^\top (\theta - \theta_2) \geq 0 \quad (30)$$

By substituting $\theta = \theta_2$ into Eq. (29), and $\theta = \theta_1$ into Eq. (30), the following equations can be obtained.

¹ When $\lambda_1 = \lambda_{max}$, θ_1 can be easily obtained by using Eq. (20).

$$\left(\theta_1 - \frac{1}{\lambda_1}\right)^\top (\theta_2 - \theta_1) \geq 0 \quad (31)$$

$$\left(\theta_2 - \frac{1}{\lambda_2}\right)^\top (\theta_2 - \theta_1) \leq 0 \quad (32)$$

In the preceding equations, θ_1 , λ_1 , and λ_2 are known. Therefore, Eq. (31) defines a n dimensional halfspace and Eq. (32) defines a n dimensional hyperball. Since θ_2 needs to satisfy both equations, it must reside in the region formed by the intersection of the halfspace and the hyperball. Obviously, this region is a closed convex set, and can be used as the \mathbf{K} to bound $|\theta_2^\top \hat{\mathbf{f}}|$.

Fig. 1 shows an example of the \mathbf{K} in a two dimensional space. In the figure, $\left(\theta_1 - \frac{1}{\lambda_1}\right)^\top (\theta_2 - \theta_1) = 0$ defines the blue line. And $\left(\theta_2 - \frac{1}{\lambda_2}\right)^\top (\theta_2 - \theta_1) = 0$ defines the red circle. And \mathbf{K} is indicated by the shaded area.

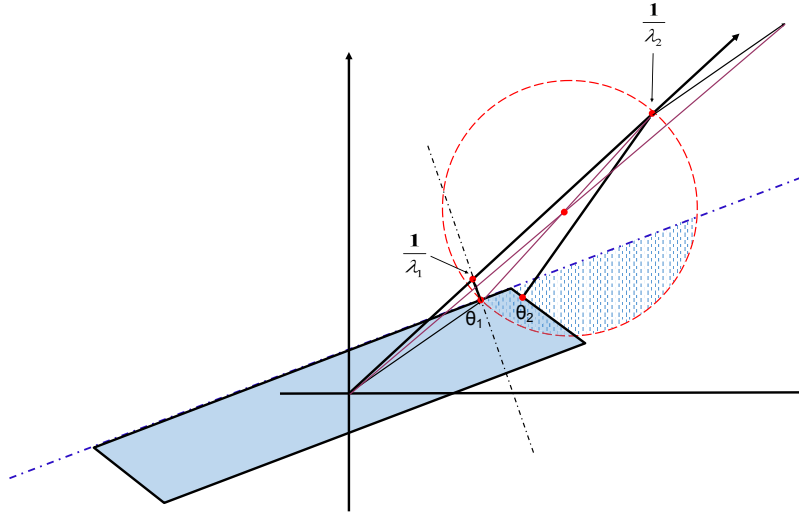


Fig. 1: The \mathbf{K} in a 2D space. It is indicated by the shaded area.

Besides the n dimensional hyperball defined in Eq. (32), it is possible to derive a series of hyperball by combining Eq. (31) and Eq. (32). Assume that θ^* is the optimal solutions of Eq. (19) and $t \geq 0$, it is easy to verify that θ^* is

also the optimal solution of the following problem.

$$\begin{aligned} \min_{\theta} \quad & \left\| \theta - \left(t \frac{\mathbf{1}}{\lambda} + (1-t) \theta^* \right) \right\|_2^2 \\ \text{s.t.} \quad & \|\hat{\mathbf{f}}_j^\top \theta\| \leq 1, \quad j = 1, \dots, m \\ & \sum_{i=1}^n \theta_i y_i = 0 \\ & \theta \succcurlyeq \mathbf{0} \end{aligned} \quad (33)$$

By applying Proposition 6.1 to the objective function defined in Eq. (33) for θ_1 , and θ_2 , the following results can be obtained.

$$\left(\theta_1 - \left(t_1 \frac{\mathbf{1}}{\lambda_1} + (1-t_1) \theta_1 \right) \right)^\top (\theta - \theta_1) \geq 0 \quad (34)$$

$$\left(\theta_2 - \left(t_2 \frac{\mathbf{1}}{\lambda_2} + (1-t_2) \theta_2 \right) \right)^\top (\theta - \theta_2) \geq 0 \quad (35)$$

Let $t = \frac{t_1}{t_2} \geq 0$. By substituting $\theta = \theta_2$ and $\theta = \theta_1$ into Eq. (34) and Eq. (35), respectively, and then combining the two obtained equations, the following equation can be obtained.

$$\begin{aligned} \mathbf{B}_t &= \left\{ \theta_2 : (\theta_2 - \mathbf{c})^\top (\theta_2 - \mathbf{c}) \leq l^2 \right\} \\ \mathbf{c} &= \frac{1}{2} \left(t \theta_1 - t \frac{\mathbf{1}}{\lambda_1} + \frac{\mathbf{1}}{\lambda_2} + \theta_1 \right), \quad l = \frac{1}{2} \left\| t \theta_1 - t \frac{\mathbf{1}}{\lambda_1} + \frac{\mathbf{1}}{\lambda_2} - \theta_1 \right\|_2 \end{aligned} \quad (36)$$

As the value of t change from 0 to ∞ , Eq. (36) generates a series of hyperball. When $t = 0$, $\mathbf{c} = \frac{1}{2} \left(\frac{\mathbf{1}}{\lambda_2} + \theta_1 \right)$ and $l = \frac{1}{2} \left\| \frac{\mathbf{1}}{\lambda_2} - \theta_1 \right\|_2$. This corresponds to the hyperball defined by Eq. (32). The following theorems provide some insights about the properties of the hyperballs generated by Eq. (36).

