Safe Triplet Screening for Distance Metric Learning

Tomoki Yoshida
Nagoya Institute of Technology

Ichiro Takeuchi
Nagoya Institute of Technology
National Institute for Material Science
RIKEN Center for Advanced Intelligence Project

Masayuki Karasuyama
Nagoya Institute of Technology
National Institute for Material Science
Japan Science and Technology Agency

ABSTRACT

We study *safe screening* for metric learning. Distance metric learning can optimize a metric over a set of triplets, each one of which is defined by a pair of same class instances and an instance in a different class. However, the number of possible triplets is quite huge even for a small dataset. Our safe triplet screening identifies triplets which can be *safely* removed from the optimization problem without losing the optimality. Compared with existing safe screening studies, triplet screening is particularly significant because of (1) the huge number of possible triplets, and (2) the semi-definite constraint in the optimization. We derive several variants of screening rules, and analyze their relationships. Numerical experiments on benchmark datasets demonstrate the effectiveness of safe triplet screening.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; • **Mathematics of computing** → *Convex optimization*;

KEYWORDS

metric learning; safe screening; convex optimization

ACM Reference Format:
Tomoki Yoshida, Ichiro Takeuchi, and Masayuki Karasuyama. 2018. Safe Triplet Screening for Distance Metric Learning. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3220037

1 INTRODUCTION

Distance metric learning (e.g., [8, 14, 25, 32]) is a widely accepted technique to acquire the optimal metric from observed data. The most standard problem setting is to learn the following parameterized Mahalanobis distance:

$$d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) := \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{x}_j)},$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are d-dimensional feature vectors, and $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ is a positive semi-definite matrix. Using a better distance metric can provide better prediction performance for a variety of machine learning tasks including classification [32], clustering [34] and ranking [19]. Further, the metric optimization has also attracted wide interest even from recent deep network studies [12, 24].

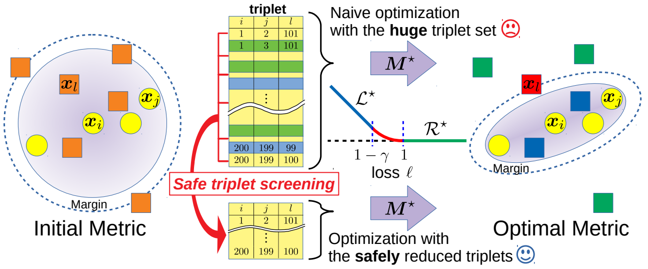


Figure 1: Metric learning with safe triplet screening. The naive optimization needs to minimize the sum of loss function values for a huge number of triplets (i, j, l). Safe triplet screening identifies a subset of $\mathcal{L}^\star$ (blue points in the right drawing) and $\mathcal{R}^\star$ (green points in the right drawing), corresponding to the location of the loss function on which each triplet lies by using the optimal $M^\star$. This enables reducing the number of triplets in the optimization problem.

The seminal work of distance metric learning [32] shows a *triplet* based formulation. A triplet (i, j, l) is defined by a pair of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ which have a same label (same class), and $\boldsymbol{x}_l$ which has a different label (different class). For a triplet (i, j, l), a desirable metric would satisfy $d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) < d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_l)$, meaning that the same class pair is closer than the pair in different classes. For each one of triplets, [32] defines a loss function penalizing violations of this constraint, which has been widely used as a standard approach to metric learning. Although pairwise approaches have also been considered (e.g., [8]), the triplet based loss is the current standard since the relative evaluation $d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) < d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_l)$ is more appropriate for most metric learning application tasks such as nearest neighbor classification [32], and similarity search [13].

However, a set of triplets is quite huge even for a small dataset. For example, considering a two class problem having 100 instances in each class, the number of possible triplets is 1, 980, 000. Since dealing with a huge number of triplets causes prohibitive computations, a small subset of triplets are sometimes used in practice (e.g., [5]) though the optimality of such a sub-sampling strategy is not clearly understood. Our *safe triplet screening* enables the identification of triplets which can be *safely* removed from the optimization problem without losing the optimality of the resulting metric. This

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08...$15.00
https://doi.org/10.1145/3219819.3220037

means that our approach can accelerate the time-consuming metric learning optimization with the optimality guarantee. Figure 1 shows a schematic illustration of safe triplet screening.

Safe screening and its extensions have been actively studied in machine learning community [9–11, 15, 17, 20–23, 27–31, 33, 37–39]. Safe screening is originally proposed for the feature selection by LASSO [10], in which unnecessary features are identified by the following procedure: (**Step 1**) Identifying a bounded region in which the optimal dual solution is guaranteed to exist, and (**Step 2**) For each one of features, verifying possibility to be selected under the condition created by **Step 1**. This procedure is useful to mitigate the optimization difficulty of LASSO for high dimensional problems, and so many papers further propose a variety of approaches to creating bounded regions for obtaining a tighter bound which results in higher screening performance [9, 17, 31, 33]. As another direction of the research, the screening idea has been applied to other learning methods including SVM non-support vector screening [22], nuclear norm regularization subspace screening [38], and group LASSO group screening [21]. To the best of our knowledge, however, no studies have considered screening for metric learning, and our safe triplet screening is particularly significant compared with those exiting studies due to (1) the huge number of possible triplets, and (2) the *semi-definite constraint* on $\boldsymbol{M}$. Our technical contributions are summarized into the following:

- Deriving six sphere regions in which the optimal $\boldsymbol{M}^\star$ must lie based on three different approaches, and analyzing their relationships
- Deriving three types of screening rules, each one of which employs a different approach for the semi-definite constraint
- Building an extension for the *regularization path* calculation

We further demonstrate the effectiveness of our approach based on several benchmark datasets having a huge number of triplets.

Notation

We denote by $[n]$ the set $\{1, 2, \ldots, n\}$ for any integer $n \in \mathbb{N}$. The inner product of the matrices is denoted by $\langle \boldsymbol{A}, \boldsymbol{B} \rangle := \sum_{ij} A_{ij} B_{ij} = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B})$. The squared Frobenius norm is represented by $\|\boldsymbol{A}\|_F^2 := \langle \boldsymbol{A}, \boldsymbol{A} \rangle$. The positive semi-definite matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ is denoted by $\boldsymbol{M} \succeq \boldsymbol{O}$ or $\boldsymbol{M} \in \mathbb{R}_+^{d\times d}$. Through eigenvalue decomposition of matrix $\boldsymbol{M} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, matrices $\boldsymbol{M}_+$ and $\boldsymbol{M}_-$ are defined as follows:

$$\boldsymbol{M} = \boldsymbol{V}\underbrace{(\boldsymbol{\Lambda}_+ + \boldsymbol{\Lambda}_-)}_{\boldsymbol{\Lambda}}\boldsymbol{V}^\top = \underbrace{\boldsymbol{V}\boldsymbol{\Lambda}_+\boldsymbol{V}^\top}_{=:\boldsymbol{M}_+} + \underbrace{\boldsymbol{V}\boldsymbol{\Lambda}_-\boldsymbol{V}^\top}_{=:\boldsymbol{M}_-},$$

where $\boldsymbol{\Lambda}_+$ and $\boldsymbol{\Lambda}_-$ are constructed only by the positive and negative components of the diagonal matrix $\boldsymbol{\Lambda}$. Note that $\langle \boldsymbol{M}_+, \boldsymbol{M}_- \rangle = \mathrm{tr}(\boldsymbol{V}\boldsymbol{\Lambda}_+\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Lambda}_-\boldsymbol{V}^\top) = \mathrm{tr}(\boldsymbol{V}\boldsymbol{O}\boldsymbol{V}^\top) = 0$, and $\boldsymbol{M}_+$ is *projection* of $\boldsymbol{M}$ onto the semi-definite cone, i.e., $\boldsymbol{M}_+ = \mathrm{argmin}_{\boldsymbol{A}\succeq\boldsymbol{O}} \|\boldsymbol{A} - \boldsymbol{M}\|_F^2$.

2 DISTANCE METRIC LEARNING

We first formulate a general form of metric learning problem as a *regularized triplet loss minimization* (RTLM) problem. For later analysis, we derive primal and dual formulations, and to discuss the optimality of the learned metric, we focus on the convex formulation of RTLM in this paper.

2.1 Primal Problem

Let $\{(\boldsymbol{x}_i, y_i) \mid i \in [n]\}$ be n pairs of a d dimensional feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$ and a label $y_i \in \mathcal{Y}$, where $\mathcal{Y}$ is a discrete label space. We consider learning the following Mahalanobis distance:

$$d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) := \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j)},$$

where $\boldsymbol{M} \in \mathbb{R}_+^{d\times d}$ is a positive semi-definite matrix which parameterizes distance. We define a *triplet* of instances as follows:

$$\mathcal{T} = \{(i,j,l) \mid (i,j) \in \mathcal{S},\ y_i \neq y_l,\ l \in [n]\},$$

where $\mathcal{S} = \{(i,j) \mid y_i = y_j,\ i \neq j,\ (i,j) \in [n]\times[n]\}$. The set $\mathcal{S}$ contains index pairs from the same class, and $\mathcal{T}$ represents a triplet of indices consisting of $(i,j) \in \mathcal{S}$, and l which is in a different class from i and j. We refer to the following loss as *triplet loss*:

$$\ell\left(d_{\boldsymbol{M}}^2(\boldsymbol{x}_i, \boldsymbol{x}_l) - d_{\boldsymbol{M}}^2(\boldsymbol{x}_i, \boldsymbol{x}_j)\right), \text{ for } (i,j,l) \in \mathcal{T},$$

where $\ell : \mathbb{R} \to \mathbb{R}$ is some loss function. For the triplet loss, we consider the hinge function $\ell(x) = \max\{0, 1-x\}$, or the smoothed hinge function

$$\ell(x) = \begin{cases} 0, & x > 1, \\ \frac{1}{2\gamma}(1-x)^2, & 1-\gamma \leq x \leq 1, \\ 1 - x - \frac{\gamma}{2}, & x < 1-\gamma, \end{cases}$$

where $\gamma > 0$ is a parameter. Note that the smoothed hinge includes the hinge function as a special case ($\gamma \to 0$). The triplet loss produces a penalty if a pair $(i,j) \in \mathcal{S}$ is more distant than the threshold compared with a pair i and l which are in difference classes. The both of two loss functions contain the "zero part", in which no penalty is imposed, and the "linear part", in which penalty is given linearly. Using the standard squared regularization, we consider the following RTLM as a general form of metric learning:

$$\min_{\boldsymbol{M}\succeq\boldsymbol{O}} P_\lambda(\boldsymbol{M}) := \sum_{ijl} \ell\left(\langle \boldsymbol{M}, \boldsymbol{H}_{ijl} \rangle\right) + \frac{\lambda}{2}\|\boldsymbol{M}\|_F^2, \qquad \text{(Primal)}$$

where $\sum_{ijl}$ denotes $\sum_{(i,j,l)\in\mathcal{T}}$, $\boldsymbol{H}_{ijl} := (\boldsymbol{x}_i - \boldsymbol{x}_l)(\boldsymbol{x}_i - \boldsymbol{x}_l)^\top - (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$, and $\lambda > 0$ is a regularization parameter.

2.2 Dual Problem

The dual problem is written as

$$\max_{\boldsymbol{0}\leq\boldsymbol{\alpha}\leq\boldsymbol{1},\ \boldsymbol{\Gamma}\succeq\boldsymbol{O}} D_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) := -\frac{\gamma}{2}\|\boldsymbol{\alpha}\|_2^2 + \boldsymbol{\alpha}^\top\boldsymbol{1} - \frac{\lambda}{2}\|\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})\|_F^2, \qquad \text{(Dual1)}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{T}|}$, which contains α_{ijl} for $(i,j,l) \in \mathcal{T}$, and $\boldsymbol{\Gamma} \in \mathbb{R}^{d\times d}$ are dual variables, and

$$\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) := \frac{1}{\lambda}\Big[\sum_{ijl} \alpha_{ijl}\boldsymbol{H}_{ijl} + \boldsymbol{\Gamma}\Big]. \qquad (1)$$

We omit the derivation due to the space limitation (see Appendix A [36]). Since the last term $\max_{\boldsymbol{\Gamma}\succeq\boldsymbol{O}} -\frac{1}{2}\|\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})\|_F^2$ is equivalent to the projection onto a semi-definite cone [4, 18], the above problem (Dual1) can be simplified as

$$\max_{\boldsymbol{0}\leq\boldsymbol{\alpha}\leq\boldsymbol{1}} D_\lambda(\boldsymbol{\alpha}) := -\frac{\gamma}{2}\|\boldsymbol{\alpha}\|_2^2 + \boldsymbol{\alpha}^\top\boldsymbol{1} - \frac{\lambda}{2}\left\|\boldsymbol{M}_\lambda(\boldsymbol{\alpha})\right\|_F^2, \qquad \text{(Dual2)}$$

where

$$M_\lambda(\boldsymbol{\alpha}) := \frac{1}{\lambda}\Big[\sum_{ijl}\alpha_{ijl}\boldsymbol{H}_{ijl}\Big]_+.$$

For the optimal $\boldsymbol{M}^\star$, each one of triplets in $\mathcal{T}$ can be categorized into the following three groups:

$$\begin{aligned}\mathcal{L}^\star &:= \{(i,j,l) \in \mathcal{T} \mid \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}^\star\rangle < 1-\gamma\},\\ \mathcal{C}^\star &:= \{(i,j,l) \in \mathcal{T} \mid 1-\gamma \le \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}^\star\rangle \le 1\},\\ \mathcal{R}^\star &:= \{(i,j,l) \in \mathcal{T} \mid \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}^\star\rangle > 1\}.\end{aligned} \tag{2}$$

This indicates that triplets in $\mathcal{R}^\star$ is in "zero part", and triplets in $\mathcal{L}^\star$ is in "linear part" of the loss function. The well-known KKT condition provides the following relation between the optimal dual variable and the derivative of the loss function (see Appendix A [36] for detail):

$$\alpha^\star_{ijl} = \nabla\ell(\langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl}\rangle). \tag{3}$$

Considering this equation, (2), and the definition of the loss function, we obtain the following rules:

$$\begin{aligned}(i,j,l) \in \mathcal{L}^\star &\Rightarrow \alpha^\star_{ijl} = 1,\\ (i,j,l) \in \mathcal{C}^\star &\Rightarrow \alpha^\star_{ijl} \in [0,1],\\ (i,j,l) \in \mathcal{R}^\star &\Rightarrow \alpha^\star_{ijl} = 0.\end{aligned} \tag{4}$$

3 SAFE TRIPLET SCREENING

The nonlinear semi-definite programming problem of RTLM can be solved by the gradient methods including the primal-based [32], or the dual-based approach [26]. However, prohibitive amount of computations can be necessary because of the huge number of triplets. Naive calculation of the objective function requires $O(d^2|\mathcal{T}|)$ computations for both of the primal and the dual cases. Our *safe triplet screening* can reduce the number of triplets by identifying a part of $\mathcal{L}^\star$ and $\mathcal{R}^\star$ before solving the optimization problem.

Let $\hat{\mathcal{L}} \subseteq \mathcal{L}^\star$ and $\hat{\mathcal{R}} \subseteq \mathcal{R}^\star$ be subsets of $\mathcal{L}^\star$ and $\mathcal{R}^\star$, respectively. When we have $\hat{\mathcal{L}}$ and $\hat{\mathcal{R}}$, the optimization problem (Primal) can be transformed into

$$\tilde{P}_\lambda(\boldsymbol{M}) = \sum_{(i,j,l)\in\mathcal{T}-\hat{\mathcal{L}}-\hat{\mathcal{R}}} \ell(\langle \boldsymbol{M}, \boldsymbol{H}_{ijl}\rangle) + \frac{\lambda}{2}\|\boldsymbol{M}\|_F^2 + \left(1-\frac{\gamma}{2}\right)|\hat{\mathcal{L}}| - \langle \boldsymbol{M}, \sum_{(i,j,l)\in\hat{\mathcal{L}}} \boldsymbol{H}_{ijl}\rangle.$$

This problem differs from the original (Primal) as follows

- In the first term, we remove $\hat{\mathcal{R}}$ which does not produce any penalty at the optimal solution
- The loss function for $\hat{\mathcal{L}}$ is fixed at "linear part" of the loss function by which the sum over triplets can be calculated beforehand (the last two terms)

Note that this problem has the same optimal $\boldsymbol{M}^\star$ as the original $P_\lambda(\boldsymbol{M})$. Therefore, if the large number of $\hat{\mathcal{L}}$ and $\hat{\mathcal{R}}$ can be detected beforehand, the metric learning optimization can be accelerated dramatically. In the case of the dual problem, the dual variables α_{ijl} for $(i,j,l) \in \hat{\mathcal{L}}$ and $(i,j,l) \in \hat{\mathcal{R}}$ can be fixed by the rule (4), and the number of variables to be optimized is reduced.

Our safe triplet screening identifies $\hat{\mathcal{L}}$ and $\hat{\mathcal{R}}$ by the following procedure:

Step 1 Identifying a sphere region, in which the optimal solution $\boldsymbol{M}^\star$ must lie, based on a current feasible solution which we call *reference solution*

Step 2 For each one of triplets $(i,j,l) \in \mathcal{T}$, verifying possibility of $(i,j,l) \in \mathcal{L}^\star$ or $(i,j,l) \in \mathcal{R}^\star$ under the condition that $\boldsymbol{M}^\star$ is in the region

In Section 3.1, we first describe **Step 2** of this procedure, and subsequently, we derive sphere shaped regions which must contain $\boldsymbol{M}^\star$, required for **Step 1**, in Section 3.2.