Theorem 6.2. Let $\mathbf{a} = \frac{\theta_1 - \frac{\mathbf{1}}{\lambda_1}}{\left\| \theta_1 - \frac{\mathbf{1}}{\lambda_1} \right\|_2}$, the radius of the hyperball generated by Eq. (36) reaches it minimum when,

$$t = 1 - \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right) \mathbf{a}^\top \mathbf{1}. \quad (37)$$

Let $\hat{\mathbf{c}}$ be the center of the ball and l be the radius, in this case,

$$\hat{\mathbf{c}} = \frac{1}{2} \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right) P_{\mathbf{a}}(\mathbf{1}) + \theta_1, \quad l = \frac{1}{2} \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right) \|P_{\mathbf{a}}(\mathbf{1})\|. \quad (38)$$

Here, $P_{\mathbf{u}}(\mathbf{v})$ is a operator projects \mathbf{v} to the null-space of \mathbf{u} :

$$P_{\mathbf{u}}(\mathbf{v}) = \mathbf{v} - \frac{\mathbf{v}^\top \mathbf{u}}{\|\mathbf{u}\|_2^2} \mathbf{u}. \quad (39)$$

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. (36). \square

Theorem 6.3. *Let the intersection of the hyperplane $\left(\theta_1 - \frac{1}{\lambda_1}\right)^\top (\theta_2 - \theta_1) = 0$ and the hyperball defined by Eq. (36) be \mathbf{P}_t . The following equation holds.*

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

Proof. The hyperballs defined in Eq. (36) can be rewritten in the form:

$$\mathbf{B}_t = \left\{ \theta_2 : \left(\theta_2 - \frac{1}{\lambda_2} \right)^\top (\theta_2 - \theta_1) - t \left(\theta_1 - \frac{1}{\lambda_1} \right)^\top (\theta_2 - \theta_1) \leq 0 \right\} \quad (40)$$

The intersect between \mathbf{B}_t and $\left(\theta_1 - \frac{1}{\lambda_1}\right)^\top (\theta_2 - \theta_1) = 0$ is:

$$\mathbf{P}_t = \left\{ \theta_2 : \left(\theta_2 - \frac{1}{\lambda_2} \right)^\top (\theta_2 - \theta_1) \text{ and } \left(\theta_1 - \frac{1}{\lambda_1} \right)^\top (\theta_2 - \theta_1) = 0 \right\} \quad (41)$$

Since \mathbf{P}_t is independent to t , we have $\mathbf{P}_{t_1} = \mathbf{P}_{t_2}$, for $\forall t_1, t_2 \geq 0, t_1 \neq t_2$. \square

This theorem shows that the intersection between the hyperball \mathbf{B}_t and the hyperplane $\left(\theta_1 - \frac{1}{\lambda_1}\right)^\top (\theta_2 - \theta_1) = 0$ is the same for different t values.

Theorem 6.4. *Let the intersection of the half space $\left(\theta_1 - \frac{1}{\lambda_1}\right)^\top (\theta_2 - \theta_1) \geq 0$ and the hyperball defined by Eq. (36) be \mathbf{Q}_t . The following inequality holds.*

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

Proof. The intersect between \mathbf{B}_t and $\left(\theta_1 - \frac{1}{\lambda_1}\right)^\top (\theta_2 - \theta_1) \geq 0$ is:

$$\mathbf{Q}_t = \left\{ \theta_2 : \left(\theta_2 - \frac{1}{\lambda_2} \right)^\top (\theta_2 - \theta_1) \leq t \left(\theta_1 - \frac{1}{\lambda_1} \right)^\top (\theta_2 - \theta_1) \right\} \quad (42)$$

Since both t and $\left(\theta_1 - \frac{1}{\lambda_1}\right)^\top (\theta_2 - \theta_1)$ are nonnegative, it is obvious that for $\forall t_1, t_2 \geq 0$ and $t_1 \leq t_2$, if $\theta_2 \in \mathbf{Q}_{t_1}$, we must have $\theta_2 \in \mathbf{Q}_{t_2}$. \square

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

Fig. 2 shows two circles in a 2D space. The circle with red color corresponds to the one obtained by setting $t_1 = 0$ in Eq. (36). And the circle with blue color corresponds to the one obtained by setting $t_2 = 1 - \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right) \mathbf{a}^\top \mathbf{1}$ in Eq. (36). It can be observed in the figure that the intersections of the two circles and the line

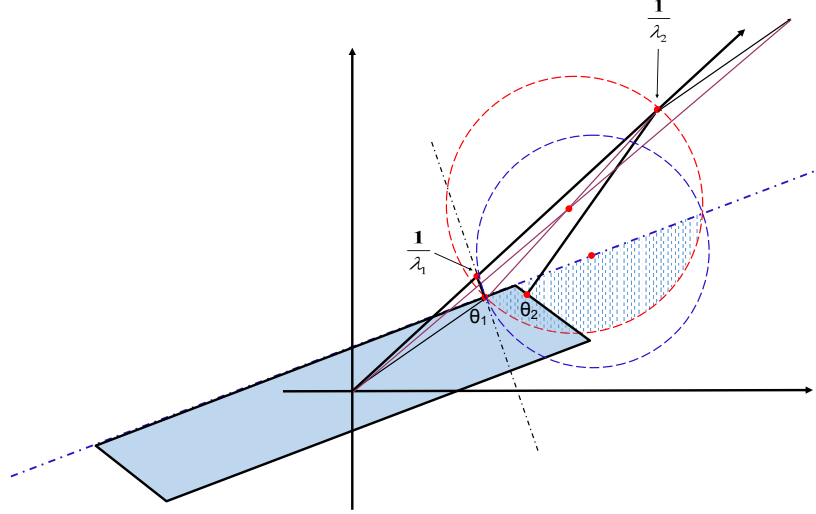


Fig. 2: The \mathbf{K} in a 2D space when different t values are used. The circle with red color corresponds to $t = 0$, and the circle with blue color corresponds to $t = 1 - \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right) \mathbf{a}^\top \mathbf{1}$.

$\left(\theta_1 - \frac{1}{\lambda_1}\right)^\top (\theta_2 - \theta_1) = 0$ are the same, and this is consistent with Theorem 6.3. Also since $t_1 \leq t_2$, $\mathbf{Q}_{t_1} \subseteq \mathbf{Q}_{t_2}$, which is consistent with Theorem 6.4.

Theorem 6.4 suggests to use the $\mathbf{Q}_{t=0}$ to construct \mathbf{K} , since when $t = 0$, the volumn of \mathbf{Q}_t is minimized. The equality $\theta^\top \mathbf{y} = 0$ in Eq. (19) of the dual formulation can also be to further reduce the volumn of \mathbf{K} .

$$\mathbf{K} = \left\{ \theta_2 : (\theta_2 - \mathbf{c})^\top (\theta_2 - \mathbf{c}) \leq l^2, \left(\theta_1 - \frac{1}{\lambda_1}\right)^\top (\theta_2 - \theta_1) \geq 0, \theta_2^\top \mathbf{y} = 0 \right\}$$

$$\text{where, } \mathbf{c} = \frac{1}{2} \left(\frac{1}{\lambda_2} + \theta_1 \right), l = \frac{1}{2} \left\| \frac{1}{\lambda_2} - \theta_1 \right\|_2$$