3.1 Screening Rule

Letting $\mathcal{B}$ be a region which contains $\boldsymbol{M}^\star$, the following screening rule can be derived from (2):

$$\max_{\boldsymbol{X}\in\mathcal{B}} \langle \boldsymbol{X}, \boldsymbol{H}_{ijl}\rangle < 1-\gamma \Rightarrow (i,j,l) \in \mathcal{L}^\star \tag{R1}$$

$$\min_{\boldsymbol{X}\in\mathcal{B}} \langle \boldsymbol{X}, \boldsymbol{H}_{ijl}\rangle > 1 \Rightarrow (i,j,l) \in \mathcal{R}^\star. \tag{R2}$$

We will show how to evaluate these rules efficiently. Since (R1) can be evaluated in the same way as (R2), we only deal with (R2) hereafter.

3.1.1 *Sphere Rule.*

Suppose that the optimal $\boldsymbol{M}^\star$ lies in a hypersphere defined by a center $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ and a radius $r \in \mathbb{R}_+$. To evaluate the condition of (R2), we consider the following minimization problem (P1):

$$\min_{\boldsymbol{X}} \langle \boldsymbol{X}, \boldsymbol{H}_{ijl}\rangle \text{ s.t. } \|\boldsymbol{X}-\boldsymbol{Q}\|_F^2 \le r^2. \tag{P1}$$

Letting $\boldsymbol{Y} := \boldsymbol{X} - \boldsymbol{Q}$, this problem is transformed into

$$\min_{\boldsymbol{Y}} \langle \boldsymbol{Y}, \boldsymbol{H}_{ijl}\rangle + \langle \boldsymbol{Q}, \boldsymbol{H}_{ijl}\rangle \text{ s.t. } \|\boldsymbol{Y}\|_F^2 \le r^2.$$

Since $\langle \boldsymbol{Q}, \boldsymbol{H}_{ijl}\rangle$ is a constant, this optimization problem is to minimize the inner product $\langle \boldsymbol{Y}, \boldsymbol{H}_{ijl}\rangle$ under the norm constraint. The optimal $\boldsymbol{Y}^\star$ of this optimization problem is easily derived as

$$\boldsymbol{Y}^\star = -r\boldsymbol{H}_{ijl}/\|\boldsymbol{H}_{ijl}\|_F,$$

and then the minimum value of (P1) is $\langle \boldsymbol{H}_{ijl}, \boldsymbol{Q}\rangle - r\|\boldsymbol{H}_{ijl}\|_F$. This derives the following *sphere rule*:

$$\langle \boldsymbol{H}_{ijl}, \boldsymbol{Q}\rangle - r\|\boldsymbol{H}_{ijl}\|_F > 1 \Rightarrow (i,j,l) \in \mathcal{R}^\star. \tag{5}$$

Obviously, this condition can be calculated immediately for given $\boldsymbol{Q}$ and r without any iterative procedure.

3.1.2 *Sphere Rule with Semi-definite Constraint.*

Since sphere rule does not utilize the positive semi-definiteness of $\boldsymbol{M}^\star$, a stronger rule can be constructed by incorporating semi-definite constraint into (P1):

$$\min_{\boldsymbol{X}} \langle \boldsymbol{X}, \boldsymbol{H}_{ijl}\rangle \text{ s.t. } \|\boldsymbol{X}-\boldsymbol{Q}\|_F^2 \le r^2,\ \boldsymbol{X} \succeq \boldsymbol{O}. \tag{P2}$$

Although the analytical solution is not available, (P2) can be solved efficiently by being transformed into the *Semi-Definite Least Squares* (SDLS) problem [18].

Suppose that a feasible solution $\boldsymbol{X}_0$ of (P2) satisfies $\langle \boldsymbol{X}_0, \boldsymbol{H}_{ijl}\rangle > 1$, because if $\langle \boldsymbol{X}_0, \boldsymbol{H}_{ijl}\rangle \le 1$, we immediately see that this triplet does not satisfy the condition of (R2). Under this condition, we consider the following problem instead of (P2):

$$\min_{\boldsymbol{X}\in\mathbb{R}^{d\times d}} \|\boldsymbol{X}-\boldsymbol{Q}\|_F^2 \text{ s.t. } \langle \boldsymbol{X}, \boldsymbol{H}_{ijl}\rangle = 1,\ \boldsymbol{X} \succeq \boldsymbol{O}. \tag{SDLS}$$

If the optimal value of this problem is greater than r^2, i.e., $\|X^\star - Q\|_F^2 > r^2$, we can deduce

$$\left\{ X \mid \langle X, H_{ijl} \rangle \leq 1,\ \|X - Q\|_F^2 \leq r^2,\ X \succeq O \right\} = \emptyset,$$

which indicates that the condition of (R2) is satisfied.

We derive the following dual problem of (SDLS) based on [18]:

$$\max_y \ D_{\mathrm{SDLS}}(y) := -\left\| [Q + yH_{ijl}]_+ \right\|_F^2 + 2Cy + \|Q\|_F^2\,,$$

where $y \in \mathbb{R}$ is a dual variable, and $C = 1$ for (R2) and $C = 1 - \gamma$ for (R1). Unlike the primal problem, the dual is an unconstrained problem which only has one variable y, and thus standard gradient-based algorithms rapidly converge. We call the quasi-Newton optimization for this problem *SDLS dual ascent method*. During the dual ascent, we can stop the iteration before convergence if $D_{\mathrm{SDLS}}(y)$ becomes larger than r^2, since the value of the dual problem does not exceed the value of the primal problem (weak duality).

Although the computation of $[Q + yH_{ijl}]_+$ requires eigenvalue decomposition, this computational requirement can be alleviated when the center Q of the hypersphere is positive semi-definite. From the definition, H_{ijl} has at most one negative eigenvalue, and then $Q + yH_{ijl}$ also has at most one negative eigenvalue. Let $\lambda_{\min}$ be the negative (minimum) eigenvalue of $Q + yH_{ijl}$, and $q_{\min}$ be the corresponding eigenvector. The projection $[Q + yH_{ijl}]_+$ can be expressed as $[Q + yH_{ijl}]_+ = (Q + yH_{ijl}) - \lambda_{\min} q_{\min} q_{\min}^\top$. Computation of the minimum eigenvalue and eigenvector is much easier than the full eigenvalue decomposition [16].

As a special case, when M is a diagonal matrix, the semi-definite constraint is reduced to the non-negative constraint, and analytical calculation of the rule (P2) is possible (see Appendix B [36]).

3.1.3 Sphere Rule with Linear Constraint. To reduce computational complexity, we here consider relaxing the semi-definite constraint into a linear constraint. Suppose that a region defined by a linear inequality $\{X \in \mathbb{R}^{d \times d} \mid \langle P, X \rangle \geq 0\}$ contains the semi-definite cone, i.e., $\mathbb{R}_+^{d \times d} \subseteq \{X \in \mathbb{R}^{d \times d} \mid \langle P, X \rangle \geq 0\}$, for which we will describe how to obtain $P \in \mathbb{R}^{d \times d}$ later. Using this relaxed constraint, the condition (R2) is

$$\min_X \ \langle X, H_{ijl} \rangle \text{ s.t. } \|X - Q\|_F^2 \leq r^2,\ \langle P, X \rangle \geq 0. \tag{P4}$$

This problem can be solved analytically by considering the KKT condition as follows (Appendix C [36]).

Theorem 3.1. *(Analytical solution of* (P4)*). The optimal solution of* (P4) *is as follows:*

$$\langle H_{ijl}, X^\star \rangle = \begin{cases} 0, & \text{if } H_{ijl} = aP, \\ \langle H_{ijl}, Q \rangle - r\|H_{ijl}\|_F, & \text{if } \langle P, Q - r\frac{H_{ijl}}{\|H_{ijl}\|_F} \rangle \geq 0, \\ \langle H_{ijl}, \frac{\beta P - H_{ijl}}{\alpha} + Q \rangle, & \text{otherwise,} \end{cases}$$

where a is a constant, and

$$\alpha = \sqrt{\frac{\|P\|_F^2 \|H_{ijl}\|_F^2 - \langle P, H_{ijl} \rangle^2}{r^2 \|P\|_F^2 - \langle P, Q \rangle^2}}, \beta = \frac{\langle P, H_{ijl} \rangle - \alpha \langle P, Q \rangle}{\|P\|_F^2}.$$

A simple way to obtain P is to utilize the projection onto the semi-definite cone. Let $A \in \mathbb{R}^{d \times d}$ be a matrix which is in outside of the semi-definite cone as illustrated in Figure 2. In the figure, A_+ is the projection of A onto the semi-definite cone. For example, when the projected gradient for the primal problem [32] is used as an optimizer, A can be an update of gradient descent $A = M - \eta \nabla P_\lambda(M)$ with some step size $\eta > 0$. Since $M - \eta \nabla P_\lambda(M)$ is projected onto the semi-definite cone at every iteration of the optimization, no additional calculation is required to obtain A and A_+. Defining $A_- := A - A_+$, for any $X \succeq O$, we obtain

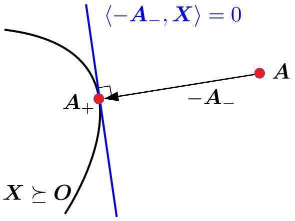


Figure 2: Linear relaxation of semi-definite constraint. From the projection of A to A_+, the supporting hyperplane $\langle -A_-, X \rangle = 0$ is constructed, and the halfspace $\{X \mid \langle -A_-, X \rangle \geq 0\}$ contains the semi-definite cone $X \succeq O$.

$$\langle A_+ - A, X - A_+ \rangle \geq 0 \ \Leftrightarrow \ \langle -A_-, X \rangle \geq 0.$$

The left inequality is from a property of a supporting hyperplane [3], and for the right inequality, we use $\langle A_+, A_- \rangle = 0$. By setting $P = -A_-$, we obtain a linear approximation of the semi-definite constraint which is a superset of the original semi-definite cone.

3.2 Sphere Bound

In previous section, we assume that the sphere region which contains the optimal $M^\star$ is available. In this section, we show that six variants of the regions created by three-types of different approaches. We here omit detailed derivation, and see Appendix in [36] for the proofs.

3.2.1 Gradient Bound (GB). We first introduce a hypersphere which we call *Gradient Bound* (GB) because the center and radius of the hypersphere are represented by the subgradient of the objective function:

Theorem 3.2. *(GB). Given any feasible solution $M \succeq O$, the optimal solution $M^\star$ for λ exists in the following hypersphere:*

$$\left\| M^\star - Q^{\mathrm{GB}}(M) \right\|_F^2 \leq \left(\frac{1}{2\lambda} \|\nabla P_\lambda(M)\|_F \right)^2,$$

where $Q^{\mathrm{GB}}(M) := M - \frac{1}{2\lambda} \nabla P_\lambda(M)$.

Sketch of proof. From the standard optimality condition of the convex optimization problem [2] (shown as Theorem D.1 in our Appendix D [36]), we obtain

$$\langle \nabla P_\lambda(M^\star), M^\star - M \rangle \leq 0,\ \forall M \succeq O. \tag{6}$$

In addition to this condition, we use the following two inequalities derived from the convexity of ℓ:

$$\ell(\langle M^\star, H_{ijl} \rangle) \geq \ell(\langle M, H_{ijl} \rangle) + \langle \Xi_{ijl}(M), M^\star - M \rangle,$$
$$\ell(\langle M, H_{ijl} \rangle) \geq \ell(\langle M^\star, H_{ijl} \rangle) + \langle \Xi_{ijl}(M^\star), M - M^\star \rangle,$$

where $\Xi_{ijl}(\boldsymbol{M})$ is an arbitrary subgradient at $\boldsymbol{M}$ of the loss function $\ell(\langle \boldsymbol{M}, \boldsymbol{H}_{ijl}\rangle)$. Theorem 3.2 is derived by combining three inequality shown above. See Appendix D [36] for the proof. □

This theorem is an extension of the sphere for SVM [28], which can be treated as a simple unconstrained problem.

Even when we substitute the optimal $\boldsymbol{M}^\star$ into the reference solution $\boldsymbol{M}$, the radius of GB is not guaranteed to be 0. By projecting the center of GB onto the feasible region (i.e., semi-definite cone), another GB based hypersphere can be derived, which has a radius converging to 0 at the optimal. We call this extension *Projected Gradient Bound* (PGB) for which a schematic illustration is shown as Figure 3(a). In Figure 3(a), the center of GB $\boldsymbol{Q}^{\mathrm{GB}}$ (abbreviation of $\boldsymbol{Q}^{\mathrm{GB}}(\boldsymbol{M})$) is projected onto the semi-definite cone which becomes a center of PGB $\boldsymbol{Q}_+^{\mathrm{GB}}$. The sphere of PGB can be written as follows:

Theorem 3.3. *(PGB). Given any feasible solution $\boldsymbol{M} \geq \boldsymbol{O}$, the optimal solution $\boldsymbol{M}^\star$ for λ exists in the following hypersphere:*

$$\left\|\boldsymbol{M}^\star - \left[\boldsymbol{Q}^{\mathrm{GB}}(\boldsymbol{M})\right]_+\right\|_F^2 \leq \left(\frac{1}{2\lambda}\left\|\nabla P_\lambda(\boldsymbol{M})\right\|_F\right)^2 - \left\|\left[\boldsymbol{Q}^{\mathrm{GB}}(\boldsymbol{M})\right]_-\right\|_F^2.$$

See Appendix E [36] for the proof. PGB contains the projections onto the positive and the negative semi-definite cone in the center and the radius, respectively. These projections require the eigenvalue decomposition of $\boldsymbol{M} - \frac{1}{2\lambda}\nabla P_\lambda(\boldsymbol{M})$. This decomposition, however, is necessary to perform only once for evaluating screening rules of all triplets. In the standard optimization procedures of RTLM, including [32], the eigenvalue decomposition of the $d \times d$ matrix is calculated at every iteration, and thus the computational complexity is not increased by PGB.

The following theorem shows a superior convergence property of PGB compared to GB:

Theorem 3.4. *There exist a subgradient $\nabla P_\lambda(\boldsymbol{M}^\star)$ such that the radius of PGB is* 0.

For the hinge loss, which is not differentiable at the "kink", the optimal dual variables provide subgradients which make the radius 0. This theorem is an immediate consequence from Appendix H [36], which is a proof for the relation between PGB and the other bound derived in section 3.2.3.

From Figure 3(a), we see that the half space $\langle -\boldsymbol{Q}_-^{\mathrm{GB}}, \boldsymbol{X}\rangle \geq 0$, where $\boldsymbol{Q}_-^{\mathrm{GB}} = \boldsymbol{Q}^{\mathrm{GB}} - \boldsymbol{Q}_+^{\mathrm{GB}}$, can be used as a linear relaxation of the semi-definite constraint for the linear constraint rule in section 3.1.3. Interestingly, GB with this linear constraint is tighter than PGB. This is proved in Appendix E [36], which is the proof of PGB.

3.2.2 Duality Gap Bound (DGB).

In this section, we describe *Duality Gap Bound* (DGB) in which the radius is represented by the duality gap:

Theorem 3.5. *(DGB). Let $\boldsymbol{M}$ be a feasible solution of the primal problem, and $\boldsymbol{\alpha}$ and Γ be feasible solutions of the dual problem, then the optimal solution of the primal problem $\boldsymbol{M}^\star$ exists in the following hypersphere:*

$$\left\|\boldsymbol{M}^\star - \boldsymbol{M}\right\|_F^2 \leq 2(P_\lambda(\boldsymbol{M}) - D_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}))/\lambda.$$

Sketch of proof. In general, a function $f(\boldsymbol{x})$ is m-strongly convex function if $f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|_2^2$ is a convex. Since the objective function $P_\lambda(\boldsymbol{M})$ is a λ-strongly convex function, we obtain

$$P_\lambda(\boldsymbol{M}) \geq P_\lambda(\boldsymbol{M}^\star) + \langle \nabla P_\lambda(\boldsymbol{M}^\star), \boldsymbol{M} - \boldsymbol{M}^\star\rangle + \frac{\lambda}{2}\left\|\boldsymbol{M} - \boldsymbol{M}^\star\right\|_F^2.$$

From the optimal condition (6), the second term on the right hand side is greater than or equal to 0, and from weak duality, $P_\lambda(\boldsymbol{M}^\star) \geq D_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})$. Therefore, we obtain Theorem 3.5. □

Since the radius is proportional to the square root of the duality gap, DGB obviously converges to 0 at the optimal solution (Figure 3(b)). For DGB, unlike the previous bounds, a dual feasible solution is necessary. This means that when a primal based optimization algorithm is employed, we need to create a dual feasible solution from a primal feasible solution. A simple way to create a dual feasible solution is to substitute the current $\boldsymbol{M}$ into $\boldsymbol{M}^\star$ of (3). On the other hand, when a dual based optimization algorithm is employed, a primal feasible solution can be created by (1).

For DGB, we further show that if the primal and dual reference solutions satisfy (1), the radius can be $\sqrt{2}$ times smaller. We extend a dual based screening of SVM [39] for RTLM.