Let $\theta_2 = \mathbf{c} + \mathbf{r}$, $\mathbf{a} = \frac{\theta_1 - \frac{1}{\lambda_1}}{\left\| \theta_1 - \frac{1}{\lambda_1} \right\|_2}$, and $\mathbf{b} = \frac{1}{2} \left(\frac{1}{\lambda_2} - \theta_1 \right)$, \mathbf{K} can be rewritten as:

$$\mathbf{K} = \left\{ \theta_2 : \theta_2 = \mathbf{c} + \mathbf{r}, \|\mathbf{r}\|^2 \leq \|\mathbf{b}\|^2, \mathbf{a}^\top (\mathbf{b} + \mathbf{r}) \leq 0, (\mathbf{c} + \mathbf{r})^\top \mathbf{y} = 0 \right\} \quad (43)$$

$$\text{where, } \mathbf{a} = \frac{\theta_1 - \frac{1}{\lambda_1}}{\left\| \theta_1 - \frac{1}{\lambda_1} \right\|_2}, \mathbf{b} = \frac{1}{2} \left(\frac{1}{\lambda_2} - \theta_1 \right), \mathbf{c} = \frac{1}{2} \left(\frac{1}{\lambda_2} + \theta_1 \right)$$

Theorem 6.3 shows that when the value of t varies, the intersection of the hyperball \mathbf{B}_t and the hyperplane $\left(\theta_1 - \frac{1}{\lambda_1}\right)^\top (\theta_2 - \theta_1) = 0$ keeps unchange.

This means that if the maximum value of $|\theta^\top \hat{\mathbf{f}}|$ is achieved with a θ 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. In Section 6.6, we will show that when the maximum value of $|\theta^\top \hat{\mathbf{f}}|$ is achieved with a θ on the intersection of the hyperball $\mathbf{B}_{t=0}$ and the hyperplane $(\theta_1 - \frac{1}{\lambda_1})^\top (\theta_2 - \theta_1) = 0$, we can simplify the computation by switching to \mathbf{B}_t with $t = 1 - (\frac{1}{\lambda_2} - \frac{1}{\lambda_1}) \mathbf{a}^\top \mathbf{1}$, which will enable us to derive a close form solution for the problem.

6.2 Computing the Upper Bound

Given the convex set \mathbf{K} defined in Equation (43), the maximum value of $|\theta_2^\top \hat{\mathbf{f}}| = |(\mathbf{c} + \mathbf{r})^\top \hat{\mathbf{f}}|$ can be computed by solving the following optimization problem:

$$\begin{aligned} & \max \left| (\mathbf{c} + \mathbf{r})^\top \hat{\mathbf{f}} \right| \\ \text{s.t. } & \mathbf{a}^\top (\mathbf{b} + \mathbf{r}) \leq 0, \|\mathbf{r}\|^2 - \|\mathbf{b}\|^2 \leq 0, (\mathbf{c} + \mathbf{r})^\top \mathbf{y} = 0. \end{aligned} \quad (44)$$

In the preceding equation, $\theta = \mathbf{c} + \mathbf{r}$, where \mathbf{r} is the unknown, and $\hat{\mathbf{f}}$, \mathbf{a} , \mathbf{b} , \mathbf{c} , and \mathbf{y} are known. Since the following equation holds:

$$\max |x| = \max \{-\min(x), \max(x)\} = \max \{-\min(x), -\min(-x)\}, \quad (45)$$

$\max |(\mathbf{c} + \mathbf{r})^\top \hat{\mathbf{f}}|$ can be decomposed to the following two sub-problems:

$$\begin{aligned} m_1 &= -\min \theta_2^\top \hat{\mathbf{f}} = -\min \mathbf{r}^\top \hat{\mathbf{f}} - \mathbf{c}^\top \hat{\mathbf{f}} \\ \text{s.t. } & \mathbf{a}^\top (\mathbf{b} + \mathbf{r}) \leq 0, \|\mathbf{r}\|^2 - \|\mathbf{b}\|^2 \leq 0, (\mathbf{c} + \mathbf{r})^\top \mathbf{y} = 0, \end{aligned} \quad (46)$$

$$\begin{aligned} m_2 &= \max \theta^\top \hat{\mathbf{f}} = -\min \theta_2^\top (-\hat{\mathbf{f}}) = -\min \mathbf{r}^\top (-\hat{\mathbf{f}}) - \mathbf{c}^\top (-\hat{\mathbf{f}}) \\ \text{s.t. } & \mathbf{a}^\top (\mathbf{b} + \mathbf{r}) \leq 0, \|\mathbf{r}\|^2 - \|\mathbf{b}\|^2 \leq 0, (\mathbf{c} + \mathbf{r})^\top \mathbf{y} = 0, \end{aligned} \quad (47)$$

and

$$\max \left| \theta_2^\top \hat{\mathbf{f}} \right| = \max \left| (\mathbf{c} + \mathbf{r})^\top \hat{\mathbf{f}} \right| = \max(m_1, m_2). \quad (48)$$

Therefore, our key is to solve the following problem:

$$\begin{aligned} & \min \mathbf{r}^\top \hat{\mathbf{f}} \\ \text{s.t. } & \mathbf{a}^\top (\mathbf{b} + \mathbf{r}) \leq 0, \|\mathbf{r}\|^2 - \|\mathbf{b}\|^2 \leq 0, (\mathbf{c} + \mathbf{r})^\top \mathbf{y} = 0. \end{aligned} \quad (49)$$

Its Lagrangian multiplier can be written as:

$$L(\mathbf{r}, \alpha, \beta, \rho) = \mathbf{r}^\top \hat{\mathbf{f}} + \alpha \mathbf{a}^\top (\mathbf{b} + \mathbf{r}) + \frac{1}{2} \beta (\|\mathbf{r}\|_2^2 - \|\mathbf{b}\|_2^2) + \rho (\mathbf{c} + \mathbf{r})^\top \mathbf{y}. \quad (50)$$

The corresponding Karush-Kuhn-Tucker (KKT) conditions are:

$$\alpha \geq 0, \quad (\text{dual feasibility}) \quad (51)$$

$$\beta \geq 0, \quad (52)$$

$$\|\mathbf{r}\|_2^2 - \|\mathbf{b}\|_2^2 \leq 0, \quad (\text{primal feasibility}) \quad (53)$$

$$\mathbf{a}^\top (\mathbf{b} + \mathbf{r}) \leq 0, \quad (54)$$

$$(\mathbf{c} + \mathbf{r})^\top \mathbf{y} = 0, \quad (55)$$

$$\alpha \mathbf{a}^\top (\mathbf{b} + \mathbf{r}) = 0, \quad (\text{complementary slackness}) \quad (56)$$

$$\beta (\|\mathbf{r}\|_2^2 - \|\mathbf{b}\|_2^2) = 0, \quad (57)$$

$$\nabla_{\mathbf{r}} L(\mathbf{r}, \alpha, \beta, \rho) = 0. \quad (\text{stationarity}) \quad (58)$$