Theorem 3.6. *(CDGB). Let $\boldsymbol{\alpha}$ and Γ be the feasible solutions of the dual problem, then the optimal solution of the primal problem $\boldsymbol{M}^\star$ exists in the following hypersphere:*

$$\left\|\boldsymbol{M}^\star - \boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})\right\|_F^2 \leq G_{D_\lambda}(\boldsymbol{\alpha}, \boldsymbol{\Gamma})/\lambda.$$

Sketch of proof. Let $G_{D_\lambda}(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) := P_\lambda(\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})) - D_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})$ be the duality gap as a function of the dual feasible solutions $\boldsymbol{\alpha}$ and $\boldsymbol{\Gamma}$. On the other hand, the following equation is the duality gap as a function of the primal feasible solution $\boldsymbol{M}$ in which the dual solutions are optimized:

$$G_{P_\lambda}(\boldsymbol{M}) := \min_{\substack{\boldsymbol{0} \leq \boldsymbol{\alpha} \leq \boldsymbol{1}, \\ \boldsymbol{\Gamma} \geq \boldsymbol{O}, \\ \boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) = \boldsymbol{M}}} G_{D_\lambda}(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) \quad = P_\lambda(\boldsymbol{M}) - \max_{\substack{\boldsymbol{0} \leq \boldsymbol{\alpha} \leq \boldsymbol{1}, \\ \boldsymbol{\Gamma} \geq \boldsymbol{O}, \\ \boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) = \boldsymbol{M}}} D_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}).$$

From the definition, we obtain

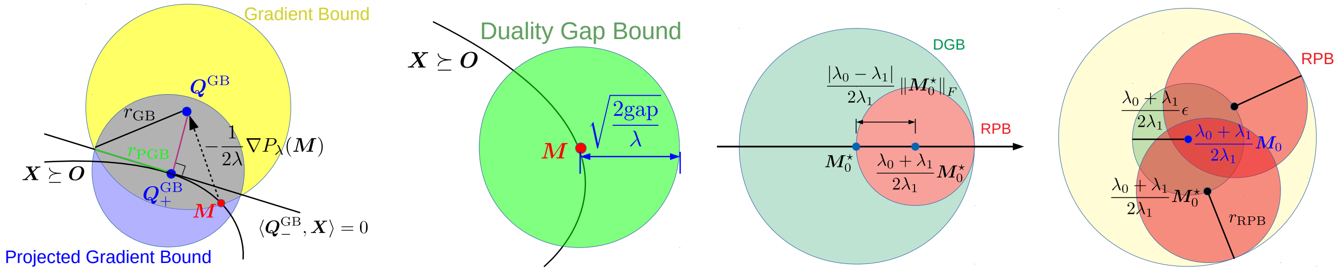
$$G_{D_\lambda}(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) \geq G_{P_\lambda}(\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})). \tag{7}$$

From the strong convexity of G_{P_λ} (Appendix F.1 [36]), we obtain

$$\begin{aligned} &G_{P_\lambda}(\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})) \geq G_{P_\lambda}(\boldsymbol{M}^\star) \\ &\quad + \langle \nabla G_{P_\lambda}(\boldsymbol{M}^\star), \boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) - \boldsymbol{M}^\star\rangle + \lambda\left\|\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) - \boldsymbol{M}^\star\right\|_F^2. \end{aligned} \tag{8}$$

Considering the optimality of $G_{P_\lambda}(\boldsymbol{M}^\star)$ and combining (7) and (8), Theorem 3.6 can be derived. See Appendix F.2 [36] for the proof. □

We call this bound *Constrained Duality Gap Bound* (CDGB). Since CDGB also has a radius proportional to the square root of the duality gap, the radius converges to 0 at the optimal solution. For primal based optimizers, additional calculation is necessary for $P_\lambda(\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}))$, while dual based optimizers calculates this term in the optimization process.


Gradient Bound
Q^{GB}
r_{GB}
$-\frac{1}{2\lambda}\nabla P_\lambda(M)$
r_{PGB}
$X \succeq O$
Q_+^{GB}
M
$\langle Q_-^{\mathrm{GB}}, X\rangle = 0$
Projected Gradient Bound
Duality Gap Bound
$X \succeq O$
M
$\sqrt{\frac{2\mathrm{gap}}{\lambda}}$
DGB
$\frac{|\lambda_0-\lambda_1|}{2\lambda_1}\|M_0^\star\|_F$
RPB
$M_0^\star$
$\frac{\lambda_0+\lambda_1}{2\lambda_1}M_0^\star$
RPB
$\frac{\lambda_0+\lambda_1}{2\lambda_1}\epsilon$
$\frac{\lambda_0+\lambda_1}{2\lambda_1}M_0$
$\frac{\lambda_0+\lambda_1}{2\lambda_1}M_0^\star$
r_{RPB}

(a) Projected Gradient Bound (PGB). The center of GB $\boldsymbol{Q}^{\mathrm{GB}}$ is projected onto the semi-definite cone. The radius is calculated as $r_{\mathrm{PGB}}^2 = r_{\mathrm{GB}}^2 - \|\boldsymbol{Q}^{\mathrm{GB}} - \boldsymbol{Q}_+^{\mathrm{GB}}\|_F^2$.

(b) Duality Gap Bound (DGB). The center is the reference solution, and the radius is determined from the duality gap.

(c) Relation between DGB and RPB (Theorem 3.9).

(d) Relaxed Regularization Path Bound (RRPB). RRPB "blurs" the center of RPB based on the current approximate solution.

Figure 3: Illustrations of sphere bounds.

3.2.3 Regularization Path Bound (RPB). In [30], a hypersphere specific for *regularization path* is proposed, in which the optimization problem should be solved for a sequence of λs. Suppose that the optimization for λ_0 is already finished and the optimization for λ_1 is necessary to solve. Then, the same approach as [30] is applicable to our RTLM, which derives a bound depending on the optimal solution for λ_0 as a reference solution:

Theorem 3.7. *(RPB). Let $\boldsymbol{M}_0^\star$ be the optimal solution for λ_0, the optimal solution $\boldsymbol{M}_1^\star$ for λ_1 exists in the following hypersphere:*

$$\left\|\boldsymbol{M}_1^\star - \frac{\lambda_0+\lambda_1}{2\lambda_1}\boldsymbol{M}_0^\star\right\|_F^2 \le \left(\frac{\lambda_0-\lambda_1}{2\lambda_1}\|\boldsymbol{M}_0^\star\|_F\right)^2.$$

Sketch of proof. Let $\boldsymbol{\alpha}_i^\star$ and $\boldsymbol{\Gamma}_i^\star$ be the optimal dual solutions for $\lambda_i (i \in \{0,1\})$. From the optimality condition of the convex optimization problem [2], which is also called *variational inequality* [30], to (Dual1), we obtain the following two inequalities

$$\nabla_{\boldsymbol{\alpha}} D_{\lambda_0}(\boldsymbol{\alpha}_0^\star, \boldsymbol{\Gamma}_0^\star)^\top(\boldsymbol{\alpha}_1^\star - \boldsymbol{\alpha}_0^\star) + \langle \nabla_{\boldsymbol{\Gamma}} D_{\lambda_0}(\boldsymbol{\alpha}_0^\star, \boldsymbol{\Gamma}_0^\star), \boldsymbol{\Gamma}_1^\star - \boldsymbol{\Gamma}_0^\star\rangle \le 0,$$
$$\nabla_{\boldsymbol{\alpha}} D_{\lambda_1}(\boldsymbol{\alpha}_1^\star, \boldsymbol{\Gamma}_1^\star)^\top(\boldsymbol{\alpha}_0^\star - \boldsymbol{\alpha}_1^\star) + \langle \nabla_{\boldsymbol{\Gamma}} D_{\lambda_1}(\boldsymbol{\alpha}_1^\star, \boldsymbol{\Gamma}_1^\star), \boldsymbol{\Gamma}_0^\star - \boldsymbol{\Gamma}_1^\star\rangle \le 0.$$

Note that dual variables should be in the feasible region $\boldsymbol{0} \le \boldsymbol{\alpha} \le \boldsymbol{1}, \boldsymbol{\Gamma} \succeq \boldsymbol{O}$. Combining these two inequalities, we obtain the Theorem 3.7. See Appendix G [36] for the proof. □

We call this bound *Regularization Path Bound* (RPB).

RPB requires the theoretically optimal solution $\boldsymbol{M}_0^\star$, which is numerically impossible. Furthermore, since the reference solution is fixed on $\boldsymbol{M}_0^\star$, RPB can be performed only once for a specific pair of λ_0 and λ_1 even if the optimal $\boldsymbol{M}_0^\star$ is available. The other bounds can be performed multiple times during the optimization by regarding the current approximate solution as a reference solution. On the other hand, RPB provides interesting insights about relations with PGB and DGB. The following theorem describes the relation between PGB and RPB:

Theorem 3.8. *(Relationship between PGB and RPB). Suppose that the optimal solution $\boldsymbol{M}_0^\star$ for λ_0 is substituted into the reference solution $\boldsymbol{M}$ of PGB. Then, there exist a subgradient $\nabla P_{\lambda_1}(\boldsymbol{M}_0^\star)$ by which PGB and RPB provides the same center and the radius for $\boldsymbol{M}_1^\star$.*

See Appendix H [36] for the proof. The following theorem describes the relation between DGB and RPB:

Theorem 3.9. *(Relationship between DGB and RPB). Suppose that the optimal solutions $\boldsymbol{M}_0^\star, \boldsymbol{\alpha}_0^\star, \boldsymbol{\Gamma}_0^\star$ for λ_0 are substituted into the reference solutions $\boldsymbol{M}, \boldsymbol{\alpha}$ and $\boldsymbol{\Gamma}$ of DGB. Then, the radius of DGB and RPB for λ_1 has a relation $r_{\mathrm{DGB}} = 2\,r_{\mathrm{RPB}}$, and the hypersphere of RPB is included in the hypersphere of DGB.*