Since the problem specified in Eq. (49) is lower bounded by $-\|\mathbf{b}\|_2\|\mathbf{f}\|_2$, it is clear that $\min_{\mathbf{r}} L(\mathbf{r}, \alpha, \beta, \rho)$ must also be bounded from below. In the following we study the four cases listed below:

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. $\beta > 0, \alpha > 0.$

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

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

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

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

$$\frac{P_{\mathbf{y}}(\mathbf{a})^\top P_{\mathbf{y}}(\hat{\mathbf{f}})}{\|P_{\mathbf{y}}(\mathbf{a})\| \|P_{\mathbf{y}}(\hat{\mathbf{f}})\|} = -1. \quad (59)$$

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

$$\alpha = -\frac{P_{\mathbf{y}}(\mathbf{a})^\top P_{\mathbf{y}}(\hat{\mathbf{f}})}{\|P_{\mathbf{y}}(\mathbf{a})\|_2^2} = \frac{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2}. \quad (60)$$

Similarly, the value of ρ can be computed by:

$$\rho = -\frac{\hat{\mathbf{f}}^\top \mathbf{y}}{\|\mathbf{y}\|_2^2} - \alpha \frac{\mathbf{a}^\top \mathbf{y}}{\|\mathbf{y}\|_2^2} = -\frac{\hat{\mathbf{f}}^\top \mathbf{y}}{\|\mathbf{y}\|_2^2} - \frac{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2} \frac{\mathbf{a}^\top \mathbf{y}}{\|\mathbf{y}\|_2^2} \quad (61)$$

By plugging $\beta = 0$ and the obtained value of α and ρ into Eq. (50), it follows:

$$\begin{aligned} L(\mathbf{r}, \alpha, 0, \rho) &= \alpha \mathbf{a}^\top \mathbf{b} + \mathbf{c}^\top (\rho \mathbf{y}) \\ &= \alpha \mathbf{a}^\top \mathbf{b} + \mathbf{c}^\top (-\hat{\mathbf{f}} - \alpha \mathbf{a}) \\ &= \alpha \mathbf{a}^\top (\mathbf{b} - \mathbf{c}) - \mathbf{c}^\top \hat{\mathbf{f}} \\ &= -\frac{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2} \mathbf{a}^\top \boldsymbol{\theta}_1 - \mathbf{c}^\top \hat{\mathbf{f}} \end{aligned} \quad (62)$$

It can be verified that in this case, all the KKT conditions specified in Eq. (51)-Eq. (58) are all satisfied. Since the problem defined in Eq. (46) is convex with a convex domain, Eq. (62) defines its minimum.

The following theorem summarize the result for the case $\beta = 0$.

Theorem 6.5. When $\frac{P_{\mathbf{y}}(\mathbf{a})^\top P_{\mathbf{y}}(\hat{\mathbf{f}})}{\|P_{\mathbf{y}}(\mathbf{a})\| \|P_{\mathbf{y}}(\hat{\mathbf{f}})\|} = -1$, $\mathbf{r}^\top \hat{\mathbf{f}}$ achieves its minimum value at $\beta = 0$, and this minimum value can be computed as:

$$\min_{\mathbf{r}} \mathbf{r}^\top \hat{\mathbf{f}} = -\frac{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2} \mathbf{a}^\top \boldsymbol{\theta}_1 - \mathbf{c}^\top \hat{\mathbf{f}}. \quad (63)$$

And in this case, we have:

$$\alpha = \frac{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2}, \beta = 0, \rho = -\frac{\hat{\mathbf{f}}^\top \mathbf{y}}{\|\mathbf{y}\|_2^2} - \frac{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2} \frac{\mathbf{a}^\top \mathbf{y}}{\|\mathbf{y}\|_2^2}. \quad (64)$$

In this case, since $\alpha = \frac{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2} > 0$, the minimum value is achieved on the hyperplane defined by $\mathbf{a}^\top (\mathbf{b} + \mathbf{r}) = 0$. To compute Eq. (59) and Eq. (63), $\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2$, $\hat{\mathbf{f}}^\top \mathbf{y}$, $\hat{\mathbf{f}}^\top \mathbf{1}$, $\mathbf{y}^\top \mathbf{y}$, and $\mathbf{y}^\top \mathbf{1}$ are independent to λ_1 , λ_2 and $\boldsymbol{\theta}_1$, therefore, can be precomputed. $\|P_{\mathbf{y}}(\mathbf{a})\|_2$ and $\mathbf{a}^\top \boldsymbol{\theta}_1$ can be shared by all features. These properties can be used to accelerate the computation of the screening rule. For each feature, the only expensive computation is $\hat{\mathbf{f}}^\top \boldsymbol{\theta}_1$, and it can be accelerated by utilizing the sparse structure of $\boldsymbol{\theta}_1$.

Corollary 6.6. When $\frac{|P_{\mathbf{y}}(\mathbf{a})^\top P_{\mathbf{y}}(\hat{\mathbf{f}})|}{\|P_{\mathbf{y}}(\mathbf{a})\| \|P_{\mathbf{y}}(\hat{\mathbf{f}})\|} = 1$, $\mathbf{r}^\top \hat{\mathbf{f}}$ achieve its maximum value at $\beta = 0$, and in this case $-\min \boldsymbol{\theta}^\top \hat{\mathbf{f}}$ can be computed as:

$$-\min \boldsymbol{\theta}_2^\top \hat{\mathbf{f}} = -\min \mathbf{r}^\top \hat{\mathbf{f}} - \mathbf{c}^\top \hat{\mathbf{f}} = \frac{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}{\|P_{\mathbf{y}}(\mathbf{a})\|_2} \mathbf{a}^\top \boldsymbol{\theta}_1. \quad (65)$$

6.5 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, it corresponds to the arc of the red circle under the blue line. By plugging $\alpha = 0$ in Eq. (50), it can be obtained:

$$L(\mathbf{r}, 0, \beta, \rho) = \mathbf{r}^\top \hat{\mathbf{f}} + \frac{1}{2}\beta (\|\mathbf{r}\|_2^2 - \|\mathbf{b}\|_2^2) + \rho(\mathbf{c} + \mathbf{r})^\top \mathbf{y} \quad (66)$$

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

$$\nabla_{\mathbf{r}} L(\mathbf{r}, 0, \beta, \rho) = \hat{\mathbf{f}} + \beta \mathbf{r} + \rho \mathbf{y} = 0 \Rightarrow \mathbf{r} = -\frac{1}{\beta} (\hat{\mathbf{f}} + \rho \mathbf{y}). \quad (67)$$

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

$$\beta = \frac{\|\hat{\mathbf{f}} + \rho \mathbf{y}\|_2}{\|\mathbf{b}\|_2} \quad (68)$$

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

$$g(\rho) = \min_{\mathbf{r}} L(\mathbf{r}, 0, \beta, \rho) = -\|\mathbf{b}\|_2 \|\hat{\mathbf{f}} + \rho \mathbf{y}\|_2 + \rho \mathbf{c}^\top \mathbf{y}. \quad (69)$$

To maximize the dual function, we simply set $\frac{\partial g(\rho)}{\partial \rho} = 0$. Also by noticing that $\mathbf{b}^\top \mathbf{y} = \mathbf{c}^\top \mathbf{y}$, as $\theta_1^\top \mathbf{y} = 0$, the following equation can be obtained:

$$-\|\mathbf{b}\|_2 \frac{\rho \mathbf{y}^\top \mathbf{y} + \hat{\mathbf{f}}^\top \mathbf{y}}{\|\hat{\mathbf{f}} + \rho \mathbf{y}\|_2} + \mathbf{b}^\top \mathbf{y} = 0. \quad (70)$$

Taking square on both sides of the equation and simplifying it, we have:

$$\begin{aligned} 0 &= \rho^2 \mathbf{y}^\top \mathbf{y} \left(\mathbf{b}^\top \mathbf{b} \mathbf{y}^\top \mathbf{y} - (\mathbf{b}^\top \mathbf{y})^2 \right) \\ &\quad - 2\rho \hat{\mathbf{f}}^\top \mathbf{y} \left(-\mathbf{b}^\top \mathbf{b} \mathbf{y}^\top \mathbf{y} + (\mathbf{b}^\top \mathbf{y})^2 \right) \\ &\quad + \mathbf{b}^\top \mathbf{b} \left(\hat{\mathbf{f}}^\top \mathbf{y} \right)^2 - \hat{\mathbf{f}}^\top \hat{\mathbf{f}} (\mathbf{b}^\top \mathbf{y})^2. \end{aligned} \quad (71)$$

Solving the preceding equation leads to the result:

$$\rho = -\frac{\hat{\mathbf{f}}^\top \mathbf{y}}{\mathbf{y}^\top \mathbf{y}} \pm \frac{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2 \mathbf{b}^\top \mathbf{y}}{\|P_{\mathbf{y}}(\mathbf{b})\|_2 \mathbf{y}^\top \mathbf{y}}. \quad (72)$$

To obtain this equation, we used the fact:

$$\mathbf{b}^\top \mathbf{b} - \frac{(\mathbf{b}^\top \mathbf{y})^2}{\mathbf{y}^\top \mathbf{y}} = \left\| \mathbf{b} - \frac{\mathbf{b}^\top \mathbf{y}}{\mathbf{y}^\top \mathbf{y}} \mathbf{y} \right\|_2^2 = \|P_{\mathbf{y}}(\mathbf{b})\|_2^2, \quad (73)$$

$$\hat{\mathbf{f}}^\top \hat{\mathbf{f}} - \frac{(\hat{\mathbf{f}}^\top \mathbf{y})^2}{\mathbf{y}^\top \mathbf{y}} = \left\| \hat{\mathbf{f}} - \frac{\hat{\mathbf{f}}^\top \mathbf{y}}{\mathbf{y}^\top \mathbf{y}} \mathbf{y} \right\|_2^2 = \|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2^2. \quad (74)$$

Since $(\mathbf{c} + \mathbf{r})^\top \mathbf{y} = 0$ and $\mathbf{r} = -\frac{1}{\beta} (\hat{\mathbf{f}} + \rho \mathbf{y})$, we have $\beta = \frac{\hat{\mathbf{f}}^\top \mathbf{y} + \rho \mathbf{y}^\top \mathbf{y}}{\mathbf{c}^\top \mathbf{y}}$. To ensure that β is positive, we must have:

$$\rho = -\frac{\hat{\mathbf{f}}^\top \mathbf{y}}{\mathbf{y}^\top \mathbf{y}} + \frac{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}{\|P_{\mathbf{y}}(\mathbf{b})\|_2} \frac{\mathbf{b}^\top \mathbf{y}}{\mathbf{y}^\top \mathbf{y}}. \quad (75)$$

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

$$\beta = \frac{\|\hat{\mathbf{f}} + \rho \mathbf{y}\|_2}{\|\mathbf{b}\|_2} = \frac{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}{\|P_{\mathbf{y}}(\mathbf{b})\|_2} \quad (76)$$

To compute $\max_{\rho} g(\rho)$, first, we notice that Eq. (70) can be rewritten as:

$$-\|\mathbf{b}\|_2 \frac{\rho \mathbf{y}^\top \mathbf{y} + \hat{\mathbf{f}}^\top \mathbf{y}}{\|\hat{\mathbf{f}} + \rho \mathbf{y}\|_2} + \mathbf{b}^\top \mathbf{y} = 0 \Rightarrow \|\mathbf{b}\|_2 \|\hat{\mathbf{f}} + \rho \mathbf{y}\|_2 = \|\mathbf{b}\|_2^2 \frac{\rho \mathbf{y}^\top \mathbf{y} + \hat{\mathbf{f}}^\top \mathbf{y}}{\mathbf{b}^\top \mathbf{y}}. \quad (77)$$

By plugging Eq. (75) and Eq. (77) into Eq. (69) we have:

$$\begin{aligned} \max_{\rho} g(\rho) &= -\|\mathbf{b}\|_2^2 \frac{\rho \mathbf{y}^\top \mathbf{y} + \hat{\mathbf{f}}^\top \mathbf{y}}{\mathbf{b}^\top \mathbf{y}} + \rho \mathbf{b}^\top \mathbf{y} \\ &= -\|P_{\mathbf{y}}(\mathbf{b})\|_2 \left\| P_{\mathbf{y}}(\hat{\mathbf{f}}) \right\|_2 - \frac{\hat{\mathbf{f}}^\top \mathbf{y} \mathbf{b}^\top \mathbf{y}}{\mathbf{y}^\top \mathbf{y}} \end{aligned} \quad (78)$$

Since $\frac{\hat{\mathbf{f}}^\top \mathbf{y} \mathbf{b}^\top \mathbf{y}}{\mathbf{y}^\top \mathbf{y}} = \hat{\mathbf{f}}^\top \mathbf{b} - P_{\mathbf{y}}^\top(\mathbf{b}) P_{\mathbf{y}}(\hat{\mathbf{f}})$, $\max_{\rho} g(\rho)$ can also be written as:

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

It can be verified that in this case, all the KKT conditions specified in Eq. (51)-Eq. (53) and Eq. (55)-Eq.(58) are satisfied. We still need to study that under which condition Eq. (54) can be satisfied. By setting the derivative of Eq. (50) to be zero, the following equation can be obtained:

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

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

$$\alpha \geq \beta \mathbf{a}^\top \mathbf{b} - \mathbf{a}^\top \hat{\mathbf{f}} - \rho \mathbf{a}^\top \mathbf{y}. \quad (79)$$