See Appendix I [36] for the proof. Figure 3(c) illustrates the relation between DGB and RPB which shows theoretical advantage of RPB for the regularization path setting. For practical use of RPB, we modify RPB in such a way that the approximate solution can be used as a reference solution. Assuming that $\boldsymbol{M}_0$ satisfy

$$\|\boldsymbol{M}_0^\star - \boldsymbol{M}_0\|_F \le \epsilon,$$

where $\epsilon \ge 0$ is a constant. Given $\boldsymbol{M}_0$ which satisfy the above condition, we obtain *Relaxed Regularization Path Bound* (RRPB):

Theorem 3.10. *(RRPB). Let $\boldsymbol{M}_0$ be an approximate solution for λ_0 which satisfies $\|\boldsymbol{M}_0^\star - \boldsymbol{M}_0\|_F \le \epsilon$. The optimal solution $\boldsymbol{M}_1^\star$ for λ_1 exists in the following hypersphere:*

$$\left\|\boldsymbol{M}_1^\star - \frac{\lambda_0+\lambda_1}{2\lambda_1}\boldsymbol{M}_0\right\|_F^2 \le \left(\frac{|\lambda_0-\lambda_1|}{2\lambda_1}\|\boldsymbol{M}_0\|_F + \frac{|\lambda_0-\lambda_1|+\lambda_0+\lambda_1}{2\lambda_1}\epsilon\right)^2.$$

See Appendix J [36] for the proof. An intuition behind RRPB is shown in Figure 3(d), in which the approximation error for the center of RPB is depicted. In the theorem, RRPB also considers the error in the radius though it is not illustrated in the figure for simplicity. To the best of our knowledge, this approach has not been introduced in other existing screening studies.

For example, ϵ can be set from Theorem 3.5 (DGB) as follows:

$$\epsilon = \sqrt{2(P_{\lambda_0}(\boldsymbol{M}_0) - D_{\lambda_0}(\boldsymbol{\alpha}_0, \boldsymbol{\Gamma}_0))/\lambda_0}.$$

When the optimization for λ_0 terminates, the solution $\boldsymbol{M}_0$ should be accurate in terms of some stopping criterion such as the duality gap. Then, ϵ is expected to be quite small, and RRPB can provide a tight bound for λ_1, which is close to the ideal (but not computable) RPB. As a special case, by setting $\lambda_1 = \lambda_0$, RRPB is applicable to perform screening of λ_1 using any approximate solution having $\|\boldsymbol{M}_1^\star - \boldsymbol{M}\|_F \le \epsilon$, and then RRPB is equivalent to DGB.

3.3 Computational Cost

Considering computational cost of the screening procedure, the rule evaluation (**Step 2**) described in section 3.1 is often dominant, because the rule needs to be evaluated for each one of triplets. On the other hand, the sphere, constructed in **Step 1**, can be fixed during the screening procedure as long as the reference solution is fixed.

Sphere Rule (section 3.1.1) needs $O(d^2)$ computations for the inner product $\langle \boldsymbol{H}_{ijl}, \boldsymbol{Q} \rangle$, but we can reuse this term from objective function $P_\lambda(\boldsymbol{M})$ calculation in the case of DGB, RPB and RRPB. The computational cost of Sphere Rule with Semi-definite Constraint (section 3.1.2) is that of SDLS algorithm. SDLS algorithm needs $O(d^3)$ because of the eigenvalue decomposition at every iteration, which may cause large computational cost. The calculation cost of Sphere Rule with Linear Constraint (section 3.1.3) takes $O(d^2)$.

4 RANGE BASED EXTENSION OF TRIPLET SCREENING

The screening rules shown in section 3.1 provides the conditions for the problem of a fixed λ. In this section, by considering λ as a variable, we derive a range of λ in which the screening rule is guaranteed to be satisfied. This is particularly useful for the regularization path calculation for which we need to optimize the metric for a sequence of λs. If a screening rule is satisfied for a triplet (i, j, l) in a range (λ_a, λ_b), we can fix the triplet (i, j, l) in $\hat{\mathcal{L}}$ or $\hat{\mathcal{R}}$ as long as λ is in (λ_a, λ_b), without computing screening rules.

Let $\boldsymbol{Q} = \boldsymbol{A} + \boldsymbol{B}\frac{1}{\lambda}$ be a general form of hypersphere for some constant matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{B} \in \mathbb{R}^{d \times d}$, and $r^2 = a + b\frac{1}{\lambda} + c\frac{1}{\lambda^2}$ be a general form of the radius for some constants $a \in \mathbb{R}$, $b \in \mathbb{R}$ and $c \in \mathbb{R}$. GB, DGB, RPB and RRPB can be in this form (see Appendix K.1 [36] for detail). The sphere rule for $\mathcal{R}^\star$ (5) is equivalent to the intersection of the following two inequalities because of $r \geq 0$ and $\|\boldsymbol{H}_{ijl}\|_F \geq 0$:

$$\langle \boldsymbol{H}_{ijl}, \boldsymbol{Q} \rangle - 1 > 0, \ (\langle \boldsymbol{H}_{ijl}, \boldsymbol{Q} \rangle - 1)^2 > r^2 \|\boldsymbol{H}_{ijl}\|_F^2.$$

Since $\langle \boldsymbol{H}_{ijl}, \boldsymbol{Q} \rangle = \langle \boldsymbol{H}_{ijl}, \boldsymbol{A} \rangle + \langle \boldsymbol{H}_{ijl}, \boldsymbol{B} \rangle \frac{1}{\lambda}$, The first and second inequalities can be transformed into linear and quadratic functions of λ respectively, for which it is easy to find the range of λ satisfying these two inequalities. The following theorem shows the range for the case of RRPB given a reference solution $\boldsymbol{M}_0$ which is an approximate solution for λ_0:

Theorem 4.1. *(Range Based Extension of RRPB). Assuming $\langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle - 2 + \|\boldsymbol{H}_{ijl}\|_F \|\boldsymbol{M}_0\|_F > 0$ and $\|\boldsymbol{M}_0^\star - \boldsymbol{M}_0\|_F \leq \epsilon$, a triplet (i, j, l) is guaranteed to be in $\mathcal{R}^\star$ for the following range of λ:*

$$\lambda \in (\lambda_a, \lambda_b),$$

where

$$\lambda_a = \frac{\lambda_0 \left(\|\boldsymbol{M}_0\|_F \|\boldsymbol{H}_{ijl}\|_F - \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle + 2\epsilon \|\boldsymbol{H}_{ijl}\|_F \right)}{\langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle - 2 + \|\boldsymbol{H}_{ijl}\|_F \|\boldsymbol{M}_0\|_F},$$

$$\lambda_b = \frac{\lambda_0 \left(\|\boldsymbol{M}_0\|_F \|\boldsymbol{H}_{ijl}\|_F + \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle \right)}{\|\boldsymbol{H}_{ijl}\|_F \|\boldsymbol{M}_0\|_F - \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle + 2 + 2\epsilon \|\boldsymbol{H}_{ijl}\|_F}.$$

See Appendix K.2 [36] for the proof.

Table 1: Summary of datasets. ∗1 :The dimension was reduced by AutoEncoder. ∗2 :The dimension was reduced by PCA. #triplet and $\lambda_{\min}$ are the average value for sub-sampled random trials.

	#dimension	#sample	#classes	k	#triplet	$\lambda_{\max}$	$\lambda_{\min}$
segment	19	2310	7	20	832000	2.5e+6	4.2e+0
phishing	68	11055	2	7	487550	5.0e+3	2.0e−1
SensIT Vehicle	100	78823	3	3	638469	1.0e+4	2.9e+0
a9a^{*1}	16	32561	2	5	732625	1.2e+5	3.1e+2
mnist^{*1}	32	60000	10	5	1350025	7.0e+3	9.6e−1
cifar10^{*1}	200	50000	10	2	180004	2.0e+3	3.3e+1
rcv1.multiclass^{*2}	200	15564	53	3	126018	3.0e+2	6.0e−4

5 EXPERIMENT

We evaluate performance of safe triplet screening using the benchmark datasets shown in Table 1, which are from LIBSVM [6] and Keras Dataset [7]. To create a set of triplets, we follow the approach by [26], in which k neighborhoods in the same class $\boldsymbol{x}_j$ and k neighborhoods in different class $\boldsymbol{x}_l$ are sampled for each $\boldsymbol{x}_i$. We employed the regularization path setting in which RTLM is optimized for a sequence of $\lambda_0, \lambda_1, \ldots, \lambda_T$. The initial $\lambda_0 = \lambda_{\max}$ was set by a sufficiently large value in which $\mathcal{R}^\star$ starts increasing from the empty set. To generate the next value of λ, we used $\lambda_t = 0.9\lambda_{t-1}$, and the path terminated when the following condition is satisfied:

$$\frac{\mathrm{loss}(\lambda_{t-1}) - \mathrm{loss}(\lambda_t)}{\mathrm{loss}(\lambda_{t-1})} \times \frac{\lambda_{t-1}}{\lambda_{t-1} - \lambda_t} < 0.01,$$

where $\mathrm{loss}(\lambda_t)$ is the loss function value at λ_t. We randomly selected 90% of the instances of each dataset 5 times, and the average is shown as the experimental result. As a base optimizer, we employed the projected gradient descent of the primal problem, and the iteration terminated when the duality gap becomes less than 10^{-6}. For the loss function ℓ, we used the smoothed hinge loss of $\gamma = 0.05$ (We also provides results for the hinge loss in Appendix L.1 [36]). We performed safe triplet screening every ten iterations of the gradient descent. We refer to the first screening for a specific λ_t, in which the solution of previous λ_{t-1} is used for the reference solution, as the *regularization path screening*. On the other hand, the screening performed during the optimization process (after regularization path screening) is called *dynamic screening*. For all experiments, we performed both of these screening procedures. As a baseline, we call the RTLM optimization without screening *naive optimization*. When the regularization coefficient changes, $\boldsymbol{M}$ starts from the previous solution $\hat{\boldsymbol{M}}$ (warm start). The step size of the gradient descent was determined by

$$\frac{1}{2} \left| \frac{\langle \Delta \boldsymbol{M}, \Delta \boldsymbol{G} \rangle}{\langle \Delta \boldsymbol{G}, \Delta \boldsymbol{G} \rangle} + \frac{\langle \Delta \boldsymbol{M}, \Delta \boldsymbol{M} \rangle}{\langle \Delta \boldsymbol{M}, \Delta \boldsymbol{G} \rangle} \right|,$$

where $\Delta \boldsymbol{M} = \boldsymbol{M}_t - \boldsymbol{M}_{t-1}$, $\Delta \boldsymbol{G} = \nabla P_\lambda(\boldsymbol{M}_t) - \nabla P_\lambda(\boldsymbol{M}_{t-1})$ [1]. In SDLS dual ascent, we used the conjugate gradient method [35] for finding the minimum eigenvalue.

5.1 Comparing GB Based Rules

We first validate the screening performance (screening rate and CPU time) of each screening rule introduced in the section 3.1. We here use GB and PGB as spheres, and observe the effect of the

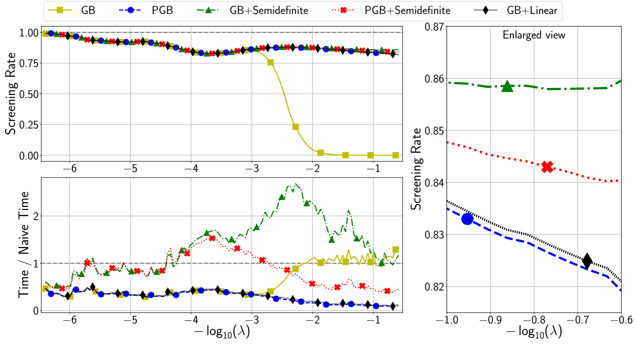


Figure 4: Screening rule comparison on segment dataset. The top left plot shows performance of regularization path screening, and the bottom left plot shows the ratio of the CPU time compared with the naive optimization for each λ. The right plot enlarges the upper left plot for a range $-\log_{10}(\lambda) \in [-1, -0.6]$.

semi-definite constraint in the rules. As a representative result, comparison on the segment data is shown in Figure 4.

First of all, we see that the rules except for GB keep the high screening rate for the entire regularization path shown as the top left plot. Note that this rate is only for regularization path screening, meaning that dynamic screening can further increase screening rate during the optimization as we see next subsection. The bottom left plot of the same figure shows PGB and GB+Linear are most efficient which achieved about 2-10 times faster CPU time than the naive optimization. The screening rate of GB severely dropped on the later half of the regularization path. As illustrated in Figure 3(a), the center of GB can be outside of the semi-definite cone by which the sphere of GB contains a larger proportion of the region violating the constraint $\boldsymbol{M} \succeq \boldsymbol{O}$, compared with the spheres having their center inside the semi-definite cone. This causes performance deterioration particularly for smaller λ, because the minimum of the loss term is usually in outside of the semi-definite cone.

The screening rate of GB+Linear and GB+Semidefinite are slightly higher than that of PGB (the right plot), which can be seen from the geometrical relation of them illustrated in Figure 3(a). GB+Semidefinite achieved the highest screening rate, but the eigenvalue decomposition is necessary to calculate repeatedly in SDLS, by which the CPU time increased in the later half of the path. Although PGB+Semidefinite is also tighter than PGB, the CPU time increased around from $-\log_{10}(\lambda) \approx -4$ to -3. Since the center of PGB is positive semi-definite, only the minimum eigenvalue is required (see section 3.1.2), but it still can increase the CPU time.

Among screening methods compared here, our empirical analysis suggests that the sphere rule with PGB is most cost-effective, in which semi-definite constraint is implicitly incorporated at the projection process. We did not observe that the other approach to considering the semi-definite (or relaxed linear) constraint in the rule substantially outperform PGB in terms of the CPU time despite their high screening rate. We observed the same tendency for DGB. The screening rate did not largely change even if the semi-definite constraint is explicitly taken into account (see Appendix L.2 [36]).

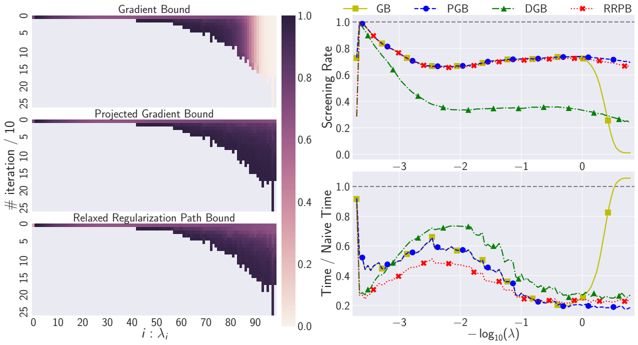


Figure 5: Comparison of sphere bounds on phishing dataset. The left heatmaps show the screening rate of dynamic screening. The vertical axis is the number of screening performed for λ_i (once every ten iterations of gradient descent). The top right plot shows the rate of regularization path screening, and the bottom right plot shows the ratio of the CPU time compared with the naive optimization.

5.2 Comparing Bounds

We next compare screening performance (screening rate and CPU time) of each bound introduced in the section 3.2. Based on the results in the previous section, we employed the sphere rule. The result of the phishing dataset are shown in Figure 5. Screening rate (the top right plot) of GB again dropped from the middle compared with the other spheres. Screening rate (the top right plot) of GB again dropped from the middle compared with the other spheres. The other spheres also have lower screening rates for small λs. As we mention in section 4, the radiuses of GB, DGB, RPB and RRPB have the form $r^2 = a + b\frac{1}{\lambda} + c\frac{1}{\lambda^2}$, meaning that if $\lambda \to 0$ then $r \to \infty$. For PGB, although dependency on λ can not be written explicitly, the same tendency was observed. We see that PGB and RRPB have similar results as suggested by Theorem 3.8, and the screening rate of DGB is lower than RRPB as suggested by Theorem 3.9. Comparing PGB and RRPB, PGB achieved the higher screening rate, but RRPB shows the faster CPU time (the bottom right plots), because PGB requires a matrix inner product calculation for each triplet. We see that the bounds other than GB are more than two times faster than the naive calculation for most of λs.

Comparing the dynamic screening rate (the left three plots of Figure 5) of PGB and RRPB, PGB has the higher screening rate. For the regularization path screening (the top right), RRPB and PGB have similar screening rate, but for the dynamic screening, PGB has the higher rate. For the later half of the regularization path, the number of gradient descent iterations increases, by which the dynamic screening significantly effects on the CPU time, and PGB becomes faster despite its additional computation for the inner product. In Appendix L.3 [36], we show the CPU time for the entire path with some additional datasets.

We further evaluate performance of the range based extension described in section 4. Figure 6 shows the rate of the range based screening for the segment dataset. We see that a wide range of λ can be screened particularly for small λ, and for large λ, although the range is smaller than the small λ cases, high screening rate was

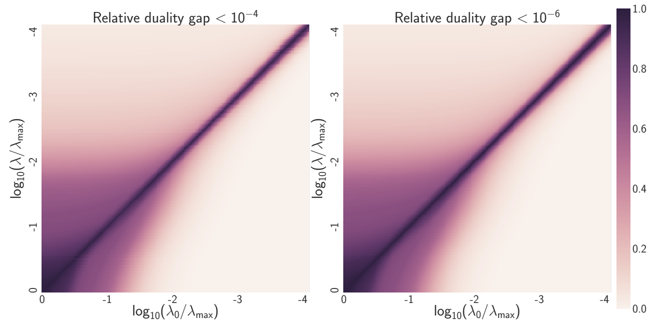


Figure 6: Screening rate of the range based screening on the segment dataset. The color indicates the screening rate for λ in the vertical axis based on the reference solution at λ_0 in the horizontal axis. The accuracy of the reference solution is 10^{-4} for the left plot and 10^{-6} for the right plot.