If $\beta \mathbf{a}^\top \mathbf{b} - \mathbf{a}^\top \hat{\mathbf{f}} - \rho \mathbf{a}^\top \mathbf{y} > 0$, we must have $\alpha > 0$, according to complementary slackness condition, we have $\mathbf{a}^\top (\mathbf{b} + \mathbf{r}) = 0$. Therefore $\alpha = \beta \mathbf{a}^\top \mathbf{b} - \mathbf{a}^\top \hat{\mathbf{f}} - \rho \mathbf{a}^\top \mathbf{y}$. On the other hand, if $\beta \mathbf{a}^\top \mathbf{b} - \mathbf{a}^\top \hat{\mathbf{f}} - \rho \mathbf{a}^\top \mathbf{y} \leq 0$, we must have $\alpha = 0$. Since, if $\alpha > 0$, we will have $\alpha = \beta \mathbf{a}^\top \mathbf{b} - \mathbf{a}^\top \hat{\mathbf{f}} - \rho \mathbf{a}^\top \mathbf{y} \leq 0$, which forms a contradiction. Therefore, to ensure that Eq. (54) is satisfied, we need to have $\beta \mathbf{a}^\top \mathbf{b} - \mathbf{a}^\top \hat{\mathbf{f}} - \rho \mathbf{a}^\top \mathbf{y} \leq 0$. By plugging the obtained β and ρ , we have:

$$\beta \mathbf{a}^\top \mathbf{b} - \mathbf{a}^\top \hat{\mathbf{f}} - \rho \mathbf{a}^\top \mathbf{y} = \left\| P_{\mathbf{y}}(\hat{\mathbf{f}}) \right\|_2 P_{\mathbf{y}}(\mathbf{a})^\top \left(\frac{P_{\mathbf{y}}(\mathbf{b})}{\|P_{\mathbf{y}}(\mathbf{b})\|_2} - \frac{P_{\mathbf{y}}(\hat{\mathbf{f}})}{\left\| P_{\mathbf{y}}(\hat{\mathbf{f}}) \right\|_2} \right) \quad (80)$$

Therefore, if $P_{\mathbf{y}}(\mathbf{a})^\top \left(\frac{P_{\mathbf{y}}(\mathbf{b})}{\|P_{\mathbf{y}}(\mathbf{b})\|_2} - \frac{P_{\mathbf{y}}(\hat{\mathbf{f}})}{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2} \right) \leq 0$, we must have $\alpha = 0$. And in this case, the KKT condition $\mathbf{a}^\top (\mathbf{b} + \mathbf{r}) \geq 0$ is also satisfied.

The following theorem summarize the result for the case $\beta > 0$, $\alpha = 0$.

Theorem 6.7. *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 $\alpha = 0$:*

$$\min_{\mathbf{r}} \mathbf{r}^\top \hat{\mathbf{f}} = -\|P_{\mathbf{y}}(\mathbf{b})\|_2 \|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2 + P_{\mathbf{y}}(\mathbf{b})^\top P_{\mathbf{y}}(\hat{\mathbf{f}}) - \hat{\mathbf{f}}^\top \mathbf{b} \quad (81)$$

In this case, we have:

$$\alpha = 0, \beta = \frac{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}{\|P_{\mathbf{y}}(\mathbf{b})\|_2}, \rho = -\frac{\hat{\mathbf{f}}^\top \mathbf{y}}{\mathbf{y}^\top \mathbf{y}} - \frac{\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2}{\|P_{\mathbf{y}}(\mathbf{b})\|_2} \frac{\mathbf{b}^\top \mathbf{y}}{\mathbf{y}^\top \mathbf{y}}. \quad (82)$$

Note that in Eq. (81) and Eq. (82), $\hat{\mathbf{f}}^\top \hat{\mathbf{f}}$, $\hat{\mathbf{f}}^\top \mathbf{y}$, $\mathbf{y}^\top \mathbf{y}$, and $\|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2$ does not rely on λ_2 and θ_1 , therefore, can be precomputed. $\mathbf{b}^\top \mathbf{b}$, $\mathbf{b}^\top \mathbf{y}$ and $\|P_{\mathbf{y}}(\mathbf{b})\|_2$, although relying on λ_2 or θ_1 , are shared by all features. These properties can be used to accelerate computation when implementing the screening rule.

Corollary 6.8. *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 $-\min \theta^\top \hat{\mathbf{f}}$ can be computed as:*

$$\begin{aligned} -\min \theta_2^\top \hat{\mathbf{f}} &= -\min \mathbf{r}^\top \hat{\mathbf{f}} - \mathbf{c}^\top \hat{\mathbf{f}} \\ &= \|P_{\mathbf{y}}(\mathbf{b})\|_2 \|P_{\mathbf{y}}(\hat{\mathbf{f}})\|_2 - P_{\mathbf{y}}(\mathbf{b})^\top P_{\mathbf{y}}(\hat{\mathbf{f}}) - \hat{\mathbf{f}}^\top \theta_1 \end{aligned} \quad (83)$$

6.6 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 red points on the intersection of the red circle and the blue line. It turns out that, in the case $\beta > 0$, $\alpha > 0$, deriving a closed form solution for the problem specified in Eq. (46) is not easy. Theorem 6.3 suggests that when the minimum value is achieved on the intersection of the hyperball and the hyperplane, we could switch the hyperball used in Eq. (46) to simplify the computation. Below, we show that a closed form solution can be obtained by using the hyperball \mathbf{B}_t with $t = 1 - \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right) \mathbf{a}^\top \mathbf{1}$. This corresponds to the hyperball defined in Theorem 6.2. As proved in Theorem 6.3, the intersections of different \mathbf{B}_t and $\left(\theta_1 - \frac{1}{\lambda_1} \right)^\top (\theta_2 - \theta_1) = 0$ are identical. Therefore, switching the hyperball \mathbf{B}_t in this case does not change the maximum value of $|\theta^\top \hat{\mathbf{f}}|$.