Table 2: Total CPU time (sec) evaluation with the active set method. The results with * indicates the fastest method.

Method\Dataset	phishing	SensIT	a9a	mnist	cifar10	rcv
ActiveSet	7989.5	16352.1	758.7	3788.1	11085.7	94996.3
ActiveSet+RRPB	*2126.2	3555.6	*70.1	*871.1	1431.3	43174.9
ActiveSet+RRPB+PGB	2133.2	*3046.9	72.1	897.9	*1279.7	*38231.1

observed for λ close to λ_0. A significant advantage of this approach is that, for triplets screened by the range, we do not need to evaluate screening rule anymore as long as λ is in the range.

5.3 Practical Performance Evaluation

As a computationally more expensive setting, we consider investigating the regularization path in more detail by setting $\lambda_t = 0.99\lambda_{t-1}$. To evaluate practical performance, we combine our safe triplet screening with the well-known *active set* heuristics. In the active set method, only a subset of triplets whose loss is greater than 0 are treated as the active set. The gradient is calculated by only using the active set, and the overall optimality is confirmed when the iteration converges. We employed the active set update strategy shown by Weinberger et al. [32], in which the active set is updated once every ten iterations.

Table 2 shows the CPU time comparison for the entire regularization path. Based on the the results in the previous section, we employed RRPB and RRPB+PGB (evaluating rules based on both spheres) for the triplet screening. Further, the range based screening described in section 4 is also performed using RRPB, for which we evaluate the range at the beginning of the optimization for each λ. We see that our safe triplet screening accelerates the optimization process about up to 10 times from the simple active set method. The results for higher dimensional datasets with diagonal $\boldsymbol{M}$ are also shown in Appendix L.4 [36].

6 SUMMARY

We introduced *safe triplet screening* for large margin metric learning. The three screening rules and the six sphere bounds were derived, and their relation was analyzed. We further proposed range based extension for the regularization path calculation. Our screening technique for metric learning is particularly significant compared with other screening studies due to massiveness of triplets and the semi-definite constraint. Our numerical experiments verified effectiveness of safe triplet screening using several benchmark datasets.

ACKNOWLEDGMENTS

This work was supported by MEXT KAKENHI to I.T. (16H06538, 17H00758) and M.K. (16H06538, 17H04694); from JST CREST awarded to I.T. (JPMJCR1302, JPMJCR1502) and PRESTO awarded to M.K. (JPMJPR15N2); from the MI2I project of the Support Program for Starting Up Innovation Hub from JST awarded to I.T., and M.K.; and from RIKEN Center for AIP awarded to I.T.

REFERENCES

[1] J. Barzilai and J. M. Borwein. 1988. Two-point step size gradient methods. *IMA journal of numerical analysis* 8, 1 (1988), 141–148.
[2] D. P. Bertsekas. 1999. *Nonlinear programming.* Athena scientific Belmont.
[3] S. Boyd and L. Vandenberghe. 2004. *Convex optimization.* Cambridge university press.
[4] S. Boyd and L. Xiao. 2005. Least-squares covariance matrix adjustment. *SIAM J. Matrix Anal. Appl.* 27, 2 (2005), 532–546.
[5] H. L. Capitaine. 2016. Constraint selection in metric learning. *arXiv preprint arXiv:1612.04853* (2016).
[6] C.-C. Chang and C.-J. Lin. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 27.
[7] F. Chollet et al. 2015. Keras. https://github.com/keras-team/keras. (2015).
[8] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. 2007. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning.* ACM, 209–216.
[9] O. Fercoq, A. Gramfort, and J. Salmon. 2015. Mind the duality gap: safer rules for the lasso. *arXiv preprint arXiv:1505.03410* (2015).
[10] L. E. Ghaoui, V. Viallon, and T. Rabbani. 2010. Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219* (2010).
[11] H. Hanada, A. Shibagaki, J. Sakuma, and I. Takeuchi. 2018. Efficiently Monitoring Small Data Modification Effect for Large-Scale Learning in Changing Environment. In *Proceedings of The 32nd AAAI Conference on Artificial Intelligence.* 1314–1321.
[12] E. Hoffer and N. Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition.* Springer, 84–92.
[13] P. Jain, B. Kulis, I. S. Dhillon, and K. Grauman. 2009. Online metric learning and fast similarity search. In *Advances in neural information processing systems.* 761–768.
[14] B. Kulis et al. 2013. Metric learning: A survey. *Foundations and Trends® in Machine Learning* 5, 4 (2013), 287–364.
[15] S. Lee and E. P. Xing. 2014. Screening rules for overlapping group lasso. *arXiv preprint arXiv:1410.6880* (2014).
[16] R. B. Lehoucq and D. C. Sorensen. 1996. Deflation techniques for an implicitly restarted Arnoldi iteration. *SIAM J. Matrix Anal. Appl.* 17, 4 (1996), 789–821.
[17] J. Liu, Z. Zhao, J. Wang, and J. Ye. 2014. Safe Screening with Variational Inequalities and Its Application to Lasso. In *International Conference on Machine Learning.* 289–297.
[18] J. Malick. 2004. A dual approach to semidefinite least-squares problems. *SIAM J. Matrix Anal. Appl.* 26, 1 (2004), 272–284.
[19] B. McFee and G. R. Lanckriet. 2010. Metric learning to rank. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10).* 775–782.
[20] K. Nakagawa, S. Suzumura, M. Karasuyama, K. Tsuda, and I. Takeuchi. 2016. Safe pattern pruning: An efficient approach for predictive pattern mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 1785–1794.
[21] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. 2016. Gap safe screening rules for sparse-group lasso. In *Advances in Neural Information Processing Systems.* 388–396.
[22] K. Ogawa, Y. Suzuki, and I. Takeuchi. 2013. Safe screening of non-support vectors in pathwise SVM computation. In *Proceedings of the 30th International Conference on Machine Learning.* 1382–1390.
[23] S. Okumura, Y. Suzuki, and I. Takeuchi. 2015. Quick sensitivity analysis for incremental data modification and its application to leave-one-out cv in linear classification problems. In *Proceedings of the 21th ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining. ACM, 885–894.

[24] F. Schroff, D. Kalenichenko, and J. Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[25] M. Schultz and T. Joachims. 2004. Learning a distance metric from relative comparisons. In *Advances in neural information processing systems*. 41–48.

[26] C. Shen, J. Kim, F. Liu, L. Wang, and A. Van Den Hengel. 2014. Efficient dual approach to distance metric learning. *IEEE transactions on neural networks and learning systems* 25, 2 (2014), 394–406.

[27] A. Shibagaki, M. Karasuyama, K. Hatano, and I. Takeuchi. 2016. Simultaneous safe screening of features and samples in doubly sparse modeling. In *Proceedings of the 33rd International Conference on Machine Learning*. 1577–1586.

[28] A. Shibagaki, Y. Suzuki, M. Karasuyama, and I. Takeuchi. 2015. Regularization path of cross-validation error lower bounds. In *Advances in Neural Information Processing Systems 2015*. 1675–1683.

[29] T. Takada, H. Hanada, Y. Yamada, J. Sakuma, and I. Takeuchi. 2016. Secure Approximation Guarantee for Cryptographically Private Empirical Risk Minimization. In *Proceedings of The 8th Asian Conference on Machine Learning*. 126–141.

[30] J. Wang, P. Wonka, and J. Ye. 2014. Scaling SVM and least absolute deviations via exact data reduction. In *International Conference on Machine Learning*. 523–531.

[31] J. Wang, J. Zhou, P. Wonka, and J. Ye. 2013. Lasso screening rules via dual polytope projection. In *Advances in Neural Information Processing Systems*. 1070–1078.

[32] K. Q. Weinberger and L. K. Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10, Feb (2009), 207–244.

[33] Z. J. Xiang, Y. Wang, and P. J. Ramadge. 2017. Screening tests for lasso problems. *IEEE transactions on pattern analysis and machine intelligence* 39, 5 (2017), 1008–1027.

[34] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng. 2003. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*. 521–528.

[35] H. Yang. 1993. Conjugate gradient methods for the Rayleigh quotient minimization of generalized eigenvalue problems. *Computing* 51, 1 (1993), 79–94.

[36] T. Yoshida, I. Takeuchi, and M. Karasuyama. 2018. Safe Triplet Screening for Distance Metric Learning. *arXiv preprint arXiv:1802.03923* (2018).

[37] W. Zhang, B. Hong, W. Liu, J. Ye, D. Cai, X. He, and J. Wang. 2016. Scaling Up Sparse Support Vector Machines by Simultaneous Feature and Sample Reduction. *arXiv preprint arXiv:1607.06996* (2016).

[38] Q. Zhou and Q. Zhao. 2015. Safe subspace screening for nuclear norm regularized least squares problems. In *International Conference on Machine Learning*. 1103–1112.

[39] J. Zimmert, C. S. de Witt, G. Kerg, and M. Kloft. 2015. Safe screening for support vector machines. In *NIPS 2015 Workshop on Optimization in Machine Learning (OPT)*.