When \mathbf{B}_t with $t = 1 - \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right) \mathbf{a}^\top \mathbf{1}$ is used and assume that the minimum is achieved on the boundary of the hyperball and the hyperplane, the problem

specified in Eq. (46) can be rewritten as:

$$\begin{aligned} & \arg_{\mathbf{r}} \min \mathbf{r}^\top \hat{\mathbf{f}} \\ \text{s.t. } & \mathbf{a}^\top \mathbf{r} = 0, \quad \|\mathbf{r}\|_2^2 - l^2 \leq 0, \quad (\hat{\mathbf{c}} + \mathbf{r})^\top \mathbf{y} = 0. \end{aligned} \quad (84)$$

And its Lagrangian multiplier can be written as:

$$L(\mathbf{r}, \alpha, \beta, \rho) = \mathbf{r}^\top \hat{\mathbf{f}} + \alpha \mathbf{a}^\top \mathbf{r} + \frac{1}{2} \beta (\|\mathbf{r}\|_2^2 - l^2) + \rho (\hat{\mathbf{c}} + \mathbf{r})^\top \mathbf{y}, \quad (85)$$

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

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

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

$$\nabla_{\mathbf{r}} L(\mathbf{r}, \alpha, \beta, \rho) = \hat{\mathbf{f}} + \alpha \mathbf{a} + \beta \mathbf{r} + \rho \mathbf{y} = 0 \Rightarrow \mathbf{r} = -\frac{1}{\beta} (\hat{\mathbf{f}} + \alpha \mathbf{a} + \rho \mathbf{y}). \quad (86)$$

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

$$\beta = \frac{\|\hat{\mathbf{f}} + \alpha \mathbf{a} + \rho \mathbf{y}\|_2}{l} \quad (87)$$

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

$$\alpha = -\mathbf{a}^\top (\hat{\mathbf{f}} + \rho \mathbf{y}) \quad (88)$$

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

$$\begin{aligned} g(\rho) = \min_{\mathbf{r}} L(\mathbf{r}, \alpha, \beta, \rho) &= -l \|\hat{\mathbf{f}} + \alpha \mathbf{a} + \rho \mathbf{y}\|_2 + \rho \hat{\mathbf{c}}^\top \mathbf{y} \\ &= -l \|\hat{\mathbf{f}} - \mathbf{a}^\top \hat{\mathbf{f}} \mathbf{a} + \rho \mathbf{y} - \mathbf{a}^\top \mathbf{y} \mathbf{a}\|_2 + \rho \hat{\mathbf{c}}^\top \mathbf{y} \\ &= -l \|P_{\mathbf{a}}(\hat{\mathbf{f}}) + \rho P_{\mathbf{a}}(\mathbf{y})\|_2 + \rho \hat{\mathbf{c}}^\top \mathbf{y}. \end{aligned} \quad (89)$$

To maximize $g(\rho)$, we simply set $\frac{\partial g(\rho)}{\partial \rho} = 0$, which leads to the equation:

$$l \frac{\rho P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y}) + P_{\mathbf{a}}(\hat{\mathbf{f}})^\top P_{\mathbf{a}}(\mathbf{y})}{\|P_{\mathbf{a}}(\hat{\mathbf{f}}) + \rho P_{\mathbf{a}}(\mathbf{y})\|_2} = \hat{\mathbf{c}}^\top \mathbf{y}. \quad (90)$$

Take square on both sides of the equation and do some simplification. The following equation can be obtained:

$$\begin{aligned} 0 &= \rho^2 P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y}) \left(\left(P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{y}) \right)^2 - P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{1}) P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y}) \right) \\ &\quad - 2\rho P_{\mathbf{a}}(\hat{\mathbf{f}})^\top P_{\mathbf{a}}(\mathbf{y}) \left(P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{1}) P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y}) - \left(P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{y}) \right)^2 \right) \\ &\quad + \left(P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{y}) \right)^2 P_{\mathbf{a}}(\hat{\mathbf{f}})^\top P_{\mathbf{a}}(\hat{\mathbf{f}}) - \left(P_{\mathbf{a}}(\hat{\mathbf{f}})^\top P_{\mathbf{a}}(\mathbf{y}) \right)^2 P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{1}). \end{aligned}$$

To obtain the preceding equation, we used that fact that

$$\hat{\mathbf{c}}^\top \mathbf{y} = \frac{1}{2} \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right) P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{1}) \text{ and } l^2 = \frac{1}{4} \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right)^2 P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{1}).$$

Solving the problem results a closed form solution for ρ in the following form:

$$\rho = -\frac{P_{\mathbf{a}}(\hat{\mathbf{f}})^\top P_{\mathbf{a}}(\mathbf{y})}{P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})} \pm \frac{\|P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\hat{\mathbf{f}}))\|_2 P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{y})}{\|P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\mathbf{1}))\|_2 P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})} \quad (91)$$

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

$$\rho = -\frac{P_{\mathbf{a}}(\hat{\mathbf{f}})^\top P_{\mathbf{a}}(\mathbf{y})}{P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})} - \frac{\|P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\hat{\mathbf{f}}))\|_2 P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{y})}{\|P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\mathbf{1}))\|_2 P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})} \quad (92)$$

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

$$\beta = \frac{\|P_{\mathbf{a}}(\hat{\mathbf{f}}) + \rho P_{\mathbf{a}}(\mathbf{y})\|_2}{l} = 2 \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right)^{-1} \frac{\|P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\hat{\mathbf{f}}))\|_2}{\|P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\mathbf{1}))\|_2} \quad (93)$$

To compute $\max_{\rho} g(\rho)$, first, we notice that Eq. (90) can be rewritten as:

$$\begin{aligned} \hat{\mathbf{c}}^\top \mathbf{y} &= l \frac{\rho P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y}) + P_{\mathbf{a}}(\hat{\mathbf{f}})^\top P_{\mathbf{a}}(\mathbf{y})}{\|P_{\mathbf{a}}(\hat{\mathbf{f}}) + \rho P_{\mathbf{a}}(\mathbf{y})\|_2} \\ \Rightarrow l \|P_{\mathbf{a}}(\hat{\mathbf{f}}) + \rho P_{\mathbf{a}}(\mathbf{y})\|_2 &= l^2 \frac{\rho P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y}) + P_{\mathbf{a}}(\hat{\mathbf{f}})^\top P_{\mathbf{a}}(\mathbf{y})}{\hat{\mathbf{c}}^\top \mathbf{y}}. \end{aligned} \quad (94)$$

By plugging Eq. (92) and Eq. (94) into Eq. (89) we have:

$$\begin{aligned} \max_{\rho} g(\rho) &= \frac{1}{2} \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right) \left(-\|P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\hat{\mathbf{f}}))\|_2 \|P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\mathbf{1}))\|_2 \right. \\ &\quad \left. - \frac{P_{\mathbf{a}}(\hat{\mathbf{f}})^\top P_{\mathbf{a}}(\mathbf{y}) P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{y})}{P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})} \right). \end{aligned} \quad (95)$$

Since $P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{f}) - \frac{P_{\mathbf{a}}(\hat{\mathbf{f}})^\top P_{\mathbf{a}}(\mathbf{y}) P_{\mathbf{a}}(\mathbf{1})^\top P_{\mathbf{a}}(\mathbf{y})}{P_{\mathbf{a}}(\mathbf{y})^\top P_{\mathbf{a}}(\mathbf{y})} = P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\mathbf{1}))^\top P_{P_{\mathbf{a}}(\mathbf{y})}(P_{\mathbf{a}}(\hat{\mathbf{f}}))$,

Eq. (95) can also be written in the following form:

$$\begin{aligned} \max_{\rho} g(\rho) = & \frac{1}{2} \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right) \left(- \left\| P_{P_{\mathbf{a}}(\mathbf{y})} \left(P_{\mathbf{a}}(\hat{\mathbf{f}}) \right) \right\|_2 \left\| P_{P_{\mathbf{a}}(\mathbf{y})} \left(P_{\mathbf{a}}(\mathbf{1}) \right) \right\|_2 \right. \\ & \left. + P_{P_{\mathbf{a}}(\mathbf{y})} \left(P_{\mathbf{a}}(\mathbf{1}) \right)^{\top} P_{P_{\mathbf{a}}(\mathbf{y})} \left(P_{\mathbf{a}}(\hat{\mathbf{f}}) \right) - P_{\mathbf{a}}(\mathbf{1})^{\top} P_{\mathbf{a}}(\mathbf{f}) \right). \end{aligned}$$

The following theorem summarize the result for the case $\beta > 0$, $\alpha > 0$.

Theorem 6.9. *When $\mathbf{r}^{\top} \hat{\mathbf{f}}$ achieves its minimum value at $\beta > 0$ and $\alpha > 0$, this value can be computed as:*

$$\begin{aligned} \min_{\mathbf{r}} \mathbf{r}^{\top} \hat{\mathbf{f}} = & \frac{1}{2} \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right) \left(- \left\| P_{P_{\mathbf{a}}(\mathbf{y})} \left(P_{\mathbf{a}}(\hat{\mathbf{f}}) \right) \right\|_2 \left\| P_{P_{\mathbf{a}}(\mathbf{y})} \left(P_{\mathbf{a}}(\mathbf{1}) \right) \right\|_2 \right. \\ & \left. + P_{P_{\mathbf{a}}(\mathbf{y})} \left(P_{\mathbf{a}}(\mathbf{1}) \right)^{\top} P_{P_{\mathbf{a}}(\mathbf{y})} \left(P_{\mathbf{a}}(\hat{\mathbf{f}}) \right) - P_{\mathbf{a}}^{\top}(\mathbf{1}) P_{\mathbf{a}}(\mathbf{f}) \right). \quad (96) \end{aligned}$$

Corollary 6.10. *When $\mathbf{r}^{\top} \hat{\mathbf{f}}$ achieves its minimum value at $\beta > 0$ and $\alpha > 0$, the corresponding $-\min \theta^{\top} \hat{\mathbf{f}}$ can be computed as:*

$$\begin{aligned} -\min \theta_2^{\top} \hat{\mathbf{f}} &= -\min \mathbf{r}^{\top} \hat{\mathbf{f}} - \hat{\mathbf{c}}^{\top} \hat{\mathbf{f}} \\ &= \frac{1}{2} \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right) \left(\left\| P_{P_{\mathbf{a}}(\mathbf{y})} \left(P_{\mathbf{a}}(\hat{\mathbf{f}}) \right) \right\|_2 \left\| P_{P_{\mathbf{a}}(\mathbf{y})} \left(P_{\mathbf{a}}(\mathbf{1}) \right) \right\|_2 \right. \\ &\quad \left. - P_{P_{\mathbf{a}}(\mathbf{y})} \left(P_{\mathbf{a}}(\mathbf{1}) \right)^{\top} P_{P_{\mathbf{a}}(\mathbf{y})} \left(P_{\mathbf{a}}(\hat{\mathbf{f}}) \right) - \hat{\mathbf{f}}^{\top} \theta_1 \right). \quad (97) \end{aligned}$$

6.7 The Feature Screening Algorithm

Algorithm 1 shows the procedure of screening features for L1-Regularized L2-Loss Support Vector Machine. Given λ_1 , λ_2 , and θ_1 , the algorithm returns a list \mathbb{L} , which contains the indices of the features that are potential to have nonzero weights when λ_2 is used as the regularization parameter.

For each feature, in Line 3, the algorithm weight the feature using \mathbf{Y} . Then, in Line 4 and Line 5, it computes $\max |\hat{\mathbf{f}}^{\top} \theta|$. If the value is larger than 1, it adds the index of the feature to \mathbb{L} in Line 7. The function $\text{neg_min}(\hat{\mathbf{f}})$ computes $-\min \theta_2^{\top} \hat{\mathbf{f}}$ using the results obtained in the preceding subsections.

Since $P_{\mathbf{u}}(-\mathbf{v}) = -P_{\mathbf{u}}(\mathbf{v})$, it is easy to see that the intermediate results generated when computing $\text{neg_min}(\hat{\mathbf{f}})$ can be used to accelerate the computation of $\text{neg_min}(-\hat{\mathbf{f}})$. Also it is easy to verify that in the worst case, the computational cost for evaluating one feature is $O(n)$. Therefore, to evaluate all m features the total computational cost is $O(m \times n)$.

```

Input:  $\mathbf{X} \in \mathbb{R}^{n \times m}$ ,  $\mathbf{y} \in \mathbb{R}^n$ ,  $\lambda_1$ ,  $\lambda_2$ ,  $\theta_1 \in \mathbb{R}^n$ .
Output:  $\mathbb{L}$ , the kept feature list.
1  $\mathbb{L} = \emptyset$ ,  $i = 1$ ,  $\mathbf{Y} = \text{diag}(\mathbf{y})$ ;
2 for  $i \leq m$  do
3    $\hat{\mathbf{f}} = \mathbf{Y}\mathbf{f}_i$ ;
4    $m_1 = \text{neg\_min}(\hat{\mathbf{f}})$ ,  $m_2 = \text{neg\_min}(-\hat{\mathbf{f}})$ ;
5    $m = \max\{m_1, m_2\}$ ;
6   if  $m \geq 1$  then
7      $\mathbb{L} = \mathbb{L} \cup \{i\}$ ;
8   end
9    $i = i + 1$ ;
10 end
11 return  $\mathbb{L}$ ;

12 Function  $\text{neg\_min}(\hat{\mathbf{f}})$ 
13   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
14     compute  $m$  using Eq. (65);
15     return  $m$ ;
16   end
17   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
18     compute  $m$  using Eq. (83);
19     return  $m$ ;
20   end
21   compute  $m$  using Eq. (97);
22   return  $m$ ;
23 end

```

Algorithm 1: The procedure of screening features for L1-Regularized L2-Loss Support Vector Machine (SVM).

References

- [1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [2] J. L. Lions and G. Stampacchia. Variational inequalities. *Communications on Pure and Applied Mathematics*, 20, (3):493–519, 1967.